Biology of Blood and Marrow Transplantation 12:252-266 (2006)

1083-8791/06/1203-0002$32.00/0
doi:10.1016/j.bbmt.2006.01.008

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

Steven Z. Pavletic,[1] Paul Martin,[2] Stephanie J. Lee,[3] Sandra Mitchell,[1] David Jacobsohn,[4] Edward W. Cowen,[1] Maria L. Turner,[1] Gorgun Akpek,[5] Andrew Gilman,[6] George McDonald,[2] Mark Schubert,[2] Ann Berger,[7] Peter Bross,[8] Jason W. Chien,[2] Daniel Couriel,[9] J. P. Dunn,[10] Jane Fall-Dickson,[11] Ann Farrell,[8] Mary E. D. Flowers,[2] Hildegard Greinix,[12] Steven Hirschfeld,[8] Lynn Gerber,[7] Stella Kim,[9] Robert Knobler,[12] Peter A. Lachenbruch,[8] Frederick W. Miller,[13] Barbara Mittleman,[14] Esperanza Papadopoulos,[15] Susan K. Parsons,[16] Donna Przepiorka,[17] Michael Robinson,[18] Michael Ward,[14] Bryce Reeve,[1] Lisa G. Rider,[13] Howard Shulman,[2] Kirk R. Schultz,[19] Daniel Weisdorf,[20] Georgia B. Vogelsang[10]

[1]National Cancer Institute, National Institutes of Health, Bethesda, Maryland; [2]Fred Hutchinson Cancer Research Center, University of Washington School of Medicine, Seattle, Washington; [3]Dana-Farber Cancer Institute, Boston, Massachusetts; [4]Children's Memorial Hospital, Northwestern University School of Medicine, Chicago, Illinois; [5]University of Maryland School of Medicine, Baltimore, Maryland; [6]University of North Carolina School of Medicine, Chapel Hill, North Carolina; [7]Warren Grant Magnuson Clinical Center, National Institutes of Health, Bethesda, Maryland; [8]US Food and Drug Administration, Rockville, Maryland; [9]University of Texas M.D. Anderson Cancer Center, Houston, Texas; [10]Johns Hopkins University School of Medicine, Baltimore, Maryland; [11]National Institute of Nursing Research, National Institutes of Health, Bethesda, Maryland; [12]Medical University of Vienna, Austria; [13]National Institute of Environmental Health Sciences, National Institutes of Health, Bethesda, Maryland; [14]National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, Maryland; [15]Memorial Sloan-Kettering Cancer Center, New York, New York; [16]Tufts-New England Medical Center, Boston, Massachusetts; [17]University of Tennessee, Memphis, Tennessee; [18]National Eye Institute; National Institutes of Health; Bethesda, Maryland; [19]University of British Columbia, British Columbia Children's Hospital, Vancouver, British Columbia, Canada; [20]University of Minnesota, Minneapolis, Minnesota

Correspondence and reprint requests: Steven Z. Pavletic, MD, Graft-versus-Host and Autoimmunity Unit, Experimental Transplantation and Immunology Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-1203 (e-mail: pavletis@mail.nih.gov).

Received January 17, 2006; accepted January 18, 2006

ABSTRACT

The lack of standardized criteria for quantitative measurement of therapeutic response in clinical trials poses a major obstacle for the development of new agents in chronic graft-versus-host disease (GVHD). This consensus document was developed to address several objectives for response criteria to be used in chronic GVHD-related clinical trials. The proposed measures should be practical for use both by transplantation and nontransplantation medical providers, adaptable for use in adults and in children, and focused on the most important chronic GVHD manifestations. The measures should also give preference to quantitative, rather than semiquantitative, measures; capture information regarding signs, symptoms, and function separately from each other; and use validated scales whenever possible to demonstrate improved patient outcomes and meet requirements for regulatory approval of novel agents. Based on these criteria, we propose a set of measures to be considered for use in clinical trials, and forms for data collection are provided (http://www.asbmt.org/GvHDForms). Measures should be made at 3-month intervals and whenever major changes are made in treatment. Provisional definitions of complete response, partial response, and progression are proposed for each organ and for overall outcomes. The proposed response criteria are based on current expert consensus

opinion and are intended to improve consistency in the conduct and reporting of chronic GVHD trials, but their use remains to be demonstrated in practice.

KEY WORDS
Chronic graft-versus-host disease • Allogeneic cell transplantation • Response criteria • Consensus

INTRODUCTION

Overall survival or survival to permanent resolution of chronic graft-versus-host disease (GVHD) and discontinuation of systemic immunosuppression are long-term clinical outcomes that are accepted major end points in chronic GVHD clinical trials [1-3], but these long-term outcomes are not suitable for early-phase studies. Qualitative assessments of chronic GVHD manifestations can guide clinical decisions but are not adequate for measuring outcomes in clinical trials. To accelerate development of novel therapeutic agents in chronic GVHD, quantitative research tools are needed to measure short-term responses to treatment and to predict long-term clinical benefit.

The lack of standardized quantitative response criteria poses one of the major obstacles in pursuing therapeutic trials for chronic GVHD [4]. No generally accepted, much less validated, quantitative criteria for organ-specific or overall responses have been developed previously. The definitions of response typically used in previous studies have been global and qualitative in nature, with considerable variability from one study to the next (extensively reviewed by Gorgun Akpek in Attachment 1 at http://www.asbmt.org/GvHDForms). In addition, methods have not been developed to account for the distinction between reversible disease activity and irreversible damage.

Because no currently available database has information from patients with chronic GVHD at a sufficient level of detail, retrospective methods could not be used to identify clinical characteristics that are sensitive to change and predictive for major outcomes. The Working Group began by reviewing instruments currently used by hematopoietic stem cell transplantation physicians at Johns Hopkins, Children's Oncology Group, Fred Hutchinson Cancer Research Center, Harvard University, University of Minnesota, and National Institutes of Health. The Working Group also included specialists from other fields, including rheumatology and gastroenterology, to benefit from their experiences in developing and using chronic disease activity indices and response criteria in clinical trials [5-8].

This document is based on a broad consensus of experts and on the use of the best available data. These 2005 recommendations are intended to advance standards of chronic GVHD therapeutic trials, but they remain provisional and will need to be validated and refined according to data emerging from prospective studies. The Working Group could not entirely resolve certain intrinsic tensions between divergent goals. On the one hand, the assessments should be as simple as possible to facilitate their use by clinicians outside the field of hematopoietic cell transplantation, but on the other hand, the assessments should contain as much information as possible to support research. The former goal would require immediate item reduction and enforcement of consistency based on expert opinion, whereas the latter goal would encourage further exploration, with deferral of item reduction until data are available. For certain organs, the Working Group could not identify quantitative measures that would be suitable for use in clinical trials, even though qualitative assessments can be used for clinical management. In the end, the Working Group proposed a broad set of assessment measures that should be feasible in most academic settings, although some simplification might be needed if the assessments are to be used by medical providers outside the field of hematopoietic cell transplantation.

The differences between this document and the Diagnosis and Staging document should be noted [9]. Although there is appearance of some overlap, characteristics that could help establish the diagnosis of chronic GVHD or to assess the severity of chronic GVHD at a single time point might not serve as the most appropriate or sensitive measures for chronic GVHD disease activity. Conversely, a sensitive measure of chronic GVHD response might not necessarily serve as an appropriate diagnostic and staging tool.

PURPOSE OF THIS DOCUMENT

This document summarizes proposed measures and criteria for assessing outcomes in clinical trials involving patients with chronic GVHD. The measures and criteria do not necessarily reflect practices that might apply to routine patient care or to trials with limited resources. The measures and response criteria were developed to meet certain requirements.

1. *The instruments should be easy to use by both transplantation and nontransplantation care providers and should be limited to testing methods that are available in the outpatient setting.*
2. *The criteria should be adaptable for use in adults and in children.*

3. *The instrument should focus on the most important and most common manifestations of chronic GVHD and should not be designed to characterize all possible clinical manifestations.*
4. *Development should focus on quantitative measures as much as possible.*
5. *Measurements of symptoms, signs, global ratings, function, quality of life, or performance status should be made separately, and scales with established psychometric characteristics and desirable measurement properties should be used whenever possible [10,11].*
6. *With appropriate refinements and reliability and validation assessments, these tools should be suitable for use in clinical trials where the goals are to improve patient outcomes or to obtain regulatory approval.*

The Working Group had 3 additional goals: (1) to propose provisional definitions of complete response, partial response, and disease progression for each organ and for overall response; (2) to suggest appropriate strategies for using short-term end points in therapeutic clinical trials; and (3) to outline future research directions.

SUMMARY OF RECOMMENDATIONS

1. Proposed chronic GVHD-specific core measures include:
   A. Clinician- or patient-assessed signs and symptoms.
   B. The chronic GVHD symptom scale by Lee et al [12].
   C. The clinician- or patient-reported global rating scales (Table 1) [12-14].

   To facilitate validation studies, continuous data should be recorded as such and should not be reduced to prespecified categories.
2. Proposed chronic GVHD nonspecific ancillary measures for adults include:
   A. Measurement of grip strength [15-17] and 2-minute walk time [18].
   B. Patient-reported Human Activity Profile (HAP) questionnaire [19].
   C. Clinician-assessed Karnofsky performance status.
   D. The SF-36 version 2 questionnaire [20,21] and FACT-BMT for quality-of-life assessments (Table 1) [22].

   The ancillary chronic GVHD nonspecific measures are optional and should not be used as primary end points in chronic GVHD trials.
3. Age-appropriate modifications of existing measures should be used and explored in children with chronic GVHD [23-29].
4. Definition of response involves a comparison of chronic GVHD activity at two different time points. Provisional definitions of complete response, partial response, and progression are offered for each organ and for overall outcomes. Simple forms to be used for clinician and patient assessments are provided in Appendices A and B at http://www.asbmt.org/GvHDForms (Forms A and B). In each specific trial, irreversible baseline organ damage may be defined initially and then excluded in response assessments.
5. Measures should be made at 3-month intervals and whenever a major change is made in treatment. Permanent discontinuation of systemic immunosuppressive treatment indicates a durable response.
6. Further assistance from subspecialists will be needed to develop organ- or site-specific measures that could improve the sensitivity of chronic GVHD assessments. Specific organ or site assessments discussed by the Working Group include the following:
   A. Skin: skin-specific scoring systems [30], durometer [30-32], biopsy [31], or imaging (ultrasound, magnetic resonance imaging) [33,34].
   B. Eyes: corneal staining grading [35], conjunctival grading [36], ocular surface disease index [37].

Table 1. *Proposed Measures for Assessing Responses in Chronic GVHD Trials*

Measure	Clinician Assessed	Patient Reported
I. Chronic GVHD-specific core measures		
Signs	**Organ-specific measures**	**N/A**
Symptoms	**Clinician-assessed symptoms**	**Patient-reported symptoms Lee symptom scale [12]**
Global rating	**Mild-moderate-severe [12]**	**Mild-moderate-severe [12]**
	0-10 severity scale [13]	**0-10 severity scale [13]**
	7-point change scale [14]	**7-point change scale [14]**
II. Chronic GVHD-nonspecific ancillary measures		
Function	**Grip strength [15-17]**	**HAP [19]**
	2-min walk time [18]	**ASK in children [23-25]**
Performance status	**Karnofsky or Lansky [26]**	
Quality of life		**SF-36v.2 [20,21] or**
		FACT-BMT [22] in adults
		CHRIs in children [27-29]

ASK indicates Activities Scale for Kids; GVHD, graft-versus-host disease; N/A, not applicable; HAP, Human Activity Profile; CHRIS, Child Health Ratings Inventories.

C. Oral: Oral Mucositis Rating Scale [38].
D. Vulvar-vaginal: organ-specific staging [39,40].
E. Function: range of motion, limb volume, fatigue severity scale [41-43].

PROPOSED MEASURES OF CHRONIC GVHD RESPONSE ASSESSMENTS

The Working Group distinguished between chronic GVHD-specific core measures that directly measure organ-specific manifestations of chronic GVHD and nonspecific ancillary measures, which could reflect the overall impact of chronic GVHD and other illness on functioning or quality of life (Table 1). In future studies, these measures should be evaluated for construct validity (for Glossary see Attachment 2 at: http://www.asbmt.org/GvHDForms) and potential item reduction. In a feasibility study, 8 clinicians who had never previously used the assessment forms evaluated 4 adults with chronic GVHD [44]. The median time for each clinician evaluation was 36 minutes, and the median time needed to complete the panel of patient self-report items was 14 minutes. Results of this evaluation offered preliminary evidence of reliability, feasibility, and acceptability of the newly proposed measures.

PROPOSED CLINICIAN-ASSESSED AND PATIENT-REPORTED CHRONIC GVHD-SPECIFIC MEASURES

The following sections describe the recommended clinician-assessed and patient-reported chronic GVHD-specific measures (Table 2). Specific pediatric considerations for such situations are highlighted where appropriate. For the assessment of symptoms in younger children, depending on the child's development, assistance can be provided by the health care provider or the parent. The Working Group also recommends formal in-person training for all assessments to minimize intraobserver and interobserver variability. Instructional manual and slide set to assist with such training are available at http://www.asbmt.org/GvHDForms.

Organ-specific Assessments

Skin and skin appendages. Skin is the most frequently affected organ in chronic GVHD, and manifestations are highly variable. Skin assessments are structured to reflect 4 anatomic levels of skin involvement: (1) erythematous rash (epidermal involvement); (2) movable sclerosis (dermal involvement); (3) nonmoveable sclerosis, hidebound skin, or involvement of

Table 2. *Proposed Clinician-Assessed and Patient-Reported Chronic GVHD-Specific Measures*

Component	Items Assessed	Measure	Assessor
Skin	Erythematous rash of any sort	% Body surface area	C
	Movable sclerosis	0%-100% For each feature	C
	Nonmoveable sclerosis or subcutaneous sclerosis/fasciitis	By using rule of nines	C
	Ulcers	Largest dimension (cm) of the largest ulcer	C
	Pruritus or itching	0-10 Scale	P
Eyes	Bilateral Schirmer's tear test scores without anesthesia	Mean of both eyes, mm	C
	Main ocular symptom at the time of the visit	0-10 Scale	P
Mouth	Erythema	Total score 0-15	C
	Lichen-type hyperkeratosis		C
	Ulcerations		C
	Mucoceles		C
	Symptoms of oral pain, dryness, sensitivity	0-10 Scale	P
Hematology	Platelet count	Number/μL	C
	Eosinophils	Percent	C
GI	Upper GI symptoms	0-3 Score	C
	Esophageal symptoms	0-3 Score	C
	Diarrhea	0-3 Score	C
Liver	Total serum bilirubin	mg/dL	C
	ALT, alkaline phosphatase	U/L	C
Lungs	Bronchiolitis obliterans syndrome	FEV_1, DLCO	C
Chronic GVHD symptom scale [12]	30 items, 7 subscales, 1 summary scale	0-100	P
Global activity rating	Severity of chronic GVHD symptoms	0-10	C/P
	Perception of change	+3 to −3	C/P
	Overall severity of chronic GVHD	Mild – moderate-severe	C/P

ALT indicate alanine aminotransferase; C, assessed by the clinician; DLCO, diffusion lung capacity for carbon monoxide; FEV_1, forced expiratory volume in the first second; GI, gastrointestinal; GVHD, graft-versus-host disease; P, reported by the patient.

Vulvar-vaginal symptoms (yes or no) and patient weight should be recorded at each visit.

Range of motion of the most affected joints should be recorded depending on the availability of a physical therapist.

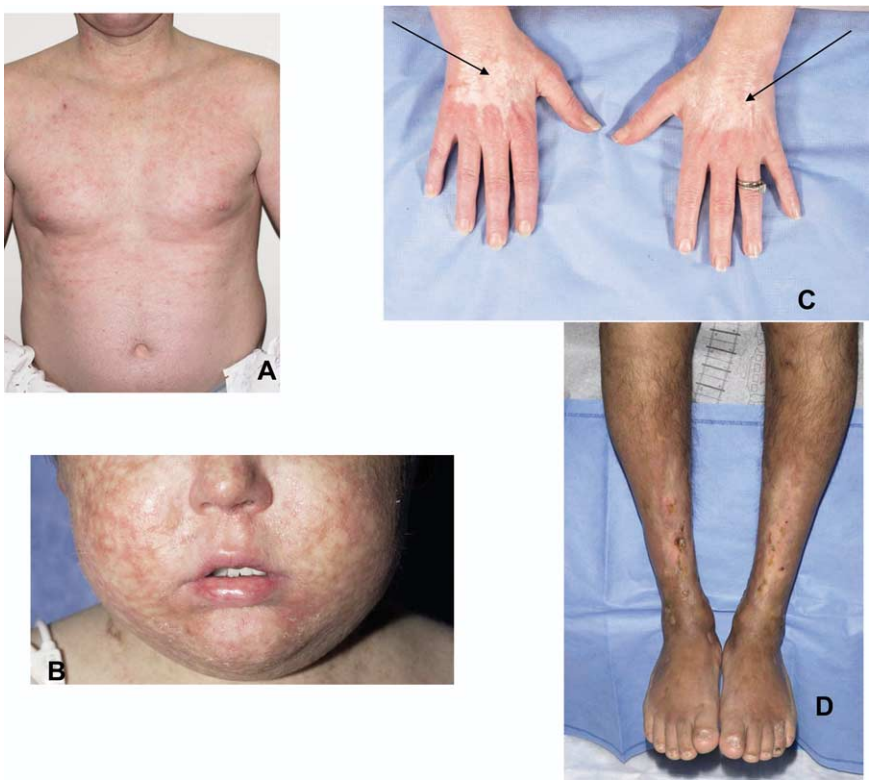

Figure 1. Skin manifestations assessed for response in chronic GVHD. A, Erythematous papular rash. B, Erythematous rash with papules and small scaly plaques. C, Dermal sclerosis. Skin is thickened, with decreased mobility to pinching but without adherence to underlying tissues. D, Subcutaneous sclerosis. Skin is hidebound, fixed to underlying tissues and cannot be pinched. Ulcers are present.

subcutaneous tissue and fascia (subcutaneous involvement); and (4) ulceration (full thickness loss of epidermal tissue) (Figure 1). Abnormalities for the first 3 points are each assessed separately according to the percent of body surface area (BSA) involved as estimated by the rule of nines for adults. A worksheet for recording the BSA involved for each of 8 skin regions is provided at: http://www.asbmt.org/GvHDForms (Attachment 3). Ulcer size is assessed by measuring the largest diameter of the largest ulcer.

The term "erythematous rash of any sort" is used as an inclusive reference to the many superficial skin eruptions of chronic cutaneous GVHD including papular, lichen planus-like, papulosquamous, poikiloderma, and keratosis pilaris-like rashes. The term "lichenoid" is not used, because this is a histopathologic diagnosis, not a clinical descriptive term.

Likewise, the term "sclerosis" or "sclerotic" is used to represent the general category of cutaneous GVHD findings associated with skin fibrosis, and to avoid confusion with the autoimmune disorder scleroderma. Superficial sclerosis (moveable) includes both lichen sclerosus-like and morphea-like lesions. Deep sclerosis includes diffuse, immovable (hidebound) sclerosis involving a wide area of skin, fibrosis of subcutaneous fat septae (rippling), and fasciitis (groove sign). Sclerotic skin manifestations may be as variable as the superficial form of the disease and are difficult to measure reliably. Sclerotic changes respond slowly to therapy and progression or regression of sclerotic lesions ideally should be assessed not only according to the total surface area involved but also according to the depth of involvement at any given site.

Because quantitative methods to measure the depth of sclerotic involvement are not available in a general oncology practice, these changes have been described in more qualitative terms related to thickening, pliability, adherence to underlying tissues, or changes in joint mobility. No validated scale exists for assessing sclerotic skin changes of chronic GVHD. Measures such as the Rodnan score for assessment of systemic sclerosis might be helpful for clinical evaluation of chronic GVHD, but this scale does not measure lichen sclerosus-like changes, subcutaneous involvement without overlying skin thickening, or fascial involvement. For this reason, the Rodnan score is not suitable for use in clinical trials. More sophisticated skin-specific scores are being developed for use by trained assessors in selected clinical trials (R. Knobler, MD, and H. Greinix, MD, oral communication, December 2005). There is an urgent need for the development of more quantifiable and reproducible measurements or imaging methods that could be used in patients with sclerotic skin manifestations of chronic GVHD [30-34].

Pigmentary changes do not indicate activity in chronic GVHD disease per se. Moreover, changes in pigmentation occur gradually and are perceptible only across long time intervals. Nonetheless, these changes should be recorded in the assessment forms, as described in the Diagnosis and Staging document [9], because they indicate the extent of previous skin involvement. Individuals who assess chronic GVHD of the skin should consult a picture atlas that is available for training and standardization (http://www.asbmt.org/GvHDForms).

The patient symptom intensity self-report profile includes the most severe itching during the past week, rated according to a 1-to-10 scale, because itching is the most frequent cutaneous symptom of chronic GVHD.

The rule of nines as an estimate of BSA involvement is intended for use in adults and is less accurate in children, particularly young children. For the sake of simplicity, we recommend using the rule of nines for all children, except for those younger than 1 year. A BSA grid for children younger than 1 year can be found at: http://www.asbmt.org/GvHDForms (Attachment 4).

Eyes. Dry eyes reflect either lacrimal dysfunction or destruction. The primary measure of lacrimal gland function in chronic GVHD is the Schirmer's test (to be performed without anesthesia) for each eye separately, as recommended by the Sjögren's syndrome consensus group [45]. Objective improvement would not be expected in cases where dry eyes and abnormal Schirmer's test result from complete lacrimal destruction. Instructions for administration of the Schirmer's test are provided with the instructional manual at: http://www.asbmt.org/GvHDForms.

The patient symptom intensity self-report profile includes the chief eye complaint rated according to a 1-to-10 scale for peak severity during the past week. The complaint can change from visit to visit, but only one chief eye complaint is graded. This method is simple to use but may impose undesirable limitations in patients with multiple complaints. In addition, ocular symptoms in patients with chronic GVHD can have causes other than chronic GHVD.

Schirmer's test without anesthesia is not recommended for children younger than 9 years, and evaluation by an ophthalmologist may be needed for objective scoring in younger children.

Mouth. Mouth assessments are conducted by using the newly proposed modification of the Schubert Oral Mucositis Rating Scale that scores oral surfaces from 0 to 15, with higher scores indicating more severe involvement. The 4 chronic GVHD manifestations assessed in this scale include: (1) mucosal erythema (0-3) grading based on the color intensity; (2) lichen-type hyperkeratosis (percent of oral surface area); (3) ulcerations (percent of oral surface area); and (4) presence of mucoceles (total number) (Figure 2). Instructions for these assessments and a photo dictionary are provided in the instructional manual on the World Wide Web: http://www.asbmt.org/GvHDForms.

The patient self-report symptom intensity profile includes dry mouth (subjective decrease in oral wetness), mouth pain in the absence of stimulation, and mouth sensitivity (irritation resulting form normally tolerated spices, foods, liquids, or flavors), each rated according to a 1-to-10 scale for peak severity during the past week.

Hematopoietic. Parameters to be evaluated for response assessments are absolute platelet count [46] and absolute eosinophil count [47]. Total white count and percent eosinophils are also recorded on the form at the time of the clinic visit.

Gastrointestinal tract. Gastrointestinal (GI) symptoms are difficult to measure in the outpatient setting. For this reason, GI symptoms during the preceding week are graded not through patient self-report forms but through interview by the examining clinician according to 0-to-3 severity scales. For upper GI symptoms of early satiety, anorexia, nausea, and vomiting, a score of 1 indicates mild, occasional symptoms, with little reduction in oral intake. A score of 2 indicates moderate, intermittent symptoms, with some reduction in oral intake, and a score of 3 indicates more severe or persistent symptoms throughout the day, with marked reduction in oral intake on most days. For esophageal symptoms of dysphagia or odynophagia, a score of 1 indicates occasionally difficult or painful swallowing of solid foods or pills. A score of 2 indicates intermittent dysphagia or odynophagia with solid foods and pills, but not for liquids or soft foods, and a score of 3 indicates dysphagia or odynophagia for almost all oral intakes on most days. Finally, for lower GI symptoms, a score of 1 indicates occasional loose or liquid stools, on some days. A score of 2 indicates intermittent loose or liquid stools throughout the day without requiring intervention to prevent or correct volume depletion, and a score of 3 indicates voluminous diarrhea requiring intervention to prevent or correct volume depletion.

Patients with chronic GVHD often have weight loss that is not always explained by GI symptoms [48]. Although the exact relationship between weight loss and chronic GVHD activity is not clear, patient weight should be recorded at each scheduled evaluation, given the simplicity of this measure and its potential importance for monitoring the success of therapy.

Liver. Liver injury should be assessed according to the most recent laboratory results for total serum bilirubin (mg/dL), alanine aminotransferase (U/L), and alkaline phosphatase (U/L). Laboratory upper limits of normal should also be recorded.

Lung. Measures that can be used to evaluate the response of bronchiolitis obliterans syndrome (BOS)

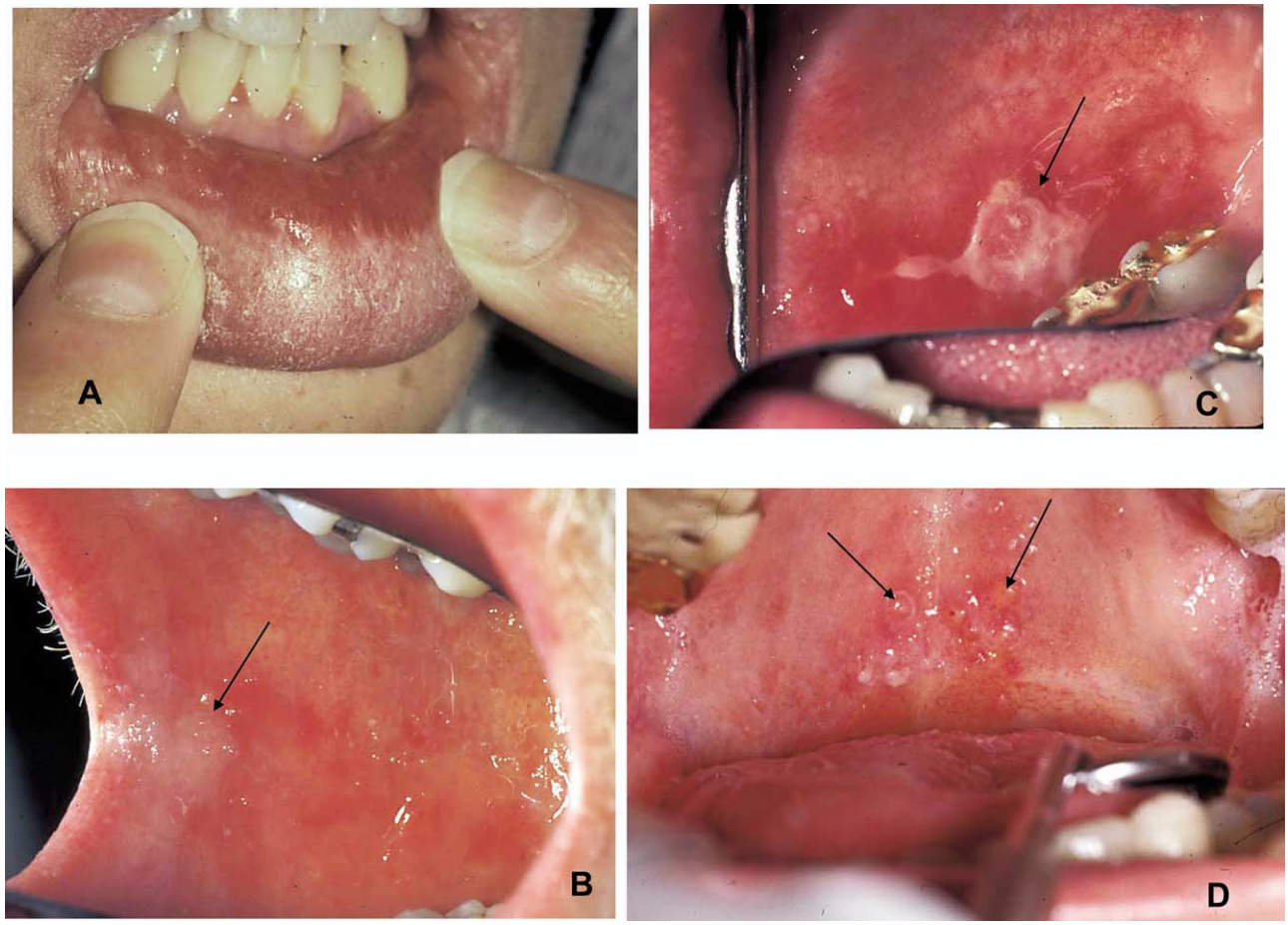

Figure 2. Oral manifestations assessed for response in chronic GVHD. A, Moderate erythema of vermilion lip. Labial mucosa shows severe erythema. B, Area of sheetlike lichenoid hyperkeratosis is present inside commissure. C, Ulcer with pseudomembranous fibrin exudates surrounded by severe erythema. D, Numerous vesicle-like mucoceles are seen at center of the palate, with patches of lichenoid hyperkeratosis and moderate erythematous changes.

after therapy are forced expiratory volume in the first second (FEV_1) and single breath diffusion lung capacity for carbon monoxide (DLCO) adjusted for hemoglobin, both of which are included in standard pulmonary function testing [49]. These two parameters are also included as components of the lung function score (LFS) that was recently developed as a predictor of respiratory failure and mortality after allogeneic hematopoietic stem cell transplantation [50]. A modified LFS is proposed as a simple measure of changes in the lung function in patients with BOS (see Table 3). Pulmonary function tests should be performed in children who are older than 5 years.

The LFS is computed according to FEV_1 and DLCO measurements compromise (>80% of predicted = 1, 70%-79% = 2, 60%-69% = 3, 50%-59% = 4, 40%-49% = 5, <40% = 6). The scores for FEV_1 and DLCO are then added together, and the sum is reduced to an overall category according to Table 3.

It is important to emphasize that the LFS has never been used in chronic GVHD response assessments, and its exact role in this setting needs to be determined. To allow validation in trials, absolute values of both FEV_1 and DLCO should be recorded on the data collection forms.

Vulva and vagina. Women should be asked specific questions relating to vulvar and vaginal symptoms, such as burning, pain, discomfort, or dyspareunia. Patients who report problems should be referred to a gynecologist. Because such symptoms could be caused by conditions other than chronic GVHD, and because proper evaluation requires a specialist examination, this information should be recorded but not scored for response assessment. Academic gynecologists interested in chronic GVHD are developing precise vulvovaginal assessment scales. These scales will be useful

Table 3. *Categories of the Lung Function Score*

Category	Lung Function	LFS
I	Normal	2
II	Mild decrease	3-5
III	Moderate decrease	6-9
IV	Severe decrease	10-12

in selected trials where vulvar and vaginal changes are the primary end points of interest [39,40].

Musculoskeletal connective tissue. Active-assisted range of joint motion could potentially serve as a very useful objective measure of chronic GVHD tissue response in patients with sclerotic changes involving large joints or the trunk. The main limitation of this tool, however, is the need for an adequately trained professional (usually a physical therapist) who can conduct the range-of-motion measurements in a standardized and reproducible fashion. If such a trained professional is available, pertinent range-of-motion measurements should be recorded sequentially, and for this purpose, trained clinicians should also be able to make serial measurements of selected sentinel joints for routine assessment purposes. Normal levels are available for adults and for children [51].

Chronic GVHD Symptoms

Lee et al [12] developed a symptom scale designed for individuals with chronic GVHD. The questionnaire asks respondents to indicate the degree of bother that they experienced during the past 4 weeks as a result of symptoms in 7 domains potentially affected by chronic GVHD (skin, eyes and mouth, breathing, eating and digestion, muscles and joints, energy, emotional distress). Published evidence supports its validity, reliability, and sensitivity to chronic GVHD severity. Items in this symptom scale can be reported in approximately 5 minutes.

The Lee chronic GVHD symptom scale has been tested only in individuals older than 18 years. Given its face validity and other desirable properties, however, this scale could be used for assessment of chronic GVHD in pediatric patients using either child or parent report, after appropriate modification and psychometric evaluation [52]. Information for the chronic GVHD symptom scale could be obtained by self-report from adolescents older than 12 years. For children who are 8 to 12 years of age, data should be obtained with the assistance of parents and the health care provider.

The Lee scale measures symptom bother as distinguished from symptom intensity, which is reported on the forms in Appendix B [53]. The degree to which patients report that they are bothered by a symptom represents a global assessment incorporating not only the intensity of the symptom and its frequency, but also the degree to which it causes emotional disturbance or interferes with functioning. The Lee scale complements the information regarding the intensity of chronic GVHD symptoms. For example, oral sensitivity may be severe, but patients may report that they are not bothered or distressed by this symptom. By contrast, skin itching may not be very intense or frequent but may cause great distress. Research is needed to determine the relationships between symptom intensity, frequency, and distress or bother in patients with chronic GVHD and to examine the degree to which these are distinct dimensions of the symptom experience.

Clinician- and Patient-Reported Global Ratings

Clinician perceptions. Physicians, nurse practitioners, or physician assistants should provide an assessment of current overall chronic GVHD severity on a 4-point scale (none, mild, moderate, severe) [12] and they can also provide an assessment of current overall chronic GVHD severity on an 11-point numeric scale (0 indicates no GVHD manifestations; 10 indicates most severe chronic GVHD symptoms possible). The categories of mild, moderate, and severe have been used in previous studies for patient and clinician assessment, where they were undefined but showed good prognostic characteristics [12,54]. Clinicians should also provide their assessments of patient chronic GVHD changes during the past month scored on a 7-point scale (very much better, moderately better, a little better, about the same, a little worse, moderately worse, very much worse) [14].

Patient perceptions. Similarly, at each patient self-assessment, patients should score their perceptions of overall chronic GVHD severity, overall severity of symptoms, and change in symptom severity compared with 1 month ago, using the same response options used by clinicians.

The exact role of global scales in chronic GVHD response assessments and their appropriate use as outcome measures in clinical trials remains to be determined. These scales could be sensitive to qualitative changes that might otherwise escape detection if the assessments were limited to quantitative measures.

PROPOSED CHRONIC GVHD NONSPECIFIC MEASURES

Nonspecific measures of function and patient-reported outcomes related to functional status and health-related quality of life could potentially offer additive objective and subjective data regarding the effects of chronic GVHD and its therapy. The GVHD nonspecific measures listed for consideration in Table 1 assess different dimensions of the patient experience. Selection of these instruments was based on the credibility and relevancy of their measurement properties (reliability, validity, responsiveness) and the availability of normative data to facilitate interpretation. Instruments that use self-report methods as opposed to interview-assisted reporting will promote feasibility in clinical trials, and the number of instruments was circumscribed to limit the burden on respondents. Consideration was also given to the availability of detailed instructions, procedure manuals,

coding algorithms and scoring systems, and background information regarding the conceptual and measurement properties of the instrument. The potential role of these nonspecific measures as outcomes in chronic GVHD therapeutic clinical trials needs to be determined in future research.

Functional Status

For an extremely complex multisystem disease such as chronic GVHD, objective measures of physical performance and patient-reported measures of functional status could represent important surrogate outcomes that might be more informative than the measures described above for assessing outcome in some situations (eg, advanced skin sclerosis). At the very least, measures of functional status can provide corroborative evidence of important changes after therapy. In other patient populations with chronic diseases [55-57], such outcomes have been extensively applied, and population norms for both physical performance measures and self-reported functional status are available. Because the use of functional end points in chronic GVHD assessment has not been extensively tested, and because these measures do not directly assess chronic GVHD manifestations, functional status outcomes can be used only as optional secondary end points in chronic GVHD trials until further information in available.

Proposed objective measures of physical performance include grip strength [15-17] measured using a hydraulic dynamometer (measured in pounds of pressure) and the 2-minute walk distance (measured as total distance in feet walked in 2 minutes) [18]. Although the measurement properties for the 2-minute walk distance have been less thoroughly examined than those of the 6-minute walk distance, the 2-minute walk may be a more feasible and efficient measure of performance in patients with chronic GVHD. Studies support the construct validity and responsiveness to change characteristics of the 2-minute walk distance [58,59]. Age-matched norms for walk time and grip strength are available for adults and for children. These simple instruments might not be available in the typical oncology clinic, but they can be obtained from rehabilitation medicine departments or purchased (eg, at: http://www.rehaboutlet.com/grip_hand_dynamometer.htm).

HAP. Recommended patient-reported measures of functional status include the HAP questionnaire (for adults) and the Activities Scale for Kids questionnaire (for children age 5-15 years) [19,23-25]. The HAP is a measure of physical activity. The 94 questions are ranked hierarchically in ascending order according to the metabolic equivalents of oxygen consumption required to perform each activity [19]. The HAP, therefore, provides a survey of the activities the patient performs independently across a wide range of metabolic demand, beginning with getting out of bed, bathing, dressing, walking using public transit, performing a series of progressively more physically demanding household chores, and ending with running or jogging 3 miles in 30 minutes or less. The recommended corollary instrument to measure self-reported function in children is the Activities Scale for Kids [23-25].

Performance scales. The Karnofsky Performance Scale is commonly used in clinical assessments of chronic GVHD and has prognostic value for survival [60]. Whether a clinician assessment that combines performance, health status, and impairment is a valid, reliable, or sensitive tool to gauge response after therapy for chronic GVHD remains to be determined. Performance scores should nonetheless be recorded as part of each assessment. Lansky Play Performance Scale scores should be recorded for children younger than 16 years [26].

Self-Reported Health-Related Quality of Life

The effects of chronic GVHD and its treatment on general physical and emotional health and quality of life are other patient-reported outcomes that may be responsive to change as a result of chronic GVHD therapy [61]. The Medical Outcomes Study Short Form 36-item Questionnaire version 2* is a measure that has had wide application and is well accepted as measure of self-reported general health and the degree to which health impairments interfere with activities of daily living and role function [21,62]. The Functional Assessment of Chronic Illness Therapy is an oncology-specific quality-of-life instrument that has well-developed psychometric properties, and population norms for healthy individuals and those with both mild and more severe chronic illnesses. An additional 18-item disease-specific module evaluates concerns common to patients who have had stem cell transplantation (FACT-BMT)* [22]. These instruments are appropriate for patients older than 18 years. In pediatric patients, the Child Health Ratings Inventories* generic core and Disease-Specific Impairment Inventory-HSCT*, a hematopoietic cell transplantation-specific module, could serve as a surrogate for FACT-BMT [27-29].

Cross-sectional studies have shown that chronic GVHD has an adverse effect on quality of life [63], but the role of quality of life as a measure of response to therapy or as a predictor of long-term outcome remains to be defined. Patient-reported quality-of-life measures cannot replace quantitative measures of chronic GVHD activity in clinical trials. Patient-reported items should be selected to address specific questions and should have relevance for chronic GVHD. Each instrument should be considered not

only for the information that it might provide in its own right but also for the information that it might add in the context of other instruments to be used in assessments. Hence, investigators should be aware of similarities and differences between instruments when making decisions about their use in clinical trials. Investigators should take care not to impose an excessive burden of self-report items on those who are participating in clinical trials. A table comparing above-discussed chronic GVHD-specific and the optional patient-reported nonspecific measures is provided at: http://www.asbmt.org/GvHDForms (Attachment 5). The recommendation to use these instruments does not imply permission for their use in clinical trials. Investigators should follow the procedure established by the organizations that hold copyright for each instrument (see Attachment 5).

CHRONIC GVHD DATA COLLECTION FORMS

Appendices A and B (http://www.asbmt.org/GvHDForms [Forms A and B]) show data collection forms for the recommended clinician-assessed and patient-reported measures. In clinical trials, data should be submitted to the study coordinating center for further calculations, processing, and interpretation of responses. It is not necessary to include recommended measures in every trial, and judgment must be used in deciding which items will best suit the needs of each study. In all studies, the measures to be made and the timing of the measures must be specified.

PROVISIONAL CRITERIA FOR DEFINITION OF RESPONSE

Protocols must specify the times when response will be assessed, and the requirement for durability of response (see forthcoming Design of Clinical Trials Working Group report). Permanent discontinuation of systemic immunosuppressive treatment indicates a durable response.

Certain changes such as dry eyes, esophageal stricture, bronchiolitis obliterans, or advanced sclerotic skin lesions may be considered irreversible and may be excluded from consideration for assessments of complete or partial response, if specified by the protocol.

To assess response, disease manifestations at two different time points must be compared, and a judgment must be made as to whether the magnitude of any change qualifies as clinical improvement or clinical deterioration. The magnitude of change required for clinical improvement or deterioration should reflect genuine clinical meaning, and the criteria should be developed and standardized as much as possible. This standardization may be relatively easy to establish for manifestations that can be measured quantitatively with little day-to-day variation but will be more difficult to establish for manifestations that can be measured only in more qualitative ways.

The statistician should be always be included early in the development of the trial design and should help to select the analyses that best fit the types of measures being collected. Because no criteria for defining meaningful improvement or clinical benefit have been validated for measures of chronic GVHD, the results of trials should include both the categorical outcomes defined below and the average change from baseline for each parametric measure. Protocols should specify whether change is to be calculated according to percent of full scale or percent of baseline. Analysis of percent changes is particularly needed for the interpretation of smaller early drug-development trials.

Pending appropriate validation studies, the Working Group proposes the following consensus definitions of complete response, partial response, and progression. The complete and partial response categories apply only to organs that have measurable and reversible GVHD-related abnormalities at baseline. For certain organs and measures, however, chronic GVHD sequelae can reflect damage that is not reversible. Some obvious examples of this problem are chronic dry eyes, esophageal stricture, bronchiolitis obliterans, or advanced skin sclerosis or contractures. For these manifestations, the category of complete organ response may not apply if protocols prespecify any such exclusion. The progression category applies to all organs.

Objective Measures of GVHD Activity

Complete organ response. The term "complete organ response" indicates resolution of all reversible manifestations related to chronic GVHD in a specific organ.

Partial organ response. The proposed general guideline for defining partial response in specific organ requires at least 50% improvement in the scale used to measure disease manifestations related to chronic GVHD. This guideline was selected as unequivocally indicating genuine clinical benefit. The criterion of 50% improvement requires some adjustment in cases where the extent of abnormality at the baseline measurement is low. For example, there would be no question that a 50% decrease in rash from 80% of BSA to 40% represents genuine clinical improvement. On the other hand, the same 50% decrease from 5% of BSA to 2.5% would represent a much less compelling clinical improvement. For this reason, when the extent of abnormality at the baseline measurement is lower than the midpoint on the scale, the minimum criterion for response should be defined as percentage (eg, 25%) of the full scale as opposed to a percentage of the starting value. To be consistent, if the extent of abnormality at the baseline measurement

is lower than the minimum percent of full-scale change needed to define a partial response (eg, 25% of the full scale), then the only possible response would be a complete response.

Organ progression. Criteria for progression in each organ must be defined, because the overall category of partial response requires the absence of progression in any organ (see below). For an organ affected by chronic GVHD at the baseline evaluation, the proposed general guideline for defining progression specifies an absolute increase of at least 25% in the scale used to measure disease manifestations related to chronic GVHD. Progression cannot be scored for manifestations with baseline values that are within 25% of the full-scale value. When baseline measures of chronic GVHD severity are 50% to 75% of full scale at baseline, the criteria for improvement require more than a 50% change from baseline (which produces more than a 25% of full-scale change), whereas a 25% of full-scale change is sufficient for progression. This asymmetry in the minimal criteria for improvement and progression is intended to ensure a high level of confidence that any improvement is clinically meaningful and to ensure early detection of any deterioration.

Proposed guidelines for calculating partial response and progression and instructions for use by study coordinating centers are available on the World Wide Web at: http://www.asbmt.org/GvHDForms.htm (Appendices C and D). The criteria proposed in these guidelines are admittedly arbitrary, because in most cases, they have never been validated for patients with chronic GVHD, and the distribution of baseline scores is unknown. For these reasons, the proposed criteria are provisional and subject to change with further clinical experience. Also, depending on the stringency of response definitions required by the specific study, these general guidelines could be modified to fit the needs of a particular protocol. Because the criteria are subject to change, we strongly recommend that data report forms should always record the actual numeric values for any measurement.

Limitations in measurement of organ responses. The response criteria in Appendix C do not account for qualitative changes. Clinical experience indicates that clinically important qualitative improvement often occurs before improvement in the measures summarized in Appendix C. For this reason, the response criteria in Appendix C should not be used as the primary guide for clinical decisions. Certain organs are not considered in Appendix C because quantitative assessments are not feasible. The response criteria also do not account for the prior trajectory of abnormalities. For example, stable disease might be considered a response when the prior trajectory was clear progression, as indicated, for example, by serial pulmonary function tests. Stable disease after prior improvement could not be considered a favorable outcome, and stable disease after prior stability cannot be considered a response.

Standardized response criteria for BOS associated with chronic GVHD have never been investigated. The hallmark of response to therapy for BOS is stabilization of lung function with no further decrease in FEV_1 during a 3-month period. A few cases of improved FEV_1 after therapy for BOS have been reported, but these outcomes could reflect disease misclassification or very early treatment.

Definitions of overall response. Three general overall categories of response are proposed: complete response, partial response, and other. Although the group recognizes the complete and partial responses as the categories of greatest interest, other summary outcomes such as stable disease or mixed response can be also included in clinical trials. Complete overall response is defined as resolution of all reversible manifestations in each organ or site, and partial overall response is defined as improvement in a measure for at least one organ or site without progression in measures for any other organ or site. We do not propose the routine use of the term "stable disease" because the interpretation depends too heavily on the prior trajectory of the disease, as discussed above.

Global Ratings, Patient-Reported Outcomes, and Performance Measures

The terms "complete response," "partial response," and "progression" do not technically apply to subjective or functional measures data. Instead, the definition of improvement or worsening for such scales is based on the reliability of the measure (the variability caused by measurement error) and is anchored against clinically perceptible changes. For global ratings and categorical scales, a 1-point change on a 3- or 7-point scale or a 2- to 3-point change (0.5 SD change) on a 0- to 10-point scale could be considered clinically meaningful, pending further evaluation in the chronic GVHD population. Unless otherwise specified, for all patient-reported measures, a change of 0.5 SD may be considered clinically meaningful [64,65]. A distribution-based analysis was used to define improvement as a change of 6 to 7 points (0.5 SD) on the chronic GVHD symptom summary scale [12].

Impairments of grip strength, walk time, and range of motion are measured by comparison with normative values. Minimal clinically meaningful improvements for these measures are provisionally defined as a 25% decrease in the level of impairment as compared with baseline. For HAP, clinically meaningful improvement is defined as a 10-point increase in the maximum activity score, because a change of this magnitude is sufficient to change the disability category at the middle of the scale.

USE OF RESPONSE ASSESSMENT AS A PRIMARY END POINT IN CLINICAL TRIALS

Beyond providing tools for assessment of response, clinical protocols must select appropriate primary and secondary end points. A primary end point represents the principal basis by which the success or failure of a treatment will be decided, whereas secondary end points are selected to be supportive of the primary end point or to demonstrate that the benefit provided with respect to the primary end point is not offset by a detrimental effect on other disease manifestations. Prespecified expectations regarding effects of a study intervention on the primary end point also provide the basis for statistical power calculations used to determine the number of patients to be enrolled. If a trial is going to be used for the marketing approval of therapy, regulatory authorities should be included early in the planning.

Table 4 summarizes the potential use of organ measures as primary end points in chronic GVHD clinical trials. Any of the listed assessments could be used as a secondary end point, with or without blinding, but the validity of subjective assessments in open-label trials will always be open to question. The list of assessments in this table is limited to measurements and scales that could be used by a general internist or pediatrician or by patients. More sophisticated assessments of certain organs such as skin, eyes, mouth, female genital tract, and joints may be needed for certain studies [30-40]. Specialized expertise will be needed for these assessments, and the criteria for measurement of response in these situations exceed the scope of the current proposal.

Some of the response scales in Table 4 measure clinical benefit, whereas others measure potential clinical benefit as reflected by a surrogate end point. For example, in cardiovascular disease, well-established surrogate end points such as blood pressure or serum cholesterol can be used for regulatory approval. Less well-established surrogate end points could be used in certain circumstances if they are reasonably likely to predict clinical benefit. Elevated serum bilirubin levels at the onset of chronic GVHD have been associated with an increased risk of nonrelapse mortality [1], but validation studies have not been carried out to show that improvement in serum bilirubin levels is associated with prolonged survival among patients with chronic GVHD. Evaluation of other liver function tests in patients with chronic GVHD has also not been reported. For this reason, the acceptability of improved liver function tests as a basis for approval remains uncertain at this time.

Some of the response scales in Table 4 involve objective assessments, whereas others involve subjective assessments. Blinding of treatment arms to prevent bias is recommended whenever feasible, especially when a subjective end point is used as a primary end point in a clinical trial. Even for objective assessments, blinding can be extremely helpful in preventing bias. For example, objective assessments of the skin and mouth can be enhanced through review of serial photographs by a panel of individuals as blinded assessors who have no other information about the patient. A similar approach could also be used in the evaluation of chronic GVHD involving the eye and female genital tract.

Table 4. *Potential Use of Chronic GVHD-specific Measures as Primary End Points in Clinical Trials*

Organ and Assessment	Clinical Benefit	Blinding Required
Skin		
Objective assessment	**Yes**	**No***
Pruritus	**Yes**	**Yes**
Eyes		
Schirmer's tear test	**Yes**	**No**
Ocular discomfort	**Yes**	**Yes**
Mouth		
Objective assessment	**Yes**	**No***
Oral pain	**Yes**	**Yes**
Oral dryness	**Yes**	**Yes**
Oral sensitivity	**Yes**	**Yes**
Hematology	**Unknown**	**No**
Gastrointestinal symptoms	**Yes**	**Yes**
Liver		
Bilirubin	**Unknown**	**No**
Alkaline phosphatase	**Unknown**	**No**
Aminotransferase levels	**Unknown**	**No**
Lungs	**Yes**	**No**
Symptom scale	**Yes**	**Yes**
Global rating scales	**Yes**	**Yes**
Range of motion	**Yes**	**No***

GVHD indicates graft-versus-host disease.

This table is limited to consideration of possible primary end points. Any of the listed assessments could be used as a secondary end point, with or without blinding.

*Objective assessments could be enhanced with the use of photographs and/or blinded assessor.

FUTURE DIRECTIONS

The proposed response criteria are expected to enhance uniformity of data collection methods and advance standards of chronic GVHD clinical trials but are only provisional and it is imperative that they be tested for reliability and validity in prospective studies. Important tasks for the immediate future include the determination of minimal clinically important changes for some of the measures proposed, determination of most relevant measures, reduction of items, and establishing an outcomes repository for data collected in clinical trials and natural history studies using these instruments. Collaborations with organ-site specialist should be strengthened to develop methods for more sensitive and objective assessment of specific organs. Future studies will be needed to determine the extent to which patient-reported outcomes and functional

measures could be used as a primary end point in chronic GVHD clinical trials. Improved methods will be needed to distinguish chronic GVHD disease activity from irreversible damage and to develop a chronic GVHD activity index for clinical trials, perhaps through the use of biomarkers [66].

NATIONAL INSTITUTES OF HEALTH CONSENSUS DEVELOPMENT PROJECT ON CRITERIA FOR CLINICAL TRIALS IN CHRONIC GVHD STEERING COMMITTEE

Steven Pavletic and Georgia Vogelsang (Project Chairs), LeeAnn Jensen (Planning Committee Chair), Lisa Filipovich (Diagnosis and Staging), Howard Shulman (Histopathology), Kirk Schultz (Biomarkers), Dan Couriel (Ancillary and Supportive Care), Stephanie Lee (Design of Clinical Trials), and James Ferrara, Mary Flowers, Jean Henslee-Downey, Paul Martin, Barbara Mittleman, Shiv Prasad, Donna Przepiorka, Douglas Rizzo, Daniel Weisdorf, and Roy Wu (Members). The project group also recognizes contributions of numerous colleagues in the field of blood and marrow transplantation, medical specialists and consultants, pharmaceutical industry, and the National Institutes of Health and US Food and Drug Administration professional staff for their intellectual input, dedication, and enthusiasm on the road toward completion of these documents.

DISCLAIMER

The opinions expressed here are those of the authors and do not represent the official position of the National Institutes of Health, Food and Drug Administration, or the US Government.

ACKNOWLEDGMENTS

This project was supported by the National Institutes of Health (NIH), National Cancer Institute, Office of the Director, Cancer Therapy Evaluation Program, Intramural Research Program, and Center for Cancer Research; National Heart Lung and Blood Institute, Division of Blood Diseases and Resources; Office of Rare Diseases, NIH, Office of the Director; National Institute of Allergy and Infectious Disease, Transplantation Immunology Branch; and the Health Resources and Services Administration, Division of Transplantation and the Naval Medical Research Center, C. W. Bill Young/Department of Defense Marrow Donor Recruitment and Research Program. The authors would also like to acknowledge the following individuals and organizations that by their participation made this project possible: American Society for Blood and Marrow Transplantation, Center for International Bone and Marrow Transplant Research, Blood and Marrow Transplant Clinical Trials Network, Canadian Blood and Marrow Transplant Group, European Group for Blood and Marrow Transplantation, Pediatric Blood and Marrow Transplant Consortium, and the representatives of the South American transplant centers (Drs Luis F. Bouzas and Vaneuza Funke). This project was conducted in coordination with the American Society for Clinical Oncology and American Society of Hematology (liaisons, Dr Michael Bishop and Mr Jeff Coughlin). The organizers are also indebted to patients and patient and research advocacy groups who made this process much more meaningful by their engagement. Special thanks also to Ms Paula Kim who coordinated these efforts.

LIST OF APPENDICES AND ATTACHMENTS AVAILABLE ON THE AMERICAN SOCIETY FOR BLOOD AND MARROW TRANSPLANTATION WORLD WIDE WEB SITE: http://www.asbmt.org/GvHDForms

Response Criteria Appendix A: Data Collection Form A–Clinican Assessment
Response Criteria Appendix B: Data Collection Form B–Patient Self-Report
Response Criteria Appendix C: Proposed Calculations for Partial Response in Chronic GVHD
Response Criteria Appendix D: Proposed Calculations for Progression in Chronic GVHD
Response Criteria Attachment 1: Literature Review of Various Response Criteria Used in Chronic Graft-versus-Host Disease Clinical Trials (By Gorgun Akpek)
Response Criteria Attachment 2: Glossary
Response Criteria Attachment 3: Skin BSA Calculation Worksheet
Response Criteria Attachment 4: BSA Assessment in Children Younger than 1 Year of Age
Response Criteria Attachment 5: Patient-Reported Outcome Measures Recommended for Chronic GVHD Response Evaluations

REFERENCES

1. Stewart BL, Storer B, Storek J, et al. Duration of immunosuppressive treatment for chronic graft-versus-host disease. *Blood*. 2004;104:3501-3506.
2. Pavletic SZ, Carter SL, Kernan NA, et al. Influence of T-cell depletion on chronic graft-versus-host disease: results of a multicenter randomized trial in unrelated marrow donor transplantation. *Blood*. 2005;106:3308-3313.
3. Koc S, Leisenring W, Flowers ME, et al. Therapy for chronic graft-versus-host disease: a randomized trial comparing cyclosporine plus prednisone versus prednisone alone. *Blood*. 2002; 100:48-51.
4. Farag SS. Chronic graft-versus-host disease: where do we go from here? *Bone Marrow Transplant*. 2004;33:569-577.
5. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity mea-

sures for rheumatoid arthritis clinical trials: the committee on outcome measures in rheumatoid arthritis clinical trials. *Arthritis Rheum*. 1993;36:729-740.
6. Ward MM, Marx AS, Barry NN. Comparison of the validity and sensitivity to change of 5 activity indices in systemic lupus erythematosus. *J Rheumatol*. 2000;27:664-670.
7. Sandborn WJ, Feagan BG, Hanauer SB, et al. A review of activity indices and efficacy endpoints for clinical trials of medical therapy in adults with Crohn's disease. *Gastroenterology*. 2002;122:512-530.
8. Rider LG, Giannini EH, Harris-Love M, et al. Defining clinical improvement in adult and juvenile myositis. *J Rheumatol*. 2003;30:603-617.
9. 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*. 2005; 11:945-956.
10. Acquadro C, Berzon R, Dubois D, et al. 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. *Value Health*. 2003; 6:522-531.
11. 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.
12. Lee S, Cook EF, Soiffer R, et al. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8:444-452.
13. Cleeland CS, Mendoza TR, Wang XS, et al. Assessing symptom distress in cancer patients: the MD Anderson symptom inventory. *Cancer*. 2000;89:1634-1646.
14. Osoba D, Rodrigues G, Myles J, et al. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol*. 1998;16:139-144.
15. Mathiowetz V, Kashman N, Volland G, et al. Grip and pinch strength: normative data for adults. *Arch Phys Med Rehabil*. 1985;66:69-74.
16. Mathiowetz V, Weber K, Volland G, et al. Reliability and validity of grip and pinch strength evaluations. *J Hand Surg [Am]*. 1984;9:222-226.
17. Mathiowetz V, Wiemer DM, Federman SM. Grip and pinch strength: norms for 6- to 19-year-olds. *Am J Occup Ther*. 1986; 40:705-711.
18. Waters RL, Lunsford BR, Perry J, et al. Energy-speed relationship of walking: standard tables. *J Orthop Res*. 1988;6:215-222.
19. Daughton DM, Fix AJ, Kass I, et al. Maximum oxygen consumption and the ADAPT quality-of-life scale. *Arch Phys Med Rehabil*. 1982;63:620-622.
20. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36), I: conceptual framework and item selection. *Med Care*. 1992;30:473-483.
21. Ware JE Jr. SF-36 health survey update. *Spine*. 2000;25:3130-3139.
22. McQuellon RP, Russell GB, Cella DF, et al. Quality of life measurement in bone marrow transplantation: development of the functional assessment of cancer therapy-bone marrow transplant (FACT-BMT) scale. *Bone Marrow Transplant*. 1997; 19:357-368.
23. Plint AC, Gaboury I, Owen J, et al. Activities scale for kids: an analysis of normals. *J Pediatr Orthop*. 2003;23:788-790.
24. Young NL, Williams JI, Yoshida KK, et al. Measurement properties of the activities scale for kids. *J Clin Epidemiol*. 2000;53:125-137.
25. Young NL, Yoshida KK, Williams JI, et al. The role of children in reporting their physical disability. *Arch Phys Med Rehabil*. 1995;76:913-918.
26. Lansky SB, List MA, Lansky LL, et al. The measurement of performance in childhood cancer patients. *Cancer*. 1987;60: 1651-1656.
27. Parsons SK, Barlow SE, Levy SL, et al. Health-related quality of life in pediatric bone marrow transplant survivors: according to whom? *Int J Cancer Suppl*. 1999;12:46-51.
28. Parsons SK, Shih MC, Duhamel KN, et al. Original research article: maternal perspectives on children's health-related quality of life during the first year after pediatric hematopoietic stem cell transplant. *J Pediatr Psychol*. 2005.
29. Parsons SK, Shih MC, Mayer DK, et al. Preliminary psychometric evaluation of the child health ratings inventory (CHRIs) and disease-specific impairment inventory-hematopoietic stem cell transplantation (DSII-HSCT) in parents and children. *Qual Life Res*. 2005;14:1613-1625.
30. Seyger MM, van den Hoogen FH, de Boo T, et al. Reliability of two methods to assess morphea: skin scoring and the use of a durometer. *J Am Acad Dermatol*. 1997;37:793-796.
31. Aghassi D, Monoson T, Braverman I. Reproducible measurements to quantify cutaneous involvement in scleroderma. *Arch Dermatol*. 1995;131:1160-1166.
32. Falanga V, Bucalo B. Use of a durometer to assess skin hardness. *J Am Acad Dermatol*. 1993;29:47-51.
33. Gottlober P, Leiter U, Friedrich W, et al. Chronic cutaneous sclerodermoid graft-versus-host disease: evaluation by 20-MHz sonography. *J Eur Acad Dermatol Venereol*. 2003;17:402-407.
34. Dumford K, Anderson JC. CT and MRI findings in scleroderma tous chronic graft vs host disease. *Clin Imaging*. 2001;25:138-140.
35. Bron AJ, Evans VE, Smith JA. Grading of corneal and conjunctival staining in the context of other dry eye tests. *Cornea*. 2003;22:640-650.
36. Robinson MR, Lee SS, Rubin BI, et al. Topical corticosteroid therapy for cicatricial conjunctivitis associated with chronic graft-versus-host disease. *Bone Marrow Transplant*. 2004;33: 1031-1035.
37. Schiffman RM, Christianson MD, Jacobsen G, et al. Reliability and validity of the ocular surface disease index. *Arch Ophthalmol*. 2000;118:615-621.
38. Schubert MM, Williams BE, Lloid ME, et al. Clinical assessment scale for the rating of oral mucosal changes associated with bone marrow transplantation: development of an oral mucositis index. *Cancer*. 1992;69:2469-2477.
39. Spinelli S, Chiodi S, Costantini S, et al. Female genital tract graft-versus-host disease following allogeneic bone marrow transplantation. *Haematologica*. 2003;88:1163-1168.
40. Stratton PTM, Turner M. Vulvar and vaginal graft versus host disease in women after hematopoietic stem cell transplantation. *J Soc Gynecol Investig*. 2004;11:162A.
41. Hann DM, Denniston MM, Baker F. Measurement of fatigue in cancer patients: further validation of the fatigue symptom inventory. *Qual Life Res*. 2000;9:847-854.
42. Buysse DJ, Reynolds CF III, Monk TH, et al. The Pittsburgh

sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res*. 1989;28:193-213.
43. Carpenter JS, Andrykowski MA. Psychometric evaluation of the Pittsburgh sleep quality index. *J Psychosom Res*. 1998;45:5-13.
44. Mitchell S, Reeve B, Cowen E, et al. Feasibility and reproducibility of the new NIH consensus criteria to evaluate response in chronic GVHD–a pilot study [abstract]. *Blood*. 2005;106(suppl 1):873a.
45. Vitali C, Bombardieri S, Jonsson R, et al. Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European consensus group. *Ann Rheum Dis*. 2002;61:554-558.
46. Lee SJ, Vogelsang G, Flowers ME. Chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2003;9:215-233.
47. Jacobsohn DA, Schechter T, Seshadri R, et al. Eosinophilia correlates with the presence or development of chronic graft-versus-host disease in children. *Transplantation*. 2004;77:1096-1100.
48. Jacobsohn DA, Margolis J, Doherty J, et al. Weight loss and malnutrition in patients with chronic graft-versus-host disease. *Bone Marrow Transplant*. 2002;29:231-236.
49. Chien JW, Madtes DK, Clark JG. Pulmonary function testing prior to hematopoietic stem cell transplantation. *Bone Marrow Transplant*. 2005;35:429-435.
50. Parimon T, Madtes DK, Au DH, et al. Pretransplant lung function, respiratory failure, and mortality after stem cell transplantation. *Am J Respir Crit Care Med*. 2005;172:384-390.
51. Norkin C, White D. *Measurement of Joint Motion*. Philadelphia: FA Davis Co; 1995.
52. Matza LS, Swensen AR, Flood EM, et al. Assessment of health-related quality of life in children: a review of conceptual, methodological, and regulatory issues. *Value Health*. 2004;7:79-92.
53. Goodell TT, Nail LM. Operationalizing symptom distress in adults with cancer: a literature synthesis. *Oncol Nurs Forum*. 2005;32:E42-E47.
54. Lee SJ, Klein JP, Barrett AJ, et al. Severity of chronic graft-versus-host disease: association with treatment-related mortality and relapse. *Blood*. 2002;100:406-414.
55. Nagashima M, Shu G, Yamamoto K, et al. The ability of disease modifying antirheumatic drugs to induce and maintain improvement in patients with rheumatoid arthritis: epidemiology of DMARDs treatment in Japan. *Clin Exp Rheumatol*. 2005;23:27-35.
56. Koller WC, Lyons KE, Truly W. Effect of levodopa treatment for parkinsonism in welders: a double-blind study. *Neurology*. 2004;62:730-733.
57. Craig J, Young CA, Ennis M, et al. A randomized controlled trial comparing rehabilitation against standard therapy in multiple sclerosis patients receiving intravenous steroid treatment. *J Neurol Neurosurg Psychiatry*. 2003;74:1225-1230.
58. Brooks D, Parsons J, Tran D, et al. The two-minute walk test as a measure of functional capacity in cardiac surgery patients. *Arch Phys Med Rehabil*. 2004;85:1525-1530.
59. Eiser N, Willsher D, Dore CJ. Reliability, repeatability and sensitivity to change of externally and self-paced walking tests in COPD patients. *Respir Med*. 2003;97:407-414.
60. Akpek G, Zahurak ML, Piantadosi S, et al. Development of a prognostic model for grading chronic graft-versus-host disease. *Blood*. 2001;97:1219-1226.
61. Wiklund I. Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life. *Fundam Clin Pharmacol*. 2004;18:351-363.
62. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health Qual Life Outcomes*. 2003;1:4.
63. Socie G, Stone JV, Wingard JR, et al. Long-term survival and late deaths after allogeneic bone marrow transplantation: late effects working committee of the international bone marrow transplant registry. *N Engl J Med*. 1999;341:14-21.
64. 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;41:582-592.
65. Norman GR, Sloan JA, Wyrwich KW. Is it simple or simplistic? *Med Care*. 2003;41:599-600.
66. Schultz K, Miklos D, Fowler D, et al. Towards biomarkers for chronic graft versus host disease. *Biol Blood Marrow Transplant*. In press.