CME article

Correlation between NIH composite skin score, patient-reported skin score, and outcome: results from the Chronic GVHD Consortium

David A. Jacobsohn,[1] Brenda F. Kurland,[2] Joseph Pidala,[3] Yoshihiro Inamoto,[2] Xiaoyu Chai,[2] Jeanne M. Palmer,[4] Sally Arai,[5] Mukta Arora,[6] Madan Jagasia,[7] Corey Cutler,[8] Daniel Weisdorf,[6] Paul J. Martin,[2] Steven Z. Pavletic,[9] Georgia Vogelsang,[10] Stephanie J. Lee,[2] and Mary E. D. Flowers[2]

[1]Division of Blood and Marrow Transplantation, Children's National Medical Center, Washington, DC; [2]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [3]Blood and Marrow Transplantation, Moffitt Cancer Center, Tampa, FL; [4]Division of Hematology/Oncology, Medical College of Wisconsin, Milwaukee, WI; [5]Division of Blood and Marrow Transplantation, Stanford University Medical Center, Stanford, CA; [6]Blood and Marrow Transplant Program, University of Minnesota, Minneapolis, MN; [7]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [8]Hematologic Malignancies, Dana-Farber Cancer Institute, Boston, MA; [9]National Cancer Institute, National Institutes of Health, Bethesda, MD; and [10]Department of Oncology, Johns Hopkins Hospital, Baltimore, MD

There are no validated criteria to measure skin response in chronic GVHD. In a prospectively assembled, multicenter cohort of patients with chronic GVHD (N = 458), we looked for correlation of change in several different scales recommended by the National Institutes of Health (NIH) Consensus with clinician and patient perception of change and overall survival. Of the clinician scales, the NIH composite 0-3 skin score was the only one that correlated with both clinician and patient perception of improvement or worsening. Of the patient-reported scales, the skin subscale of the Lee Symptom Scale was the only one that correlated with both clinician and patient perception of improvement or worsening. At study entry, NIH skin score 3 and Lee skin symptom score > 15 were both associated with worse overall survival. Worsening of NIH skin score at 6 months was associated with worse overall survival. Improvement in the Lee skin symptom score at 6 months was associated with improved overall survival. Our findings support the use of the NIH composite 0-3 skin score and the Lee skin symptom score as simple and sensitive measures to evaluate skin involvement in clinical trials as well as in the clinical monitoring of patients with cutaneous chronic GVHD. (*Blood*. 2012;120(13):2545-2552)

Medscape EDUCATION **Continuing Medical Education online**

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Medscape, LLC and the American Society of Hematology. Medscape, LLC is accredited by the ACCME to provide continuing medical education for physicians.

Medscape, LLC designates this Journal-based CME activity for a maximum of 1.0 **AMA PRA Category 1 Credit(s)™**. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

All other clinicians completing this activity will be issued a certificate of participation. To participate in this journal CME activity: (1) review the learning objectives and author disclosures; (2) study the education content; (3) take the post-test with a 70% minimum passing score and complete the evaluation at **http://www.medscape.org/journal/blood**; and (4) view/print certificate. For CME questions, see page 2774.

Disclosures

The authors, the Associate Editor Martin S. Tallman, and CME questions author Laurie Barclay, freelance writer and reviewer, Medscape, LLC, declare no competing financial interests.

Learning objectives

Upon completion of this activity, participants will be able to:

1. Describe the purpose and utility of scales to measure skin response in cutaneous chronic GVHD that correlate with clinician and patient perception of improvement or worsening, based on findings of a prospective, multicenter cohort study.
2. Describe the clinical utility of the NIH composite 0-3 skin score in patients with cutaneous chronic GVHD.
3. Describe the clinical utility of the Lee skin symptom score in patients with cutaneous chronic GVHD.

Release date: September 27, 2012; Expiration date: September 27, 2013

Submitted April 14, 2012; accepted June 19, 2012. Prepublished online as *Blood* First Edition paper, July 6, 2012; DOI 10.1182/blood-2012-04-424135.

There is an Inside *Blood* commentary on this article in this issue.

The online version of this article contains a data supplement.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Introduction

The major cause of late morbidity and nonrelapse mortality (NRM) after allogeneic hematopoietic cell transplantation (HCT) is chronic graft-versus-host disease (GVHD).[1] The risk of chronic GVHD is increased with older recipient age, use of unrelated donors, peripheral blood as the HCT source, and treatment with donor-lymphocyte infusion.[2,3] Published reports of agents targeting chronic GVHD have measured clinical benefit using scales that have not been validated and are not universally accepted.[4-7] Validated measurement tools to assess changes in the severity of chronic GVHD are needed for conduct of robust clinical trials of chronic GVHD therapies, and for comparison with prior studies.

Recognizing the lack of uniformity in reported chronic GVHD studies and absence of a widely accepted gold standard for determining activity of chronic GVHD or response to treatment, the 2005 National Institutes of Health (NIH) Consensus Conference proposed new criteria for scoring disease severity at initial diagnosis[8] and for measuring response in clinical trials.[9] In an effort to validate and refine the NIH consensus measures, the Chronic GVHD Consortium has undertaken a multicenter, prospective, longitudinal study in patients with chronic GVHD[10] which includes the evaluation of instruments proposed by the NIH Consensus and other measures reported in chronic GVHD trials. The goal of the study is to identify the most useful instruments to grade chronic GVHD severity, to capture clinically meaningful changes over time, and to correlate with overall survival (OS), NRM, and quality of life.

The skin is one of most common organs affected with chronic GVHD[11] and can cause significant morbidity. The variety of skin manifestations and the subjectivity in interpreting the degree of sclerosis and overall degree of involvement have led to poor interrater reliability between experts and clinicians.[12] In practice, clinicians have usually relied on their memory or whatever descriptions are available in the medical records to ascertain whether a patient's manifestation is objectively better or worse over time. Very few institutions include digital photography of the skin in their records, but even so, this measure does not capture texture. Use of ultrasound to quantify skin involvement has not been adopted, and would only be useful for sclerotic lesions.[13,14] Validated instruments that adequately capture the severity and changes over time (response) of cutaneous chronic GVHD are needed for clinical trials.

The current analysis evaluated several measures of cutaneous chronic GVHD. The instruments included 2 NIH recommended scales (the NIH composite 0-3 skin scoring and the NIH skin response scale),[8,9] the Vienna Skin Scale,[15] the Johns Hopkins sclerosis and fasciitis scales,[6] and 2 GVHD symptom measures including the skin subscale of the Lee Symptom Scale[16] and pruritus.[9] The aim of this study was to identify which of the evaluated skin scales or combination of instruments best correlated with clinician and patient perception of skin change, as well as with major outcomes such as survival. A simple and validated instrument to evaluate skin response in clinical trials would represent an important methodologic contribution to chronic GVHD research.

Methods

Chronic GVHD Consortium: description of study cohort

A cohort of HCT recipients with chronic GVHD was assembled prospectively in a multicenter observational study. The protocol was approved by the institutional review board at each of the 9 participating sites (Table 1), and all subjects provided written informed consent in accordance with the Declaration of Helsinki. Study participants were allogeneic HCT recipients age 2 years or older with chronic GVHD (classic and overlap subtypes) requiring systemic immunosuppressive therapy.[8] Cases were classified as incident (enrollment $<$ 3 months after chronic GVHD diagnosis) or prevalent (enrollment 3 or more months after chronic GVHD diagnosis but within 3 years of HCT). At enrollment and every 6 months thereafter, clinicians and patients report standardized information on chronic GVHD organ involvement and symptoms. Incident cases had an additional assessment time point 3 months after enrollment. The overall severity of chronic GVHD was assessed by the NIH consensus global severity scoring (based on number of organs involved and severity score in each organ),[8] by the clinician perception of severity (mild, moderate, and severe, 0-10 scale), and by the patient perception of severity (mild, moderate, and severe, 0-10 scale).[12] Standardized chart review after each visit abstracted objective medical data (including ancillary testing and laboratory results), medical complications, and medication profiles.

Skin assessment measures

Seven instruments used to assess skin involvement by GVHD were evaluated in this study. The NIH skin response scale (supplemental Figure 1A, available on the *Blood* Web site; see the Supplemental Materials link at the top of the online article) scores 8 body regions according to percentage of body surface area (BSA) involved with erythematous rash (which includes any cutaneous manifestations other than sclerosis), movable sclerosis, and nonmovable sclerosis.[9] The scale was scored as the involved percentage weighted by BSA, with possible values of 0%-100% for each manifestation. This instrument was proposed by the NIH Consensus for evaluating skin response in clinical trials. The NIH composite 0-3 skin scoring considers the extent of skin involvement, the presence of sclerotic features, and symptoms into a composite scale (supplemental Figure 1B). This composite skin score was proposed by the NIH Consensus to assess severity at initial diagnosis and staging of chronic GVHD.[8] The Vienna Skin Scale (VSS) instrument scores 10 body regions with percentage involvement of pigmentary changes, rash, and sclerosis (supplemental Figure 1C). Regional scores are summed for a Vienna Skin Total (VST) score of 0-50.[15] Note that the 3 instruments do not include identical skin manifestations, so it is possible to have involvement by one scale but not by another. In particular, the NIH 0-3 composite skin score includes functional and symptom considerations. The VSS includes alopecia, hypopigmentation, and hyperpigmentation in its scoring. The other 4 instruments analyzed were the Hopkins skin sclerosis score (0-4; supplemental Figure 1D),[6] the Hopkins fasciitis score (0-3; supplemental Figure 1E),[6] and 2 patient-reported measures: skin itching (0-10)[8] and the Lee Symptom Scale, skin subscale (5 items, 0-100).[16] A change of 15 points in the Lee skin symptom score is considered clinically significant. At follow-up visits every 6 months, patients and providers rated separately their perception of change in skin involvement on an 8-point scale, which was analyzed as improved ["(1) completely gone," "(2) very much better," "(3) moderately better"], stable ["(4) a little better," "(5) about the same," "(6) a little worse"], or worse ["(7) moderately worse," "(8) very much worse"].

Statistical methods

Patient sociodemographics, transplantation characteristics, and GVHD organ severities are presented as median and range for continuous variables, and as frequency and percentage for categorical variables.

At each study visit, presence or absence of cutaneous involvement was classified based on 2 of the clinician report measures. Presence as recorded by both the NIH skin response measure (% BSA $>$ 0 for erythematous rash, movable sclerosis, nonmovable sclerosis, fasciitis, and/or deep sclerosis) and the Vienna Skin Score (VST $>$ 0) was required to indicate skin involvement for a given patient at a given visit. Sequential change scores for skin measures were calculated by subtracting previous scores from current values. For example, change in erythema could range from -100 to 100. Multivariable regression models were constructed to examine the association between clinician and patient perception of change in skin

Table 1. Characteristics of the study cohort

Characteristics	N	Total population, N (%)		Without skin involvement, n (%)		With skin involvement, n (%)	*P*
Participating sites, no. (%)	458		207		250*		.002
Fred Hutchinson Cancer Research Center		214 (46)		95 (46)		118 (47)	
Stanford University Medical Center		65 (14)		30 (14)		35 (14)	
Dana-Farber Cancer Institute		53 (12)		17 (8)		36 (14)	
University of Minnesota		51 (11)		34 (16)		17 (7)	
Vanderbilt University Medical Center		36 (8)		17 (8)		19 (8)	
Medical College of Wisconsin		15 (3)		1 (1)		14 (6)	
Northwestern Children's Hospital		13 (3)		8 (4)		5 (2)	
Moffitt Cancer Center		8 (2)		4 (2)		4 (2)	
Washington University Medical Center		3 (1)		1 (1)		2 (1)	
Case type, no. (%)	458		207		250		.62
Incident		254 (55)		112 (54)		141 (56)	
Prevalent		204 (45)		95 (46)		109 (44)	
Patient age, y, no. (%)	458		207		250		.32
18 or younger		14 (3)		8 (4)		6 (2)	
19-49		194 (42)		93 (45)		100 (40)	
50 or older		250 (55)		106 (51)		144 (58)	
Patient sex, no. (%)	458		207		250		.03
Male		264 (58)		108 (52)		155 (62)	
Female		194 (42)		99 (48)		95 (38)	
Race, no. (%)	458		207		250		< .001
White		411 (90)		175 (85)		235 (94)	
Nonwhite		47 (10)		32 (15)		15 (6)	
Transplantation stem cell source, no. (%)	458		207		250		.01
Peripheral blood		402 (88)		171 (83)		230 (92)	
Bone marrow		34 (7)		22 (11)		12 (5)	
Cord blood		22 (5)		14 (7)		8 (3)	
Transplantation conditioning type	457		207		249		.99
Myeloablative		260 (57)		118 (57)		142 (57)	
Not myeloablative		197 (43)		89 (43)		107 (43)	
Donor type, no. (%)	457		207		249		.74
HLA-identical related		202 (45)		91 (44)		111 (45)	
HLA-matched unrelated		180 (39)		79 (38)		100 (40)	
HLA-mismatched		75 (16)		37 (18)		38 (15)	
Donor sex, no. (%)	453		205		247		.06
Female into male		130 (29)		50 (24)		80 (32)	
Other		323 (71)		155 (76)		167 (68)	
Prior acute GVHD, no. (%)	458		207		250		.28
Prior acute GVHD		311 (68)		135 (65)		175 (70)	
No prior acute GVHD		147 (32)		72 (35)		75 (30)	
Median months from transplantation to enrollment (range)	458	12.2 (3-299)	207	12.1 (3-299)	250	12.6 (3-39)	.95

*One patient missing the National Institutes of Health (NIH) skin response measure.

involvement and sequential change scores for skin measures, limiting the analysis to visits with skin involvement in the current and/or previous visit. Covariates adjusted in all models were Karnofsky performance status at chronic GVHD onset (< 80, ≥ 80), case type (incident, prevalent), and months from HCT to enrollment (< 12 months, ≥ 12 months). These covariates were all related to perceived change in univariate analysis. Linear mixed models with random patient effect were used to account for within-patient correlation, and study site (Fred Hutchinson Cancer Research Center, other sites) was an additional covariate to account for unmeasured patient selection effects. Type I error was controlled by considering a 2-sided *P* value of .01 or lower as statistically significant.

OS was calculated as months between enrollment and death, with follow-up censored at date of last contact. NRM was defined as death before relapse, with relapse treated as a competing event. OS was displayed using the Kaplan-Meier method, and NRM by cumulative incidence curves. Survivor function estimates (ie, 2-year OS probability) were calculated using the Breslow method.[17] Cox regression models were used to predict OS and NRM based on clinician skin severity score or Lee skin symptom score at enrollment. Skin assessment measure hazard ratios (HRs) were estimated adjusting for study site (Fred Hutchinson Cancer Research Center, other sites), months from HCT to cohort enrollment (< 12, ≥ 12 months), case type (incident, prevalent), platelet count at chronic GVHD onset ($\geq 100 \times 10^9$/L, $< 100 \times 10^9$/L), Karnofsky performance status at chronic GVHD onset (< 80, ≥ 80), patient age at transplantation (≥ 50, < 50 years), donor type (matched related, matched unrelated, mismatched), donor-patient gender combination (female into male, other), and prior history of acute GVHD (yes, no). These covariates were chosen as known chronic GVHD mortality risk factors[18-21] and to control for potential recruitment population differences between study sites. HRs were reported with clinician skin severity score of 0 ("no skin involvement") or Lee skin symptom score < 15 as the reference.

Additional landmark analyses fitted Cox regression models for OS and NRM from the 6- or 12-month visit, predicted by change in the NIH 0-3 composite skin score from baseline, categorized as "improved (< 0)," "stable ($= 0$)," or "worsened (> 0)" or change in the Lee skin symptom score (15 or more point change considered clinically meaningful). HRs were calculated with clinician skin severity change score of 0 ("stable") as the reference. A final series of Cox regression models treated clinician skin severity score as a time-dependent covariate. Statistical analyses were conducted using SAS/STAT Version 9.2 software (SAS Institute Inc) and R Version 2.14.2 R (Foundation for Statistical Computing).

Table 2. Chronic GVHD characteristics at study enrollment

Characteristics	N	Total population, N (%)		Without skin involvement, n (%)		With skin involvement, n (%)	P
NIH consensus global GVHD severity score	458		207		250		.002
Less than mild		2 (1)		1 (1)		1 (1)	
Mild		46 (10)		28 (14)		18 (7)	
Moderate		262 (57)		129 (62)		133 (53)	
Severe		148 (32)		49 (24)		98 (39)	
Clinician reported overall chronic GVHD severity score	458		207		250		< .001
Mild		223 (49)		127 (61)		96 (38)	
Moderate		193 (42)		72 (35)		121 (48)	
Severe		42 (9)		8 (4)		33 (13)	
Patient reported overall severity score	386		172		213		.21
None		11 (3)		8 (5)		3 (1)	
Mild		207 (53)		95 (55)		112 (53)	
Moderate		138 (36)		58 (34)		80 (38)	
Severe		30 (8)		11 (6)		18 (8)	
NIH 0-3 composite skin score	458		207		250		< .001
0		173 (38)		159 (77)		14 (6)	
1		107 (23)		26 (13)		81 (32)	
2		111 (24)		14 (7)		97 (39)	
3		67 (15)		8 (4)		58 (23)	
Skin involvement by VSS	457		207		250		NA
Involved (VST > 0)		306 (67)		56 (27)		250 (100)	
Not involved (VST = 0)		151 (33)		151 (73)		0 (0)	
Other organ involvement							
Oral	458	276 (60)	207	134 (65)	250	141 (56)	.07
Gastrointestinal	458	136 (30)	207	75 (36)	250	61 (24)	.006
Eye	458	221 (48)	207	107 (52)	250	113 (45)	.17
Joints and fascia	458	135 (29)	207	50 (24)	250	85 (34)	.02
Lung	458	228 (50)	207	114 (55)	250	113 (45)	.04
Liver	456	235 (52)	207	112 (54)	248	122 (49)	.30
Genital	416	49 (12)	187	22 (12)	228	27 (12)	.98
Platelet count at chronic GVHD onset, $\times 10^9$/L	447		207		239		.67
< 100		110 (25)		53 (26)		57 (24)	
≥ 100		337 (75)		154 (74)		182 (76)	
Karnofsky performance at onset of chronic GVHD	348		159		188		.88
< 80		74 (21)		34 (21)		39 (21)	
≥ 80		274 (79)		125 (79)		149 (79)	

NIH indicates National Institutes of Health; VSS, Vienna Skin Scale; and VST, Vienna Skin Total.

Results

Demographics and baseline characteristics

Nine participant sites in the Chronic GVHD Consortium enrolled 458 patients in the trial as of December 2010. Table 1 describes characteristics of the 458 patients at the time of enrollment (baseline), of whom 250 (55%) had skin involvement as measured both by the NIH response measure (> 0% BSA) and the VSS (VST > 0). Chronic GVHD severity measures at enrollment are summarized in Table 2. The study cohort was represented equally by incident and prevalent chronic GVHD cases (Table 1). Almost 90% of subjects received mobilized blood as the stem cell graft source. Thirty-two percent of subjects had severe chronic GVHD overall, per NIH Consensus criteria. One hundred forty-seven patients (59%) had 2 or more time points included in the analysis.

Figure 1 displays skin manifestations among the 257 patients (56%) with involvement at enrollment according to the NIH skin response scale. The most common manifestation according to NIH skin response scale was erythema alone (60%), which includes any cutaneous involvement other than sclerosis, but all combinations of erythema, movable sclerosis, and nonmovable sclerosis were observed, including presence of all 3 manifestations in 8 patients

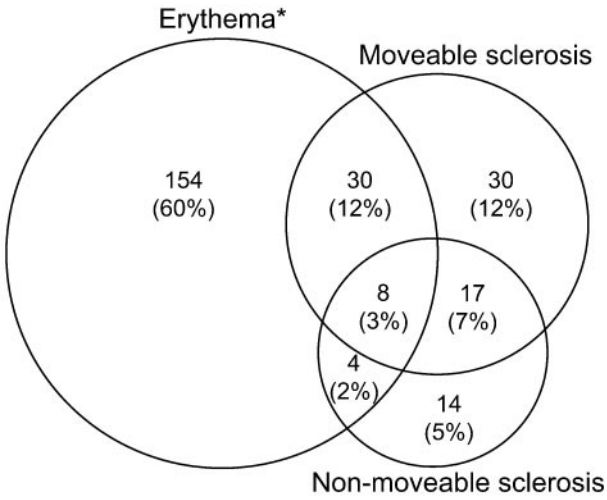

Figure 1. Cutaneous manifestations of chronic GVHD by NIH skin response criteria, for 257 (56%) of 457 cases with skin involvement at enrollment. *Any skin GVHD type except sclerosis/fasciitis.

Table 3. Multivariable linear mixed models to predict changes in skin severity according to 9 different measures

Sequential change of skin severity measures/Contrast in clinician or patient perception of skin change	Clinician		Patient	
	Estimate (95% CI)	*P*	Estimate (95% CI)	*P*
NIH response erythema* scale[9]				
Improve vs stable	**−10.49 (−14 ~ −6.97)**	**< .001**	**−5.58 (−9.81 ~ −1.36)**	**.01**
Worsen vs stable	6.42 (0.48 ~ 12.36)	.03	6.17 (−2.26 ~ 14.59)	.15
NIH response movable sclerosis scale[9]				
Improve vs stable	−2.07 (−3.82 ~ −0.31)	.02	−0.8 (−2.72 ~ 1.12)	.41
Worsen vs stable	1.35 (−1.62 ~ 4.32)	.37	−1.36 (−5.18 ~ 2.47)	.49
NIH response nonmovable sclerosis scale[9]				
Improve vs stable	−0.44 (−1.62 ~ 0.74)	.47	0.05 (−1.2 ~ 1.31)	.93
Worsen vs stable	**5.13 (3.14 ~ 7.12)**	**< .001**	1.95 (−0.56 ~ 4.46)	.13
Vienna Total Score[15]				
Improve vs stable	**−1.63 (−2.21 ~ −1.05)**	**< .001**	−0.55 (−1.24 ~ 0.13)	.11
Worsen vs stable	**1.8 (0.82 ~ 2.78)**	**< .001**	1.16 (−0.21 ~ 2.53)	.10
Hopkins sclerotic scale[6]				
Improve vs stable	−0.14 (−0.29 ~ 0.01)	.07	−0.09 (−0.26 ~ 0.07)	.27
Worsen vs stable	**0.4 (0.15 ~ 0.65)**	**.002**	0.26 (−0.08 ~ 0.59)	.13
Hopkins fascia scale[6]				
Improve vs stable	**−0.15 (−0.26 ~ −0.03)**	**.01**	−0.1 (−0.22 ~ 0.02)	.09
Worsen vs stable	0.09 (−0.1 ~ 0.27)	.36	0.04 (−0.2 ~ 0.28)	.73
Skin itching				
Improve vs stable	−0.3 (−0.85 ~ 0.25)	.28	−0.53 (−1.06 ~ −0.01)	.05
Worsen vs stable	0.81 (−0.11 ~ 1.73)	.08	**2.88 (1.82 ~ 3.94)**	**< .001**
NIH composite 0-3 score[8]				
Improve vs stable	**−0.74 (−0.93 ~ −0.56)**	**< .001**	**−0.36 (−0.58 ~ −0.14)**	**.002**
Worsen vs stable	**0.78 (0.46 ~ 1.09)**	**< .001**	**0.79 (0.35 ~ 1.24)**	**< .001**
Skin subscale, Lee Symptom Scale[16]				
Improve vs stable	**−11.9 (−16.05 ~ −7.76)**	**< .001**	**−8.75 (−12.7 ~ −4.79)**	**< .001**
Worsen vs stable	**9.89 (2.84 ~ 16.93)**	**.006**	**28.91 (20.64 ~ 37.18)**	**< .001**

The linear mixed models were used to predict clinician or patient-reported change in skin involvement according to sequential changes in skin severity by 9 different measures ($P \leq .01$ in bold). Each model includes one sequential change measure as the outcome, one perceived change measure as the covariate of interest, and controls for study site and for factors related to perceived change (Karnofsky performance status, case type, and months from HCT to enrollment).

NIH indicates National Institutes of Health; CI, confidence interval; and HCT, hematopoietic cell transplantation.

*Includes any cutaneous manifestations other than sclerosis and fasciitis.

(3%). According to the NIH composite 0-3 score, 285 patients (62%) had skin involvement at enrollment: 107 (37%) with score 1, 111 (39%) with score 2, and 67 (24%) with score 3.

Serial change in skin measures as a predictor for perceived change at follow up visits

Of the 961 follow-up visits, 543 were for patients with skin involvement at that visit or the previous visit. One hundred fifty-two (28%) were 6-month visits, 102 (19%) were 12-month visits, and the rest were 18-month or later visits. Change in skin manifestation was rated by the medical providers as improved in 45%, stable in 46%, or worse in 9%, and by patients as 58%, 35%, and 7%, respectively. The concordance between clinicians and patients was moderate (weighted $\kappa = 0.41$).

Table 3 shows the results of multivariable linear mixed models examining the association between sequential changes of skin measures and clinician or patient perception of change. While changes in several skin scores correlated with either changes reported by clinicians or patients, only changes in the NIH composite 0-3 skin score and in the skin subscale of the Lee symptom score showed sensitivity to both worsening and improvement in the perceived change by both clinicians and patients. The NIH composite 0-3 skin score decreased by an estimated average of 0.74 (95% confidence interval [CI] 0.56-0.93) point when clinicians perceived improvement versus stability, and by a corresponding 0.36 (95% CI 0.14-0.58) point when patients perceived the same improvement. The NIH composite 0-3 skin score increased by an estimated average of 0.78 (95% CI 0.46-1.09) point when clinicians perceived worsening versus stability, and by a corresponding 0.79 (95% CI 0.35-1.24) point when patients perceived the same worsening (Table 3). Examination of the patterns of change in the different skin measures showed the decreased erythema appeared best correlated with improvement in the NIH composite 0-3 score whereas worsening in all types of skin involvement (erythema, movable sclerosis, nonmovable sclerosis, worsening range of motion in joints) was seen when the NIH composite score worsened. These patterns were consistent regardless of the initial NIH skin severity.

The estimated mean changes in the Lee skin symptom score with both the clinician and patient perception of improvement or worsening compared with stability are shown in Table 3. These results are both clinically and statistically significant ($P < .01$).

Association between NIH composite 0-3 skin score and mortality

Because the NIH composite 0-3 scale was the only clinician skin measure that correlated with clinician and patient perception in both the improvement and worsening directions, we analyzed its potential association with OS and NRM. Of the 458 patients enrolled, 454 patients had subsequent follow-up visits and were included in these analyses. A total of 75 patients died, 58 without recurrent or progressive malignancy. The median follow-up time of survivors was 18.8 months (range 2.6-47.2). Figure 2 displays OS and NRM according to enrollment NIH 0-3 skin score. Two-year OS for NIH skin scores of 0, 1, 2, and 3 at study entry were 86%

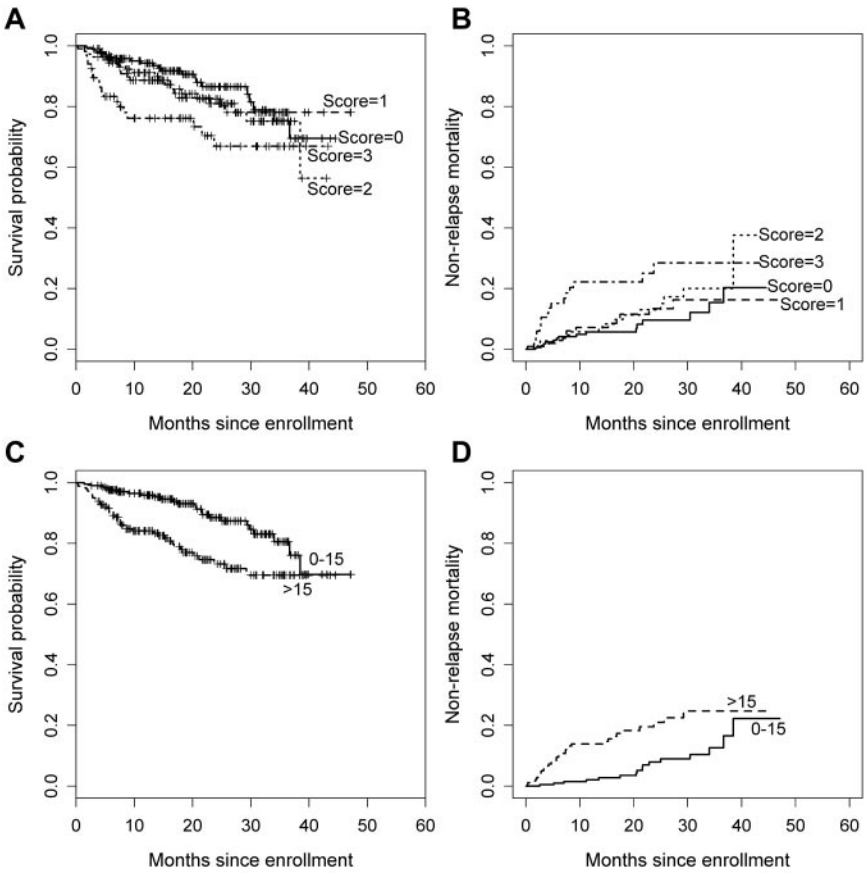

Figure 2. Outcomes according to baseline NIH skin score and Lee skin score. (A) Overall survival and (B) nonrelapse mortality according to baseline NIH composite 0-3 skin score. (C) Overall survival and (D) nonrelapse mortality according to baseline Lee skin symptom score.

Table 4. Association between the NIH composite 0-3 score (clinician measure) or Lee Symptom Scale (patient measure) and OS and NRM

	OS		NRM	
	HR (95% CI)	***P***	**HR (95% CI)**	***P***
NIH composite 0-3 skin score				
Score at enrollment				
3 vs 0	**2.9 (1.5 ~ 5.7)**	**.001**	**4.5 (2.2 ~ 9.4)**	**< .001**
Change at 6 mo				
Improve	1.2 (0.3 ~ 4.3)	.81	1.0 (0.2 ~ 3.8)	.94
Worsen	4.7 (0.9 ~ 25.1)	.07	7.2 (1.2 ~ 43.9)	.03
Change at 12 mo				
Improve	1.6 (0.3 ~ 9.6)	.58	0.9 (0.1 ~ 6.6)	.91
Worsen	1.7 (0.2 ~ 16.4)	.65	3.1 (0.3 ~ 34.9)	.37
Skin subscale, Lee Symptom Scale				
Score at enrollment				
> 15 points	**2.3 (1.3 ~ 3.8)**	**.002**	**2.9 (1.6 ~ 5.4)**	**.001**
Change at 6 mo				
Improve	**0.07 (0.008 ~ 0.6)**	**.01**	0.09 (0.01 ~ 0.8)	.03
Worsen	1.6 (0.4 ~ 6.0)	.46	2.1 (0.5 ~ 8.3)	.29
Change at 12 mo				
Improve	0.4 (0.03 ~ 5.7)	.50	0.3 (0.02 ~ 4.0)	.35
Worsen	3.4 (0.5 ~ 21.1)	.19	4.0 (0.6 ~ 28.1)	.17

Each model includes either the NIH composite skin score or the Lee Symptom Scale as the covariate of interest, and controls for study site and for factors associated with chronic GVHD health outcomes (months from HCT to cohort enrollment, case type, platelet count at chronic GVHD onset, Karnofsky performance status at chronic GVHD onset, patient age at transplantation, donor type, donor-patient gender combination, and prior history of acute GVHD).

NIH indicates National Institutes of Health; OS, overall survival; NRM, nonrelapse mortality; HCT, hematopoietic cell transplantation; HR, hazard ratio; and CI, confidence interval.

Bold values represent HRs with $P \leq .01$.

(95 CI, 80%-92%), 83% (75%-91%), 81% (73%-89%), and 69% (58%-83%), respectively. The 2-year NRM for NIH skin scores of 0, 1, 2, and 3 at study entry were 10% (5%-16%), 13% (6%-20%), 15% (8%-22%), and 30% (16%-41%), respectively. After adjusting for known chronic GVHD risk factors, a baseline NIH skin score of 3 was associated with higher overall mortality (HR 2.9, 95% CI 1.5-5.7, $P = .001$) and higher NRM (HR 4.5, 2.2-9.4, $P < .001$) compared with no skin involvement (Table 4). In addition, worsening in NIH skin score at 6 months was associated with subsequent overall mortality (HR 4.7, 0.9-25.1, $P = .07$) and higher subsequent NRM (HR 7.2, 1.2-43.9, $P = .03$) compared with stable NIH skin score among patients with cutaneous involvement at study entry. Change of NIH skin score at 12 months was not associated with subsequent OS or NRM.

Using NIH 0-3 skin score as a categorical time-dependent covariate, similar associations were found with OS and NRM. Compared with no skin involvement (score 0), NIH skin score of 3 was associated with lower subsequent OS (HR 4.3, 95% CI 2.3-8.2, $P < .001$) and higher NRM (HR 6.9,95% CI 3.3-14.3, $P < .001$). In addition, NIH skin score of 2 was associated with worse OS (HR 1.9, 95% CI 1.0-3.5, $P = .06$) but not with NRM ($P = .13$).

Incident versus prevalent chronic GVHD type was not significantly associated with OS or NRM and did not show any statistical interaction with skin involvement (data not shown).

Association between the skin subscale of the Lee Symptom Scale and mortality

Given that the skin subscale of the Lee Symptom Scale was the only patient-reported scale that correlated with clinician and patient perception of change in both directions, we looked at its association with OS and NRM. In a multivariate model with known risk factors as covariates, a score greater than 15 in the skin subscale of the Lee Symptom Scale at enrollment was associated with increased subsequent overall mortality (HR 2.3, 95% CI 1.3-3.8, $P = .002$) and with higher NRM (HR 2.9, 95% CI 1.6-5.4, $P = .001$). Figure 2C and D show OS and NRM according to Lee skin symptom score dichotomized between 0-15 versus > 15 points at study entry. Improvement in the skin subscale at 6 months correlated with improved subsequent OS ($P = .01$) and with NRM ($P = .03$), but worsening skin subscale was not associated with statistically significant worsening in subsequent OS or NRM. There was no statistically significant association with change in the skin subscale at 12 months with OS or NRM. Testing of the skin subscale as either an enrollment score or as a time-varying covariate showed similar associations with OS and NRM.

Discussion

In this prospective, longitudinal observational study, we have demonstrated that among the 5 clinician-assessment measures of burden of skin chronic GVHD, changes in the simplest scale (NIH composite 0-3 score) were the only ones that correlated well with provider- and patient- reported perception of skin changes (worsening and improved compared with stability). In addition, the NIH composite score at enrollment correlated with subsequent OS and NRM. Of the 2 patient-reported outcome measures tested, the skin subscale of the Lee Symptom Scale correlated with OS and NRM, and changes in the skin subscale over 6-month intervals correlated with provider and patient perceptions of skin changes. Thus, our results suggest that these 2 measures best capture cross-sectional severity of skin chronic GVHD, and that changes over time in these measures reflect clinical benefit. When changes occur in these measures, both providers and patients can perceive a difference and there is an association with subsequent OS and NRM. Considering that both the NIH composite score and the Lee symptom score include components of symptoms or functional status in their response options, we believe they reflect the clinical sense that symptoms are critical to assess response in cutaneous chronic GVHD and that physical findings alone are not sufficient. This may explain the superior performance of these 2 scales over the others tested in our study. Some of the other scales also correlated with changes, particularly to clinician perceived changes, but none achieved the full range of sensitivity seen with the NIH skin score or the Lee skin symptom scale.

The NIH composite 0-3 skin score, originally recommended only to assess skin severity at initial diagnosis and for subsequent scoring purposes[8] (but not response determination), is easy to use given the broad categories for recording the percentage of skin involvement, presence of movable or unmovable sclerosis, and symptoms. The incorporation of symptoms into this scale (eg, severe pruritus, decreased mobility) may explain why it correlated well with clinician and patient perception of change. A recent study showed a similar good correlation of change in the NIH composite eye score with physician and patient perception of change in ocular GVHD symptoms.[22] The lack of correlation in the NIH skin response measure (percentage of body surface area with erythema, movable, and unmovable sclerosis manifestations)[9] with clinician and patient perception of change suggests that small changes in skin manifestations did not reach the threshold for changes in clinically significant symptoms or functional impairments that were noticable to the clinician or the patient. Poor interrater reliability of the NIH skin response measure has been previously reported,[12] also suggesting that this measure may be difficult to complete compared with the NIH 0-3 composite skin score.

One limitation of the NIH composite 0-3 score is that this scale combines heterogeneous manifestations into the same score. For example, patients with any erythematous or lichen-planus–like cutaneous involvement of 25%-50% BSA are combined with those having movable sclerosis into score 2. Patients with cutaneous involvement of $> 50\%$, unmovable sclerosis, or severe pruritus are all scored a 3. Because changes in sclerosis take many months, if not years, to resolve, the 0-3 scoring system may not be sensitive enough to detect slow but meaningful changes. Future work should evaluate which elements of the composite score are the most clinically relevant under different scenarios of diagnosis and treatment. For example, a separate sclerosis scale may be needed to address the different natural history of this manifestation and to evaluate its treatments.

Among the 2 patient-reported skin scales evaluated in our study, we also have demonstrated that the skin subscale of the Lee Symptom Scale correlated well with perceived change by both clinician and patient. Furthermore, we showed that a low baseline score, as well as improvement at 6 months, is associated with better OS and lower NRM. This simple 5-item patient-reported instrument could therefore be incorporated into chronic GVHD trials as an additional measure of response.

There are number of limitations to this study. Treatment was not controlled in this observational study, but our aim was to evaluate accuracy and responsiveness of skin chronic GVHD assessments that would be applicable for treatments using a broad array of

agents. Treatment data were collected and may be used in future analyses to address different aims. The study population is primarily white, so results may not apply to nonwhite patients. The associations between the skin scores and the survival outcomes were strongest early after the assessment point and for people with the most severely involved skin. Finally, this analysis was limited to skin measurement scales and any potential correlation between changes in skin with changes in other organs affected by chronic GVHD were not addressed. Current studies using the data from the Chronic GVHD Consortium are addressing the correlation of change among multiple involved organs according to the NIH response scales and the potential value to such change in predicting long-term outcomes. In addition, intervention studies such as the Clinical Trial Network chronic GVHD trial are using NIH scales and other measures that will allow for analysis of correlation of changes with treatment response and other major outcomes.

Our findings support the use of the NIH composite skin score 0-3 and the Lee skin symptom scale as sensitive measures to evaluate skin involvement in clinical trials as well as in the clinical monitoring of patients with cutaneous chronic GVHD. The validation of skin measures for chronic GVHD is a step forward to simplify the clinician assessment of cutaneous changes over time, which we hope will rekindle the interest in clinical trials testing novel agents for response in cutaneous chronic GVHD.

Acknowledgments

This work was supported by grants CA118953 and CA163438 from the National Institutes of Health, Department of Health and Human Services and the National Cancer Institute.

Authorship

Contribution: D.A.J. and M.E.D.F. proposed the study concept, analyzed data, and wrote the manuscript; B.F.K. and X.C. performed statistical analyses and contributed to writing the manuscript; J.P., Y.I., J.M.P., S.A., M.A., M.J., C.C., D.W., P.J.M., S.Z.P., and G.V. contributed to data analysis and critical review of the manuscript; and S.J.L. contributed to the development of the study concept, data analysis, and writing of the manuscript.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Mary E. D. Flowers, MD, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, D5-290, Seattle, WA 98109; e-mail: mflowers@fhcrc.org.

References

1. 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(2):406-414.
2. Flowers ME, Inamoto Y, Carpenter PA, et al. Comparative analysis of risk factors for acute graft-versus-host disease and for chronic graft-versus-host disease according to National Institutes of Health consensus criteria. *Blood*. 2011; 117(11):3214-3219.
3. Chalandon Y, Passweg JR, Schmid C, et al. Outcome of patients developing GVHD after DLI given to treat CML relapse: a study by the Chronic Leukemia Working Party of the EBMT. *Bone Marrow Transplant*. 2010;45(3):558-564.
4. Johnston LJ, Brown J, Shizuru JA, et al. Rapamycin (sirolimus) for treatment of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2005;11(1):47-55.
5. Cutler C, Miklos D, Kim HT, et al. Rituximab for steroid-refractory chronic graft-vs.-host disease. *Blood*. 2006;108(2):756-762.
6. Jacobsohn DA, Chen AR, Zahurak M, et al. Phase II study of pentostatin in patients with corticosteroid-refractory chronic graft-versus-host disease. *J Clin Oncol*. 2007;25(27):4255-4261.
7. Flowers ME, Apperley JF, van Besien K, et al. A multicenter prospective phase 2 randomized study of extracorporeal photopheresis for treatment of chronic graft-versus-host disease. *Blood*. 2008;112(7):2667-2674.
8. 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(12):945-956.
9. Pavletic SZ, Martin P, Lee SJ, et al. Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease. IV. Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2006;12(3):252-266.
10. Chronic GVHD Consortium. Rationale and design of the chronic GVHD cohort study: improving outcomes assessment in chronic GVHD. *Biol Blood Marrow Transplant*. 2011;17(8):1114-1120.
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(2):415-419.
12. 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(11):1619-1629.
13. Leiter U, Kaskel P, Krahn G, et al. Psoralen plus ultraviolet-A-bath photochemotherapy as an adjunct treatment modality in cutaneous chronic graft versus host disease. *Photodermatol Photoimmunol Photomed*. 2002;18(4):183-190.
14. 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(4):402-407.
15. Greinix HT, Pohlreich D, Maalouf J, et al. A single-center pilot validation study of a new chronic GVHD skin scoring system. *Biol Blood Marrow Transplant*. 2007;13(6):715-723.
16. Lee S, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8(8):444-452.
17. Breslow NE. Discussion of Professor Cox's paper. *J Royal Stat Soc B*. 1972;34:216-217.
18. Arora M, Klein JP, Weisdorf DJ, et al. Chronic GVHD risk score: a Center for International Blood and Marrow Transplant Research analysis. *Blood*. 2011;117(24):6714-6720.
19. Sullivan KM, Witherspoon RP, Storb R, et al. Prednisone and azathioprine compared with prednisone and placebo for treatment of chronic graft-v-host disease: prognostic influence of prolonged thrombocytopenia after allogeneic marrow transplantation. *Blood*. 1988;72(2):546-554.
20. Vigorito AC, Campregher PV, Storer BE, et al. Evaluation of NIH consensus criteria for classification of late acute and chronic GVHD. *Blood*. 2009;114(3):702-708.
21. Wingard JR, Piantadosi S, Vogelsang GB, et al. Predictors of death from chronic graft-versus-host disease after bone marrow transplantation. *Blood*. 1989;74(4):1428-1435.
22. Inamoto Y, Chai X, Kurland BF, et al. Validation of measurement scales in ocular graft-versus-host disease. *Ophthalmology*. 2012;119(3):487-493.