Fred Hutch
Cancer Center

Stephanie J. Lee, MD, MPH
Professor & Section Head,
Hematologic Malignancies
Fred Hutchinson Cancer Center
1100 Fairview Ave. N.
P.O. Box 19024
Seattle, WA 98109
sjlee@fredhutch.org

11/27/2023

RE: Lee Symptom Scale

Dear Colleague,

Thank you for your interest in using the Lee Symptom Scale to assess chronic graft-versus-host disease symptom burden in your research project.

For the ORIGINAL 30 item, 1 month recall, Lee Symptom Scale and how to score it, see *Lee SJ, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus- host disease. Biol Blood Marrow Transplant 2002; 8:444-452.* There are seven subscales and one summary score. You can score a subscale if >50% of items are answered. You can calculate a summary score if >50% of scales were calculated. Additional validation of the scale was published in *Merkel et al. Content validity of the Lee chronic-graft-versus-host disease symptom scale as assessed by cognitive interviews. Biol Blood Marrow Transplant 2016; 22: 752-758.* Note that the headers in the survey do <u>not</u> correspond directly to the 7 subscales. See the *Teh et al.* paper below for clarification of the subscales.

The MODIFIED version of the Lee Symptom Scale, uses the “past 7 days” as the recall time instead of “past month.” Scoring of the MODIFIED LSS follows the same principles as the original scale. In some cases, 2 items are not included: “need to use oxygen” and “receiving nutrition from an intravenous line or feeding tube.” These modifications were documented in *Teh C, Onstad L, Lee SJ. Reliability and Validity of the Modified 7-Day Lee Chronic Graft-versus-Host Disease Symptom Scale. Biol Blood Marrow Transplant 2020; 26(3):562-567.* The original 7 subscales remain. Table 1 in this paper also clarifies the item groupings for scoring.

A further modification of the Lee Symptom Scale was used in the REACH3 trial, see *Zeiser R et al, Ruxolitinib for glucocorticoid-refractory chronic graft-versus-host disease. N Eng J Med 2021; 385: 228-238.* This scale replaces “bother” with “severity” and uses response options of Did not have this problem, Mild, Moderate, Severe, Very severe. It includes all 30 items. The instructions are: “Please let us know how severe any of the following problems have been in the past week.” Scoring is similar to the original scale. For future studies, we replaced “did not have this problem” with “no symptoms” and edited the instructions to: “By circling one (1) number per line, please indicate how severe your symptoms have been <u>in the past 7 days</u>:” These revisions better match the original scale.

The Lee Symptom Scale is endorsed by the 2005 and 2014 NIH Chronic GVHD Consensus Conferences as a validated patient-reported outcome measure to capture the symptom burden of chronic GVHD. The Lee Symptom Scale is mentioned in the FDA-labeling of ibrutinib, belumosudil and ruxolitinib for chronic GVHD.

Descriptions of the available scales are shown in the Table. A template database and SAS coding for scoring the instrument are available at https://research.fredhutch.org/lee/en/research.html. For permission to use the scales, please contact me at sjlee@fredhutch.org. Non-commercial use is free for academic and non-profit use but requires a license. Non-exclusive licenses and fees are required for commercial use. Some translations are available upon request.

Table. Versions of the Lee Symptom Scale

Name of scale	Recall period /concept	Response options	Reference
Original	Past month /Bother	Not at all, Slightly, Moderately, Quite a bit, Extremely	Lee, BBMT 2002; 8:444
mLSS_7_day_bother	7 days /Bother	Not at all, Slightly, Moderately, Quite a bit, Extremely	Teh, BBMT 2020; 26: 562
mLSS_7_day_severity	7 days /Severity	No symptoms, Mild, Moderate, Severe, Very Severe	Modified from Zeiser, NEJM 2021; 385: 228

Sincerely,

Stephanie J. Lee, MD, MPH
Professor and Section Head, Hematologic Malignancies, Clinical Research Division, Fred Hutchinson Cancer Center
Professor of Medicine, University of Washington
David and Patricia Giuliani/Oliver Press Endowed Chair in Cancer Research

Biology of Blood and Marrow Transplantation 8:444-452 (2002)

Development and Validation of a Scale to Measure Symptoms of Chronic Graft-versus-Host Disease

Stephanie J. Lee,[1,2] *E. Francis Cook,*[2] *Robert Soiffer,*[1,2] *Joseph H. Antin*[1,2]

[1]Department of Adult Oncology, Dana-Farber Cancer Institute; [2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Correspondence and reprint requests: Stephanie Lee, MD, MPH, Center for Outcomes and Policy Research, Dana-Farber Cancer Institute, 44 Binney St, Boston, MA 02115 (e-mail: stephanie_lee@dfci.harvard.edu).

Received April 1, 2002; accepted May 10, 2002

ABSTRACT

Chronic GVHD (cGVHD) affects 30% to 90% of surviving allogeneic transplant recipients. Thus far, no quality-of-life instruments have been developed to measure the effect of this common complication of allogeneic transplantation on patients' functioning and well-being. Using a prospective cohort of 107 patients with active cGVHD who completed the symptom scale at enrollment and at intervals of 3 and 6 months, we developed a 30-item symptom scale with 7 subscales to capture the cGVHD-specific symptom burden. The symptom scale correlated highly with patients' self-assessed mild, moderate, and severe cGVHD manifestations in cross-sectional analysis. Reliability (Cronbach's α = 0.79-0.90), test-retest (r^2 = 0.28-0.93), and convergent and discriminant validity compared to the Medical Outcomes Study Short Form 36 (SF-36) and Functional Assessment of Chronic Illness Therapy with BMT subscale (FACT-BMT) were assessed and found to be adequate. Longitudinal assessments showed that changes in overall health status correlated best with changes in quality of life as measured by the SF-36 and FACT-BMT. In contrast, changes in cGVHD severity were best detected by changes in the symptom scale. We recommend that either the SF-36 or the FACT-BMT be combined with a cGVHD-specific symptom scale to measure the impact of cGVHD on patients' quality of life and that this endpoint be included in clinical trials testing cGVHD interventions. The cGVHD symptom scale is a short, simple, and valid measure of cGVHD manifestations and can be used to follow complication-specific symptoms using patient self-administered questionnaires.

KEY WORDS

Chronic graft-versus-host disease • Stem cell transplantation • Quality of life • Outcomes research

INTRODUCTION

Chronic graft-versus-host disease (cGVHD) is a serious late complication of allogeneic stem cell transplantation and the leading cause of nonrelapse death more than 2 years after the procedure [1-6]. Thirty percent to 90% of surviving allogeneic transplant recipients develop cGVHD, which is associated with decreased quality of life (QOL) [7], impaired functional status [5,8], continued need for immunosuppressive medication [6], and increased mortality [1-6]. A higher incidence of cGVHD is observed when patients are older, when immunosuppressive medications instead of T-cell depletion are used for acute GVHD prophylaxis, or when unrelated, mismatched, or peripheral blood grafts serve as the stem cell source [9-16]. Nonmyeloablative procedures are becoming more common, but the comparable rates of acute GVHD in patients treated with nonmyeloablative and myeloablative procedures suggest that rates of cGVHD may also be similar [17-19].

No validated measures capturing the symptom burden of cGVHD have been published. Observational physician-reported data suggest that approximately 65% to 85% of patients with cGVHD have skin involvement, 60% have mouth involvement, 40% to 55% have liver involvement (although this is usually asymptomatic for patients), 25% to 45% have eye involvement, 20% to 30% have nutritional problems, and 10% to 15% have lung manifestations [20]. Thus, we developed a patient self-administered symptom scale reflective of the multiorgan manifestations of cGVHD based on a cohort accrued from 1998 to 2000. To be useful in following the QOL of patients with cGVHD enrolled in clinical studies, the symptom scale had to be sensitive to cross-sectional and longitudinal severity of cGVHD. We hypothesized that direct measurement of patient QOL and symptom burden offers the most sensitive means of following the clinical course of patients with cGVHD.

MATERIALS AND METHODS

Patients

Patients with active cGVHD following allogeneic stem cell transplantation at the Dana-Farber Cancer Institute and the Brigham and Women's Hospital Stem Cell Transplant Program were identified through physician review of patient lists every 3 months. Most patients were diagnosed clinically without histopathologic confirmation. Potentially eligible patients were mailed a cover letter, consent form, baseline self-administered questionnaire, opt-out card, and self-addressed stamped envelope in which to return their surveys. If neither a survey nor an opt-out card was received within 4 weeks of the initial mailing, a second mailing was sent. If no response was received within 3 weeks of the second mailing, the patient was contacted by phone to confirm receipt of study materials. Both incident and prevalent cases were enrolled in the cohort between 1998 and 2000. The Institutional Review Board approved the study protocol, and all patients provided signed, informed consent for participation.

Patients who had developed cGVHD in the 3 months preceding study enrollment were classified as newly diagnosed and asked to complete questionnaires every 3 months for the first year and every 6 months thereafter. Patients who were diagnosed with cGVHD more than 3 months prior to enrollment were designated as established and were surveyed every 6 months. The more intensive survey schedule for newly diagnosed patients reflected our hypothesis that the greatest changes in disease activity occur in the initial period after diagnosis.

Data Collection Instruments and Methods

The Short Form 36 (SF-36) [21,22], the Functional Assessment of Chronic Illness Therapy with bone marrow transplantation subscale (FACT-BMT) [23], and a checklist of 42 potentially bothersome symptoms of cGVHD were mailed to patients every 3 to 6 months. At each time point, patients were also asked to provide self-assessment of their current Karnofsky performance status (KPS), current level of overall health (excellent, very good, good, fair, poor), cGVHD severity (mild, moderate, or severe), and cGVHD trajectory over the last 6 months (improved, stable, worsened). If a survey or opt-out card was not returned within 3 to 4 weeks, a second cover letter and packet were mailed. If a response was not received after an additional 4 weeks, patients were contacted by phone. Sociodemographic information was collected at baseline.

The SF-36 is a validated, generic, QOL instrument measuring 8 subscales (physical, role physical, pain, general health, vitality, social functioning, role emotional, and mental health) and 2 summary scales (physical and mental). The FACT-BMT is a validated, cancer-specific QOL instrument with a transplantation-specific subscale designed to address additional areas of concern to transplant recipients. This survey is composed of 4 core domains (physical, social, emotional, and functional) and the transplantation-specific subscale (BMT module) that result in a summary measure (FACT-BMT) when combined. The checklist of 42 symptoms was developed based on review of the literature and discussion with transplantation physicians, nurses, and patients. The symptom checklist was pilot tested on 10 patients with cGVHD for ease of completion, clarity, and comprehensiveness of symptoms. Patients were asked to report whether or not they had certain symptoms and rate how bothersome they were on a 6-point Likert scale: "symptom not present," "not bothered at all," "slightly bothered," "moderately bothered," "quite a bit bothered," and "extremely bothered." Test results from 15 healthy volunteers suggested appropriate interval scaling for response categories.

Medical records were reviewed for clinical information by one physician (S. J. L.) who did not have knowledge of patients' survey responses. Data regarding type and degree of organ involvement and clinical severity of cGVHD (limited/extensive) [24] were abstracted from clinical staff notes that had been recorded within 6 weeks of each survey completion date. This restriction was intended to ensure that notes accurately reflected medical status at the time of the QOL assessment. Grades were assigned based on data available in the clinic notes, Eastern Cooperative Oncocology Group (ECOG) performance status, cGVHD severity (mild/moderate/severe), and overall health (on a 5-point Likert scale: excellent, very good, good, fair, or poor). Mild cGVHD was defined as "signs and symptoms of cGVHD do not interfere substantially with function and do not progress once appropriately treated with local therapy or standard systemic therapy (steroids and/or cyclosporine or tacrolimus)." Moderate cGVHD was defined as "signs and symptoms of cGVHD interfere somewhat with function despite appropriate therapy or are progressive through first-line systemic therapy defined as steroids and/or cyclosporine or tacrolimus." Severe cGVHD was defined as "signs and symptoms of cGVHD limit function substantially despite appropriate therapy or are progressive through second-line therapy." These definitions have not been independently validated. Information on relapse and death was obtained from the clinical transplantation database maintained at the Dana-Farber Cancer Institute.

Biostatistical Methods

Patient characteristics were reported descriptively. Subscale and summary scores for the SF-36 and FACT-BMT were calculated according to recommended methods for handling missing data and scaling responses [21-23]. Summary scores on the SF-36 were normalized so that 50 was the mean for the general population, with a standard deviation of 10. On both the SF-36 and the FACT-BMT, a higher score indicated better functioning.

The QOL and symptom scores of patients who reported mild, moderate, or severe overall cGVHD symptoms at enrollment were compared using general linear models. Spearman correlations were calculated between the cGVHD symptom scores and domains of the SF-36, FACT-BMT, other patient-reported measures, and medical chart review. Correlation coefficients were graphed (Figures 1 and 2) with the largest dot representing strong correlation and the smallest dot reflecting lack of correlation. Survival was calculated from the time of enrollment in the cohort. Cox proportional hazards modeling was used to evaluate the predictive ability of patient self-assessed severity of cGVHD at

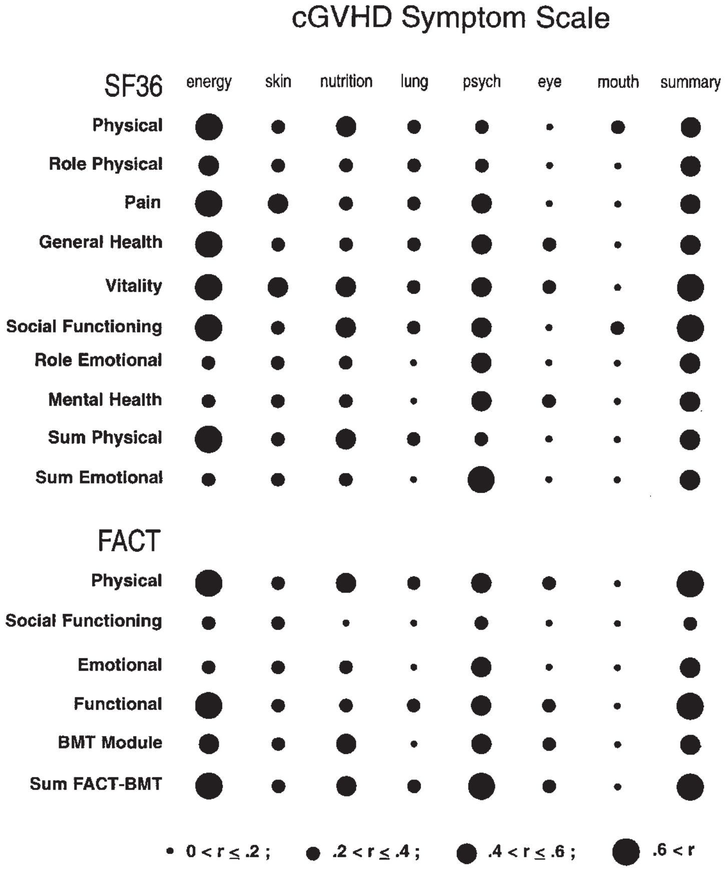

Figure 1. Spearman correlations between the cGVHD symptom scale and the SF-36 and FACT-BMT domains.

the time of enrollment into the cohort. $P < .01$ was considered significant because of multiple testing.

Generation and Psychometric Testing of the Symptom Score

Factor analysis with a promax rotation was used to reduce the 42 items on the symptom scale to 30 items. The categories of "no symptoms" and "symptoms, but not bothered at all" were combined. Twelve items were eliminated through factor analysis: dry mouth, diarrhea, ability to concentrate, nausea, memory loss, anorexia, hair loss, headache, pain, numbness, mouth pain, and sexual difficulties. A score was calculated for each factor by taking the mean of all items completed if more than 50% were answered and normalizing to a 0 to 100 scale. A summary score was similarly calculated considering all the subscales. A higher score indicated more bothersome symptoms.

Enrollment surveys were used for the majority of the psychometric testing. Internal consistency was reported as Cronbach's α. Test-retest reliability was evaluated in 22 patients who were sent the symptom scale 2 to 4 weeks after completing their enrollment surveys. Interclass correlations of the symptom scale were calculated. Convergent and discriminant validity were determined by correlation with clinical data abstracted from the medical record, concurrent patient self-report on the validated instruments (SF-36 and FACT-BMT), and patient self-assessed severity of cGVHD (mild/moderate/severe).

The sensitivity of the QOL instruments to changes in overall health and cGVHD severity was explored [25-29]. Because serial surveys were collected, we enriched for periods of change using the following algorithm: (1) for each patient, a search was made to identify the first instance that he/she reported a change in cGVHD or overall health so that a 6-month change score could be calculated; or (2) for patients who did not report any improvement or worsening over the 24 months of possible observation, the first 2 surveys 6 months apart were used to calculate change scores. Patients were represented only once in each analysis so that observations were independent. Correlation coefficients were calculated between 6-month instrument change scores and changes in overall health or cGVHD severity (categorized as improved, stable, or worsened). Effect sizes were calculated as 6-month change scores (follow-up minus baseline) divided by the standard deviation of the baseline score for each group. Effect sizes were reported so that a positive value indicated an improvement in symptoms, whereas a negative value indicated worsening symptoms.

RESULTS

Patient Characteristics

A total of 158 eligible patients were mailed baseline surveys, and 127 (80%) of 158 returned the surveys. Of the 31 nonrespondents, 15 opted out, 6 died within 2 months of survey mailing, and 10 did not return a survey or opt-out

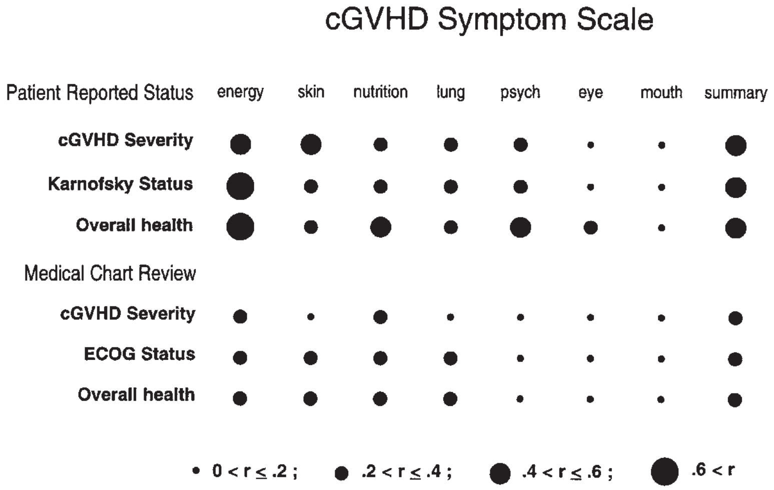

Figure 2. Spearman correlations between the cGVHD symptom scale and patient self-assessment and medical chart review.

card despite 2 mailings and phone contact. Of the 127 respondents, 20 indicated on their baseline surveys that they did not have cGVHD. These patients were excluded from further analysis. Thus, the cGVHD study population consisted of 107 patients. Eighty-three patients (78%) were married or living with a partner, and 63 (59%) had a college or postgraduate degree. Twenty-seven percent of patients were working full time, 15% were working part time, 7% were homemakers, 21% were disabled, and 15% were on medical leave. Self-assessed KPS was 80% to 100% in 62% of patients and 70% or less in 31% of patients; 7% of patients had missing data.

Eleven patients had developed cGVHD following donor lymphocyte infusion for relapse (and were in remission at the time of baseline survey completion). These patients were considered to be without current evidence of disease [30]. Six patients relapsed with their original malignancy after enrollment in the cohort. Tables 1 and 2 summarize additional clinical characteristics of the study population.

Table 1. *Patient Characteristics**

Variable	n = 107
Age, median (range), y	**40 (20-60)**
Male, n (%)	**58 (54)**
Race, n (%)	
White	**97 (91)**
Non-white	**10 (9)**
Married or living with partner, n (%)	**83 (78)**
College or postgraduate degree, n (%)	**63 (59)**
Disease, n (%)	
AML/ALL	**25 (23)**
CML	**48 (45)**
MDS	**11 (10)**
MM	**10 (9)**
NHL/HD/CLL	**10 (9)**
Other	**3 (3)**
Disease stage, n (%)	
Early	**68 (64)**
Intermediate	**31 (29)**
Advanced	**8 (7)**
Donor type, n (%)	
Related	**69 (64)**
Unrelated	**38 (36)**
Acute GVHD prophylaxis, n (%)	
T-cell depletion	**37 (35)**
Immune suppressive medications	**70 (65)**
Year of transplantation, n (%)	
1998-2001	**42 (39)**
1996-1997	**27 (25)**
1994-1995	**15 (14)**
<1994	**23 (22)**

*AML indicates acute myeloid leukemia; ALL, acute lymphoblastic leukemia; CML, chronic myeloid leukemia; MDS, myelodysplastic syndrome; MM, multiple myeloma; NHL, non-Hodgkin's lymphoma; HD, Hodgkin's disease; CLL, chronic lymphocytic leukemia.

Table 2. *Chronic GVHD Characteristics*

Incident/prevalent cases, n (%)	
Newly diagnosed	**44 (41)**
Established diagnosis	**63 (59)**
Severity of cGVHD at enrollment, n (%)	
Patient self-assessed	
Mild	**55 (51)**
Moderate	**39 (36)**
Severe	**13 (12)**
Chart review	
Limited	**35 (33)**
Extensive	**49 (46)**
Could not be determined	**23 (22)**
Current work status, n (%)	
Working full time	**29 (27)**
Working part time	**16 (15)**
Homemaker	**7 (7)**
On medical leave	**16 (15)**
Disabled	**22 (21)**
Other	**17 (16)**
Self-reported KPS, n (%)	
80%-100%	**66 (62)**
<70%	**33 (31)**
Missing	**8 (7)**
Follow-up of survivors, median (range), y	**1.8 (0-2.8)**

Table 3. *Descriptive Statistics of Score Distribution and Intercorrelations for the cGVHD Symptom Scale*

	Energy	Skin	Nutrition	Lung	Psychological	Eye	Mouth	Summary
Items, no.	7	5	5	5	3	3	2	30
Mean	29	18	6	8	18	24	17	17
SD	26	22	14	15	23	29	26	13
Median	25	10	0	0	8	13	0	15
Range	0-93	0-85	0-100	0-95	0-92	0-100	0-100	0-56
Cronbach	0.88	0.81	0.83	0.84	0.79	0.85	0.84	0.90
Floor	17%	33%	67%	63%	43%	39%	61%	3%
Ceiling	1%	2%	1%	1%	1%	2%	2%	1%
Nonresponse	2%	3%	2%	2%	2%	2%	2%	2%
Retest	0.78	0.74	0.83	0.28	0.55	0.87	0.93	0.64
Intercorrelations								
Energy		0.52*	0.47*	0.36†	0.55*	0.21	0.10	.77*
Skin			0.32†	0.28‡	0.40*	0.06	-0.03	.53*
Nutrition				0.26‡	0.31‡	0.20	0.23	.55*
Lung					0.24	0.00	0.14	.42*
Psych						0.23	0.22	.72*
Eye							0.06	.49*
Mouth								.42*

* $P \leq .0001$.
†$P \leq .001$.
‡$P \leq .01$.

Response rates at each time point varied from 78% to 93% of survivors. Specifically, response rates were 38 (93%) of 41 patients at 3 months, 74 (88%) of 84 patients at 6 months, 22 (81%) of 27 patients at 9 months, 52 (78%) of 67 patients at 12 months, 46 (87%) of 53 patients at 18 months, and 34 (83%) of 41 patients at 24 months after study enrollment. Note that only newly diagnosed patients were asked to complete forms at 3 and 9 months, accounting for the smaller denominator.

Development of the Symptom Scale

Through factor analysis, 12 questions were deleted from the original 42-item symptom survey. Seven subscales with 2 to 7 items each were derived, representing domains of skin, eye, mouth, lung function, nutrition, psychological status, and energy. Subscale scores could be calculated for 97% to 98% of patients. Based on pilot testing with 10 patients prior to the study, we estimated that it took 20 to 30 minutes to complete the entire battery of questionnaires (141 questions). Therefore we estimated that the final 30-item symptom scale took approximately 5 minutes to complete.

Reliability

Internal consistency was high (Cronbach's α, 0.79-0.85) and test-retest reproducibility was good for all subscales (0.74-0.93) except psychological status (0.55) and lung symptoms (0.28). Intercorrelations were moderate to high for the energy, skin, nutrition, lung function, and psychological subscales but very low for eye and mouth symptoms. Additional psychometric properties of the subscales are shown in Table 3.

Validity

Convergent Validity. Figures 1 and 2 show the degree of correlation between the symptom subscales and QOL as measured by the SF-36 and FACT-BMT, patient self-assessment of cGVHD severity, KPS, and overall health and physician assessment as obtained by chart review. Consistent with prior hypotheses, the energy factor correlated most closely with physical domains, whereas the psychological scale correlated best with the emotional and mental domains. Correlation with patient self-assessed overall cGVHD severity, KPS, and overall health was moderate, but correlation with information interpreted from the medical record was poor. The eye and mouth subscales were not correlated with summary measures of QOL.

Discriminant Validity. Table 4 and Figures 3 and 4 show that the scales of the SF-36, FACT-BMT, and symptom summary score and the subscales of energy, skin, nutrition, and psychological symptoms could successfully discriminate between patients with self-assessed mild, moderate, or severe cGVHD at baseline. The greatest differences were seen in physical functioning and in the symptom subscales representing energy level, skin involvement, and nutritional status. No differences were seen in the social, emotional, or mental subscales of the SF-36 and the FACT-BMT or in the lung, eye, or mouth components of the symptom scale. Discriminant validity was also supported by the lack of correlation between the cGVHD symptom scale and unrelated domains of the SF-36 and FACT-BMT shown in Figure 1.

No differences in QOL or symptoms were seen when the 35 patients with well-documented limited cGVHD were compared to the 49 patients with extensive disease. However, 23 patients (22%) lacked adequate medical record documentation to allow classification at the time of study enrollment.

Responsiveness to Change. During the study, 14 patients reported worsening cGVHD, 33 improving cGVHD, and 20 stable cGVHD. Changes in cGVHD were not correlated with cGVHD severity at enrollment (r^2 = 0.17, P = .16). Twenty-two patients reported worsening overall health, 21 improving health, and 32 stable health. Change in overall

Table 4. *SF-36, FACT-BMT, and cGVHD Symptom Scores Based on Mild, Moderate, or Severe Patient Self-Rated cGVHD at Enrollment*

	Mild, Mean (SD) (n = 55)	Moderate, Mean (SD) (n = 39)	Severe, Mean (SD) (n = 13)	*P*
SF-36				
Summary physical	42 (12)	36 (10)	28 (8)	.0002
Summary emotional	52 (10)	48 (11)	46 (11)	.07
Physical	67 (28)	57 (26)	26 (17)	<.0001
Role physical	50 (45)	30 (40)	11 (19)	.004
Pain	80 (20)	70 (22)	57 (24)	.002
General health	56 (24)	45 (20)	34 (26)	.004
Vitality	57 (22)	40 (20)	36 (21)	<.0001
Social functioning	75 (28)	66 (23)	51 (30)	.01
Role emotional	84 (31)	69 (39)	38 (43)	.0003
Mental health	77 (17)	73 (19)	69 (21)	.3
FACT-BMT				
Summary FACT-BMT	112 (20)	99 (22)	90 (23)	.0006
Physical	22 (5)	19 (5)	16 (7)	.0003
Social functioning	22 (5)	20 (6)	23 (4)	.1
Emotional	20 (4)	19 (4)	16 (6)	.01
Functional	20 (6)	17 (6)	14 (5)	.001
BMT module	29 (6)	25 (7)	22 (8)	.001
Symptoms				
Summary symptoms	12 (9)	18 (13)	34 (13)	<.0001
Energy	19 (20)	34 (27)	56 (24)	<.0001
Skin	9 (13)	22 (20)	46 (32)	<.0001
Nutrition	3 (8)	4 (7)	24 (31)	<.0001
Lung	6 (16)	8 (14)	15 (15)	.2
Psychological	13 (22)	16 (20)	41 (24)	.0004
Eye	19 (25)	27 (33)	30 (36)	.3
Mouth	16 (23)	15 (27)	29 (35)	.2

health was associated with overall health at enrollment (r^2 = 0.47, *P* = .0001) but not with cGVHD severity (r^2 = 0.17, *P* = .14). In only 58% of cases were identical time points used for both analyses because of our enrichment algorithm. Table 5 shows the effect sizes and correlation coefficients of the change scores for the SF-36, FACT-BMT, and symptom scales with patient-reported changes in cGVHD severity and changes in overall health. Changes in overall health were highly correlated with changes in the subscales of the SF-36 and FACT-BMT but not with changes in the cGVHD symptom scale. In contrast, perceived changes in cGVHD severity correlated with changes detected by the summary cGVHD symptom scale (*r* = 0.33, *P* = .007) but not by the SF-36 or FACT-BMT.

The standard deviation of the baseline summary cGVHD symptom score was 12.9. A distribution-based method based on 0.5 times the standard deviation of the baseline responses was used to estimate a clinically meaningful difference of 6 to 7 points on the cGVHD symptom scale [26,27,31].

Severity of cGVHD and Survival

Survival differed between patients with mild versus moderate or severe cGVHD. Of the 55 patients who rated their cGVHD as "mild" on their baseline surveys, only 3 (5%) died. In contrast, 9 (23%) of 39 and 5 (38%) of 13 patients in the moderate and severe categories, respectively, died. In Cox proportional hazards modeling, patients with moderate and severe cGVHD had similar survival rates, so these subgroups were combined. After a median of 1.8 years of follow-up for survivors, patients with moderate or severe cGVHD had a relative risk of death of 4.2 (95% confidence interval, 1.4-12.8; *P* = .01) compared to patients with mild cGVHD.

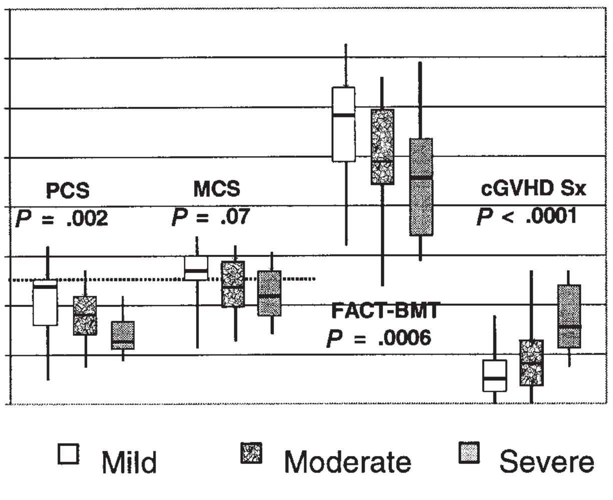

Figure 3. Cross-sectional QOL at enrollment for the SF-36 physical composite scale (PCS), SF-36 mental composite scale (MCS), FACT-BMT summary scale, and chronic GVHD symptom scale (cGVHD Sx) according to whether patient reported mild (n = 55), moderate (n = 39), or severe (n = 13) cGVHD severity. The horizontal bars represent the medians; the upper and the lower bounds of the boxes represent the 75th and 25th percentiles; and the vertical lines indicate the full range of values.

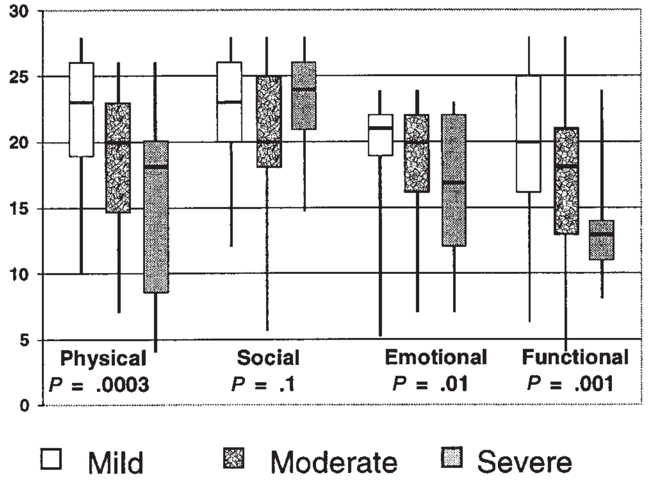

Figure 4. Cross-sectional QOL at enrollment for the domains of the FACT instrument according to whether patient reported mild (n = 55), moderate (n = 39), or severe (n = 13) cGVHD severity. The horizontal bars represent the medians; the upper and the lower bounds of the boxes represent the 75th and 25th percentiles; and the vertical lines indicate the full range of values.

DISCUSSION

We report the development and validation of a 30-item, 7-subscale symptom scale for patients with cGVHD for evaluation of adverse effects on skin, vitality, lung, nutritional status, psychological functioning, eye, and mouth. The psychometric properties of this scale show that reliability, convergent and discriminant validity, and sensitivity to change are adequate to recommend this scale for further development to assess symptomatology in patients with cGVHD. Comparisons with standardized QOL measures show that the symptom scale is more responsive to changes in patient-perceived cGVHD severity than are generic or cancer-specific instruments. Thus, we recommend that this symptom scale be added to one of the validated QOL instruments to fully capture the symptom burden and trajectory of patients with cGVHD.

With patient self-assessed cGVHD as the gold standard, the energy, skin, nutritional status, and psychological subscales showed adequate discriminant validity. Energy, skin, nutrition, and psychological subscales seemed to be driving cross-sectional cGVHD severity assessments, whereas changes in cGVHD severity across time showed greater correlation with changes in energy and lung symptoms. In contrast, eye, mouth, and lung symptoms, although undeniably aspects of cGVHD, were not correlated with patient or physician assessment of cGVHD severity. We hypothesize several reasons for this lack of sensitivity. First, the symptoms addressed on the eye subscale are dry eyes, need to use eye drops frequently, and difficulty seeing clearly, all symptoms that may result from total body irradiation and cataracts, not only the sicca syndrome associated with cGVHD. Indeed, even patients with mild cGVHD reported a high level of bothersome eye symptoms. Second, cGVHD involvement of the lung is relatively uncommon, occurring in approximately 10% to 15% of patients. Its rarity may have masked its importance in determining cGVHD severity in the population even though affected patients are highly symptomatic. A third possibility is that eye and mouth symptoms, although prevalent and bothersome, may not be viewed by patients as determinants of cGVHD severity.

We found that patient self-assessment of cGVHD, in contrast to physician information in the medical record, was reasonably well correlated with patient self-reports of other aspects of health. This finding suggests that direct patient report provides a richer measure of the impact of cGVHD on QOL and functional limitations and may be a more sensitive measure of cGVHD activity than physician assessment.

Several limitations should be noted. First, we achieved an 80% participation rate despite several attempts to enroll patients and a 78% response to the 6-month follow-up

Table 5. *Sensitivity of Scales to Changes in cGVHD Severity and Overall Health**

	Effect Size					Effect Size				
Scale and subscale	Worse cGVHD	Stable cGVHD	Improved cGVHD	R^2	P	Worse Health	Stable Health	Improved Health	R^2	P
n	14	20	33			22	32	21		
SF-36										
Summary Physical	–0.15	–0.24	0.09	.15	.25	–0.26	0.18	0.21	.26	.03
Summary Emotional	–0.13	–0.24	–0.05	.06	.63	–0.59	–0.04	0.25	.34	.004
FACT										
FACT-BMT	–0.23	–0.07	0.05	.16	.21	–0.67	0.11	0.36	.56	<.0001
Symptoms										
Summary symptoms	–0.35	0.26	0.31	.33	.007	0.20	0.16	–0.04	.10	.41
Energy	–0.10	–0.01	0.24	.25	.04	0.00	0.30	–0.03	.01	.94
Skin	–0.34	0.55	0.18	.14	.27	0.19	0.23	–0.05	.12	.32
Nutrition	0.02	0.16	0.31	.05	.71	–0.26	0.19	–0.02	.03	.83
Lung	–0.41	–0.30	0.28	.26	.04	0.21	0.20	0.08	.00	1.00
Psychological	0.0	0.0	0.14	.07	.59	–0.04	0.03	0.02	.03	.81
Eye	–0.21	0.11	–0.17	.01	.97	–0.06	–.035	–0.08	.02	.85
Mouth	0.0	0.21	0.34	.17	.16	0.13	0.23	0.00	.05	.66

*Change scores are divided by standard deviation of baseline scores for each group. SF-36, FACT, and the symptom scale have been adjusted so that positive effect sizes indicate improvement of symptoms and negative effect sizes indicate worsening.

survey. Results may not be representative of our entire cGVHD population, but should have internal validity. Also, despite efforts to have physicians identify affected patients, 20 (16%) of the 127 patients returning baseline surveys reported that they did not have cGVHD. This result could indicate either that patients are unaware they have cGVHD or that their cGVHD resolved by the time we contacted them. Without knowledge as to which possibility was the case, we excluded this group. Second, we would have liked to correlate more objective medical information and laboratory results with QOL and symptom data. However, we did not mandate concurrent physician visits at the time of survey completion and, in retrospect, found that many patients were not seen at our center within 6 weeks of each follow-up period. In addition, abstraction of a core set of medical information from clinic notes proved difficult because of limited documentation. We recommend that future studies investigating the relationship between self-reported QOL and physical manifestations of cGVHD collect such data prospectively and concurrently. Finally, we did not collect survey data on a control group without cGVHD but otherwise matched for time since transplantation and other characteristics. In retrospect, having such a group assembled would have provided valuable comparative information. These studies are ongoing.

In summary, we recommend that either the SF-36 or the FACT-BMT be combined with the cGVHD symptom scale for use in future cGVHD studies when patient QOL and symptoms are being assessed. Given the significant impact of cGVHD on patients' daily experience, we suggest that QOL should be considered an important endpoint in any study of cGVHD interventions, and a 6-to 7-point change on the cGVHD symptom scale from a preintervention survey should be considered clinically meaningful.

ACKNOWLEDGMENTS

This work was supported in part by the Amy Strelzer-Manasevit Scholars Program and the National Institutes of Health Grant No. CA75267-04.

The authors wish to thank Marisa Bueno, BA, and Christina Caron, BA, for their help with study coordination. We especially wish to thank our patients for their participation.

REFERENCES

1. Sullivan KM, Witherspoon RP, Storb R, et al. Alternating-day cyclosporine and prednisone for treatment of high-risk chronic graft-v-host disease. *Blood*. 1988;72:555-561.
2. 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;72:546-554.
3. Wingard JR, Piantadosi S, Vogelsang GB, et al. Predictors of death from chronic graft-versus-host disease after bone marrow transplantation. *Blood*. 1989;74:1428-1435.
4. Loughran TP Jr, Sullivan K, Morton T, et al. Value of day 100 screening studies for predicting the development of chronic graft-versus-host disease after allogeneic bone marrow transplantation. *Blood*. 1990;76:228-234.
5. Duell T, van Lint MT, Ljungman P, et al. Health and functional status of long-term survivors of bone marrow transplantation: EBMT Working Party on Late Effects and EULEP Study Group on Late Effects: European Group for Blood and Marrow Transplantation. *Ann Intern Med*. 1997;126:184-192.
6. 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.
7. Sutherland HJ, Fyles GM, Adams G, et al. Quality of life following bone marrow transplantation: a comparison of patient reports with population norms. *Bone Marrow Transplant*. 1997; 19:1129-1136.
8. Syrjala KL, Chapko MK, Vitaliano PP, Cummings C, Sullivan KM. Recovery after allogeneic marrow transplantation: prospective study of predictors of long-term physical and psychosocial functioning. *Bone Marrow Transplant*. 1993;11:319-327.
9. Majolino I, Saglio G, Scime R, et al. High incidence of chronic GVHD after primary allogeneic peripheral blood stem cell transplantation in patients with hematologic malignancies. *Bone Marrow Transplant*. 1996;17:555-560.
10. Storek J, Gooley T, Siadak M, et al. Allogeneic peripheral blood stem cell transplantation may be associated with a high risk of chronic graft-versus-host disease. *Blood*. 1997;90:4705-4709.
11. Urbano-Ispizua A, Garcia-Conde J, Brunet S, et al. High incidence of chronic graft versus host disease after allogeneic peripheral blood progenitor cell transplantation: the Spanish Group of Allo-PBPCT. *Haematologica*. 1997;82:683-689.
12. Solano C, Martinez C, Brunet S, et al. Chronic graft-versus-host disease after allogeneic peripheral blood progenitor cell or bone marrow transplantation from matched related donors: a case-control study: Spanish Group of Allo-PBT. *Bone Marrow Transplant*. 1998;22:1129-1135.
13. Scott MA, Gandhi MK, Jestice HK, Mahendra P, Bass G, Marcus RE. A trend towards an increased incidence of chronic graft-versus-host disease following allogeneic peripheral blood progenitor cell transplantation: a case controlled study. *Bone Marrow Transplant*. 1998;22:273-276.
14. Ustun C, Arslan O, Beksac M, et al. A retrospective comparison of allogeneic peripheral blood stem cell and bone marrow transplantation results from a single center: a focus on the incidence of graft-vs.-host disease and relapse. *Biol Blood Marrow Transplant*. 1999;5:28-35.
15. Blaise D, Kuentz M, Fortanier C, et al. Randomized trial of bone marrow versus lenograstim-primed blood cell allogeneic transplantation in patients with early-stage leukemia: a report from the Societe Francaise de Greffe de Moelle. *J Clin Oncol*. 2000;18: 537-546.
16. Cutler C, Giri S, Jeyapalan S, Paniagua D, Viswanathan A, Antin JH. Acute and chronic graft-versus-host disease after allogeneic peripheral-blood stem-cell and bone marrow transplantation: a meta-analysis. *J Clin Oncol*. 2001;19:3685-3691.
17. Khouri IF, Keating M, Korbling M, et al. Transplant-lite: induction of graft-versus-malignancy using fludarabine-based non-ablative chemotherapy and allogeneic blood progenitor-cell transplantation as treatment for lymphoid malignancies. *J Clin Oncol*. 1998;16:2817-2824.
18. Giralt S, Thall PF, Khouri I, et al. Melphalan and purine analog-containing preparative regimens: reduced-intensity conditioning for patients with hematologic malignancies undergoing allogeneic progenitor cell transplantation. *Blood*. 2001;97:631-637.

19. McSweeney PA, Niederwieser D, Shizuru JA, et al. Hematopoietic cell transplantation in older patients with hematologic malignancies: replacing high-dose cytotoxic therapy with graft-versus-tumor effects. *Blood*. 2001;97:3390-3400.
20. Lee SJ, Klein JP, Barrett AJ, et al. Severity of chronic graft-vs.-host disease: association with treatment-related mortality and relapse. *Blood*. 2002;100:406-414.
21. Ware JE, Snow KK, Kosinski M, Gandek B. *SF-36 Health Survey: A Manual and Interpretation Guide*. Boston, Mass: The Health Institute, New England Medical Center; 1993.
22. Ware JE, Kosinski M, Keller SD. *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston, Mass: The Health Institute, New England Medical Center; 1994.
23. 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.
24. Shulman HM, Sullivan KM, Weiden PL, et al. Chronic graft-versus-host syndrome in man: a long-term clinicopathologic study of 20 Seattle patients. *Am J Med*. 1980;69:204-217.
25. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989;10:407-415.
26. Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res*. 1993;2:221-226.
27. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol*. 1994;47:81-87.
28. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol*. 1998;16:139-144.
29. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. *JAMA*. 1999;282:1157-1162.
30. Craddock C, Szydlo RM, Klein JP, et al. Estimating leukemia-free survival after allografting for chronic myeloid leukemia: a new method that takes into account patients who relapse and are restored to complete remission. *Blood*. 2000;96:86-90.
31. Sloan J. Asking the obvious questions regarding patient burden. *J Clin Oncol*. 2002;20:4-6.

APPENDIX

Please let us know whether you have been bothered by any of the following problems in the past month.

	Not at all	Slightly	Moderately	Quite a bit	Extremely
SKIN:					
a. Abnormal skin color	0	1	2	3	4
b. Rashes	0	1	2	3	4
c. Thickened skin	0	1	2	3	4
d. Sores on skin	0	1	2	3	4
e. Itchy skin	0	1	2	3	4
EYES AND MOUTH:					
f. Dry eyes	0	1	2	3	4
g. Need to use eyedrops frequently	0	1	2	3	4
h. Difficulty seeing clearly	0	1	2	3	4
i. Need to avoid certain foods due to mouth pain	0	1	2	3	4
j. Ulcers in mouth	0	1	2	3	4
k. Receiving nutrition from an intravenous line or feeding tube	0	1	2	3	4
BREATHING:					
l. Frequent cough	0	1	2	3	4
m. Colored sputum	0	1	2	3	4
n. Shortness of breath with exercise	0	1	2	3	4
o. Shortness of breath at rest	0	1	2	3	4
p. Need to use oxygen	0	1	2	3	4
EATING AND DIGESTION:					
q. Difficulty swallowing solid foods	0	1	2	3	4
r. Difficulty swallowing liquids	0	1	2	3	4
s. Vomiting	0	1	2	3	4
t. Weight loss	0	1	2	3	4
MUSCLES AND JOINTS:					
u. Joint and muscle aches	0	1	2	3	4
v. Limited joint movement	0	1	2	3	4
w. Muscle cramps	0	1	2	3	4
x. Weak muscles	0	1	2	3	4
ENERGY:					
y. Loss of energy	0	1	2	3	4
z. Need to sleep more/take naps	0	1	2	3	4
aa. Fevers	0	1	2	3	4
MENTAL AND EMOTIONAL:					
bb. Depression	0	1	2	3	4
cc. Anxiety	0	1	2	3	4
dd. Difficulty sleeping	0	1	2	3	4

Biol Blood Marrow Transplant 22 (2016) 752–758

ELSEVIER

Biology of Blood and
Marrow Transplantation

journal homepage: www.bbmt.org

ASBMT™
American Society for Blood and Marrow Transplantation

Content Validity of the Lee Chronic Graft-versus-Host Disease Symptom Scale as Assessed by Cognitive Interviews

Emily C. Merkel [1], Sandra A. Mitchell [2], Stephanie J. Lee [3,*]

[1] *Department of Medicine, School of Medicine, University of Washington, Seattle, Washington*
[2] *Division of Cancer Control and Population Sciences, Healthcare Delivery Research Program, Outcomes Research Branch, National Cancer Institute, Bethesda, Maryland*
[3] *Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington*

Article history:
Received 20 November 2015
Accepted 29 December 2015

Key Words:
Allogeneic hematopoietic cell transplantation
Chronic graft-versus-host disease
Patient-reported outcome
Symptoms
FDA qualification
Cognitive interview

ABSTRACT

The Lee Chronic Graft-versus-Host Disease (cGVHD) Symptom Scale has been recommended for use by the 2005 and 2014 National Institutes of Health (NIH) Consensus Conferences to capture cGVHD symptoms. Although the cGVHD Symptom Scale was previously validated, this study aims to reexamine the instrument's content validity by exploring the clarity, comprehensibility, relevance, and ease of use in a contemporary cGVHD sample, toward Food and Drug Administration (FDA) qualification of this patient-reported outcomes (PRO) instrument as a drug development tool. Attaining FDA qualification means that an instrument has been judged to be a reliable and valid measure of clinical benefit. Twenty adult patients with a median age of 58 year (range, 31 to 79 years) participated. The median duration of cGVHD was 33 months (range, 0 to 134.4 months), and current NIH severity score was mild in 1 patient, moderate in 10 patients, and severe in 9 patients, with a median of 5.5 treatments (range, 0 to 14) ever used for cGVHD. The median summary score was 23 (range, 8 to 51), and the median time to complete the scale was 2 minutes, 7 seconds (range, 1 minute, 8 seconds to 4 minutes). Symptoms of cGVHD were well captured on the Lee cGVHD Symptom Scale, although 4 additional symptoms/signs were mentioned by 15% of the participants. Participants mostly reported that item wording was clear and provided accurate definitions of specific terminologies; however, 7 participants (35%) reported finding 1 or more items in the skin domain unclear, reporting, for example, that rashes and itchy skin seemed synonymous. Two of 19 participants (10.5%) described how their answers would have changed had they been asked about their symptoms within the past month instead of within the past week, owing to recently resolved symptoms. All participants were able to accurately explain the concept of "bother" in their own words and distinguish it from symptom severity or other related symptom attributes. In summary, participants found the Lee GVHD Symptom Scale to be a comprehensive and understandable way to report their cGVHD symptom experience. Future work will focus on options for the recall period, the phrasing of skin items, and whether some very rare symptoms (eg, feeding tube, use of oxygen) should continue to be a part of the scale.

INTRODUCTION

Chronic graft-versus-host disease (cGVHD) is a major complication of allogeneic hematopoietic stem cell transplantation (HCT) that is associated with decreased quality of life, impaired functional status, continued need for immunosuppressive medication, and increased mortality [1,2]. A recent National Institutes of Health (NIH) Consensus Conference proposed various tools for standardizing diagnosis, scoring, histopathology, biomarker assays, response assessment, and conduct of clinical trials; patient-reported measures are included in these recommendations [3,4].

In 2002, the development and validation of the Lee cGVHD Symptom Scale to measure symptoms in outpatients age >18 years with cGVHD was reported [5]. This instrument, recommended by the 2005 and 2014 NIH Consensus Conferences [4,6], is now commonly used to evaluate symptoms in clinical practice and in trials of new therapies for cGVHD. The scale contains 30 items in 7 subscales (skin, eye, mouth, lung, nutrition, energy, and psychological). Patients report

Financial disclosure: See Acknowledgments on page 757.

* Correspondence and reprint requests: Stephanie J. Lee, MD, MPH, Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, D5-290, Seattle, WA 98109.
E-mail address: sjlee@fredhutch.org (S.J. Lee).

http://dx.doi.org/10.1016/j.bbmt.2015.12.026
1083-8791/© 2016 American Society for Blood and Marrow Transplantation.

their level of symptom "bother" over the previous month on a 5-point Likert scale: not at all, slightly, moderately, quite a bit, or extremely (Appendix 1). Subscale scores and the summary score range from 0 to 100, with a higher score indicating worse symptoms. A clinically meaningful difference of 6 to 7 points has been suggested for the summary score [1].

Given the the significant impact of cGVHD on patient symptoms after HCT, it is crucial that new treatments seeking Food and Drug Administration (FDA) approval be evaluated in light of their ability to improve the common symptoms of cGVHD as well as overall patient function [7]. The FDA has not yet qualified a patient-reported outcome (PRO) measure to assess whether potential treatments for cGVHD improve patient symptoms, although helping patients "live better" is one of the criteria for FDA approval, along with "living longer."

For the FDA to qualify the Lee cGVHD Symptom Scale as a drug development tool, it must perform a rigorous evaluation of whether the Scale is a reliable and valid measure of clinical benefit in a particular context of use [8]. To date, the FDA has qualified only one PRO instrument, a tool recently qualified for measuring exacerbations in chronic obstructive pulmonary disease [9]. Qualifying an instrument means that sufficient quantitative and qualitative evidence concerning the measurement properties has been presented to the FDA such that the measure can be accepted to support a labeling claim [8,10].

Qualitative evidence of content validity is an important component of the evidence dossier for FDA qualification. Evidence in support of the content validity of a measure of cGVHD symptoms should include empiric evidence that the measure is relevant and comprehensible to patients with cGVHD. Although patient input informed the original development of the Lee cGVHD Symptom Scale, this development occurred between 1998 and 2000 [5]. The aim of the present study was to reassess the instrument's content validity using one-on-one cognitive interviews with 20 patients living with cGVHD.

METHODS

Participants and Data Collection

Between June and August 2015, 2 of the authors (S.J.L. and E.C.M.) conducted one-on-one cognitive interviews with 20 HCT survivors with cGVHD who were attending the Long-Term Follow-Up (LTFU) Clinic at the Seattle Cancer Care Alliance. Inclusion criteria included adults with active cGVHD who could communicate in English. Patients who were eligible and gave written consent participated in a 20- to 30-minute semistructured in-person interview that was audio-recorded.

The scripted, cognitive interview was based on a framework developed by the International Society for Pharmacoeconomics and Outcomes Research evaluating existing PRO instruments and their measures [11-13]. Interview questions and follow-up probes were based on the principles of cognitive interviewing articulated by Willis [14]. Interview questions explored whether the Lee cGVHD Symptom Scale accurately reflected the patient's experience of cGVHD symptoms and the impact of cGVHD on his or her life, as well as the comprehensibility, relevance, and ease of use of the response choices and recall period (Appendix 2). Recruitment continued until saturation was reached (ie, further interviews did not produce any new relevant themes or categories).

Patients were first asked to describe current and past cGVHD symptoms, to provide researchers with an unbiased list of possible symptoms against which to compare the items in the symptom scale. The participant then completed the Lee cGVHD Symptom Scale during the interview, and was asked to identify any items that were confusing or challenging to answer. Because we were interested in exploring the effects of a shorter recall period, the standard recall period of 1 month was modified in the questionnaire instructions from "the past month" to "the past week." The interviewers (E.C.M. and S.J.L.) probed the relevance and clarity of the 30 items through open-ended questions. Each respondent was also interviewed in detail about the clarity and relevance of a subset of 5 items. Items were assigned to each interview to ensure that all scale items would receive a detailed evaluation by at least 3 respondents. The semistructured interview questions also addressed comprehensibility, relevance, recall period, and understanding of the concept of "symptom bother." Interviewers supplemented the cognitive interview script by asking follow-up questions to further probe participant responses. Participants then completed 2 questions to assess their self-perceived overall cGVHD severity, along with a brief demographic questionnaire. The entire process took 15 to 25 minutes.

Following enrollment, participants' charts were reviewed to collect details about previous treatments and current cGVHD organ involvement using the 2014 NIH consensus criteria for diagnosis and scoring [15].

This study was approved by the Institutional Review Board of the Fred Hutchinson Cancer Research Center. Each participant provided signed written consent before the interview and verbally confirmed consent for the interview to be audio-recorded. The participant was given a $20 gift card after completion of the interview.

Data Analysis

Inductive thematic analysis was performed by 2 of the authors (E.C.M. and S.J.L.) to develop and iteratively modify a codebook. Interviews were transcribed verbatim, and transcripts were then coded line-by-line to facilitate analysis and identification of themes. In this report, illustrative quotes are provided to supplement narrative descriptions, with the participant identification number noted in parentheses.

Scores were calculated following the developer's instructions (Appendix 3). Specifically, the subscale scores (skin, eye, mouth, lung, nutrition, energy, and psychological status) were calculated if $\geq$50% of items in the subscale were completed. Note that the instrument is formatted for ease of completion, but the order does not exactly match the subscales, which were determined by factor analysis during development. The summary score is the mean of the subscales when $\geq$50% of the subscales were completed. The theoretical range for each subscale and the summary score was 0 to 100, with higher scores indicating greater symptom bother. Additional quantitative data were drawn from interview transcripts, Lee cGVHD Symptom Scale scores, and chart abstractions. Quantitative analysis was performed using the CORR procedure in SAS version 9.4 (SAS Institute, Cary, NC). Spearman correlation coefficients were calculated between the Lee cGVHD Symptom Scale scores and the cGVHD organ severity scores, derived through chart abstraction and using the 2014 NIH consensus criteria.

RESULTS

Participant Characteristics

A total of 20 patients (11 males [55%]; median age, 58 years; range, 31 to 79 years) were enrolled between June and August 2015. Three participants (15%) were racial or ethnic minorities. Among the participants who self-identified as a racial or ethnic minority, 1 identified as black, 1 identified as Asian, and 1 selected more than 1 race. Sixteen participants (80%) had a college or postgraduate degree, and 16 (80%) were married or living with a spouse or partner. Six participants (30%) were working full time; 3 (15%), part time. Six participants (30%) were retired, and 3 (15%) were disabled and unable to work.

All participants had received a peripheral blood stem cell transplant. Eighteen participants (90%) had a history of acute GVHD, and 18 (90%) had an established diagnosis of cGVHD before the clinic visit at which they were interviewed; 2 participants' initial diagnosis of cGVHD was confirmed at the clinic visit and before the interview took place. The median duration of cGVHD was 33 months (range, 0 to 134.4), and the global severity of cGVHD using 2014 NIH consensus scoring was mild in 1 participant, moderate in 10 participants, and severe in 9 participants. Participants had received a median of 5.5 treatments (range, 0 to 14) for their cGVHD.

The demographic, clinical, and cGVHD characteristics of the participants are summarized in Tables 1 and 2.

Lee CGVHD Symptom Scale

The duration of the interview was a median of 14 minutes and 48 seconds (range, 6 minutes, 38 seconds to 24 minutes, 18 seconds). The median time to complete the 30 items of the

Table 1
Participant Characteristics (n = 20)

Characteristic	Value
Male sex, n (%)	11 (55)
Age, yr, median (range)	58 (31-79)
Caucasian, n (%)	17 (85)
Non-white, n (%)	3 (15)
Married or living with partner, n (%)	16 (80)
College or postgraduate degree, n (%)	16 (80)
Employment status, n (%)	
Working full time	6 (30)
Working part time	2 (10)
In school full time	1 (5)
Homemaker	4 (20)
Retired	6 (30)
Disabled, unable to work	3 (15)
Unemployed, not looking for work	1 (5)
Other	2 (10)
Underlying disease, n (%)	
AML	4 (20)
ALL	5 (25)
MDS	7 (35)
HD	1 (5)
Other	3 (15)
Stage of disease at transplantation, n (%)	
Early	11 (55)
Intermediate	9 (45)
Advanced	0
Donor type, n (%)	
Related	9 (45)
Unrelated	11 (55)
Matched	15 (75)
Partially mismatched	5 (25)
Stem cell source, n (%)	
Peripheral blood stem cells	20 (100)

AML indicates acute myelogenous leukemia; ALL, acute lymphoblastic leukemia; MDS, myelodysplastic syndrome; HD, Hodgkin disease.

Lee cGVHD Symptom Scale was 2 minutes, 7 seconds (range, 1 minute, 8 seconds to 4 minutes; n = 16). Timing data for the other 4 participants were excluded because their completion times had inadvertently included the sociodemographic questions. All study participants reported that the 30 items of the symptom scale could be completed with minimal burden. The median summary score was 23 (range, 8 to 51). Figure 1 shows the percentage of participants endorsing any subscale symptoms and the median and range of each subscale score among symptomatic patients.

Table 2
Characteristics of cGVHD (n = 20)

Characteristic	n (%)
Newly diagnosed cGVHD	2 (10)
Established cGVHD diagnosis	18 (90)
History of acute GVHD diagnosis	18 (90)
Self-assessed cGVHD severity	
Mild	9 (45)
Moderate	8 (40)
Severe	3 (15)
cGVHD severity based on 2014 NIH organ scoring	
Mild	1 (5)
Moderate	10 (50)
Severe	9 (45)
2014 NIH organ score ≥1 per medical records	
Skin	14 (70)
Eye	17 (85)
Mouth	12 (60)
Lung	12 (60)
Gastrointestinal	1 (5)
Liver	0
Joint	9 (45)
Genital (women only)	3 (33)

Several patterns emerged when comparing the Lee cGVHD Symptom Scale domain scores with the NIH consensus scoring system based on chart abstraction for the same domains, participants' scoring on the energy domain on the Lee cGVHD Symptom Scale was strongly correlated ($r = 0.65$; $P = .002$) with their NIH calculated overall cGVHD severity, making energy the most predictive domain for objective overall cGVHD severity. The Lee cGVHD Symptom Scale domain score with the strongest correlation to the corresponding NIH severity score for that domain was the mouth ($r = 0.63$; $P = .003$), and 100% of participants endorsed symptoms in this domain. The eye domain also had a moderately strong correlation between the Lee cGVHD Symptom Scale Score and the NIH domain score ($r = 0.52$; $P = .019$), and 18 participants (90%) endorsed eye symptoms.

Inclusiveness of Symptom Scale Items

Before completing the Lee cGVHD Symptom Scale, participants were asked to describe their current and past GVHD symptoms. This information provided an unbiased list of possible symptoms against which to compare the items in the symptom scale. The cGVHD symptoms that participants reported spontaneously were well captured on the symptom scale, with mouth and eye symptoms the most discussed. However, edema/swelling, vaginal, liver, and fingernail symptoms (items not on the scale) were mentioned as symptoms by 3 participants (15%) each. In the interviews, several participants highlighted that vaginal symptoms are often left out of the conversation about cGVHD, and expressed a desire to see vaginal symptoms better addressed:

> "I've had vaginal GVH since early on and it's just…like I know the team would talk about there was somebody else to deal with girl parts, and I'm a nurse, and I feel like that needs to be talked about more openly. Like it's just kind of left out, so unless I bring it up, it gets skipped over. And so I'm okay with bringing it up, but I feel like a lot of people aren't, and so having stuff going on in your sexual or intimate life can be a really big deal, and it just might mean that people aren't as lucky as I am to have people to talk to about it" (11).

Participants noted that although elevated liver function test results due to cGVHD were not included in the symptom scale, these abnormalities did not have a significant impact on their quality of life, because they caused them minimal to no bother.

> "Liver function is the one that jumps out at me. Um, you know it was never annoying but something that showed up in the labs that they felt the need to address with medication. I mean, it was another pill, which I didn't love, but that was essentially a nonissue" (7).

Clarity and Relevance of Symptom Scale Items

Skin domain

During the course of the interview or immediately after completing the symptom scale, seven participants (35%) indicated that they found 1 or more items in the skin domain unclear. For example, 2 participants (10%) believed that rashes and itchy skin were synonymous. Five participants (25%) spoke of difficulties in assigning their skin symptoms to only 1 of the available categories of cGVHD skin characteristics, specifically abnormal skin color, rashes, thickened skin, sores on skin, or itchy skin.

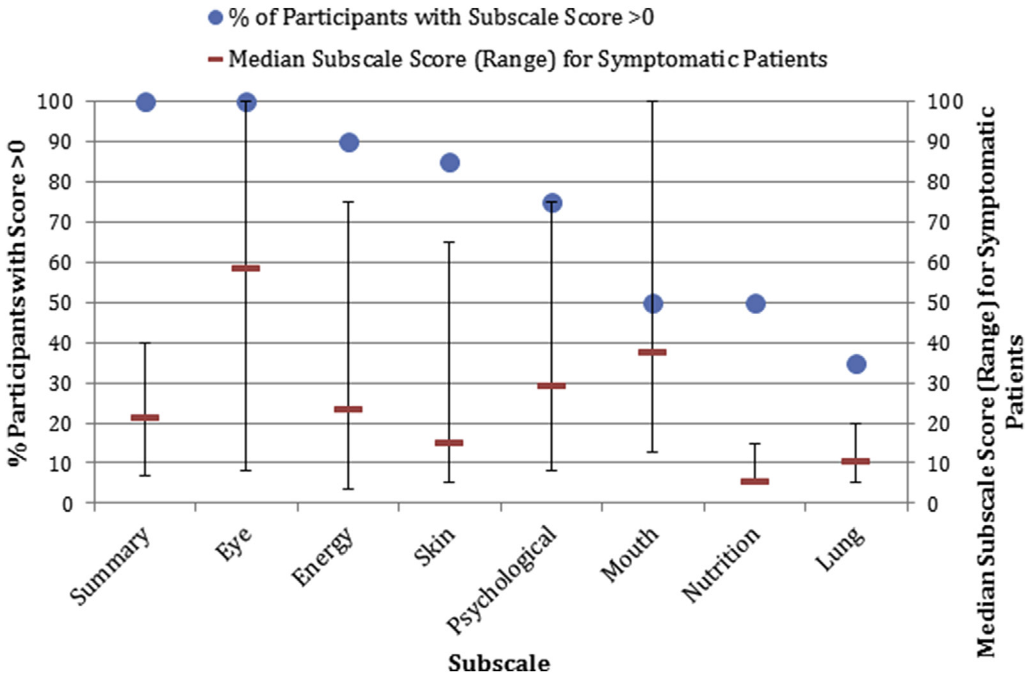

Figure 1. Proportion of respondents endorsing any subscale symptoms, and the median and range of each subscale score among symptomatic patients.

> "I don't really know where my leg lesions fall in. Maybe they fall under one of these; maybe it's here [gestures to skin section of survey]. They're not really sores, though… maybe abnormal skin color" (8).

> "I have a couple things on my skin that aren't really sores. You know, they were…kind of like little spots, and now they've turned into sort of scaly little lumps, so they kind of don't fit into any of these, but they're there" (11).

> "If I were to look at this, I kind of equate this with this [points to rashes and itchy skin]. Yeah, so there's some bleeding over there. But no, I guess you could have a rash that wasn't itchy. I don't know" (15).

Need to use oxygen and receiving nutrition from an intravenous line or feeding tube

In addition, 2 (10%) participants mentioned they found the items "need to use oxygen" and "receiving nutrition from an intravenous line or feeding tube" not meaningful, citing their extreme nature.

> "Receiving nutrition from an intravenous line or feeding tube. I don't think that's that; that's kind of way out there. It seems like it's not part of—I mean, these are things that…just seems like a real—like you're in the hospital doing that" (1).

> "I don't know about the need to use oxygen. That seems really extreme" (4).

Psychological domain

Three participants (15%) noted that answering the mental and emotional questions regarding symptoms of depression and anxiety was more challenging for them because those symptoms might have non-GVHD causes or because of difficulty pondering or evaluating these particular emotions. All participants understood what was meant by the questions, however.

> "I think that mental and emotional stuff is harder to measure, harder to evaluate" (6).

> "So I'm saying no to these 3 things [points to depression, anxiety, and difficulty sleeping] as a consequence of GVHD. I have—there are other reasons I might be having them, okay?" (18).

> "I just don't like the mental and emotional. Sometimes I don't like to answer those…it's perfectly worded" (20).

Overall clarity and relevance

A majority of the participants reported that item phrasing, including the response choices, was clear for all items, and they were able to give accurate explanations of the symptom terms on which they had received in-depth instruction. In the interviews, several participants noted the comprehensiveness and relevance of the items included in the scale and recognized that the items on the scale were common cGVHD symptoms even if they had never personally experienced those symptoms.

> "I'm not experiencing problems with breathing, but just because I'm not doesn't mean someone else isn't" (5).

> "Well, some of them don't apply to me. Loss of sleep—I don't have loss of sleep, so I just marked zero on them, but I take it they're all common symptoms of GVHD. I just don't have all of them" (13).

Recall Period

Seventeen of 19 participants (89.5%) said that their answers would not change if asked about their symptoms within the past month instead of within the past week.

Participants comprehended that the recall period "within the past week" was included in the recall of the past month.

> "It probably wouldn't change, because the past week is included in the past month, isn't it? So, I guess no, probably not" (4).

Two of 19 participants (10.5%) said that their answers would change if asked about their symptoms within the past month instead of within the past week. When asked to clarify why their answers would change and what specific symptoms would change, both noted recently resolved symptoms.

> "Well, I can mainly think of the cough, because that was—dealing with the cough a month ago. Yeah, that would be the main one" (18).

Bother

Participants did not report any confusion about the concept of bother, defining it within 4 general categories: an irritant or annoyance, an interruption, a deviation from the norm, or a discomfort.

> "I'm still struggling to find normal—normalcy, what that means to me now, and just think I'm on that path to kind of a routine and something happens that I don't expect or can't predict. That's what I would consider bothersome. Usually I can anticipate that things are going to happen…I kind of deal with it a little differently, but when its unpredictable, it happens and it's just kind of a surprise, it's a bit of a bother—a little irritant." (5)

> "It's not anything that's life-threatening, but it's just annoying" (7).

> "Just that you notice it to a point where you feel like it's kind of an annoyance or its abnormal or, you know, not what you were you used to prior to…the transplant" (8).

> "It just interferes with my daily life a lot…And then the other thing is just related to GVH and immune suppression—feeling like I can't do, you know, a lot of things I'd normally do" (11).

> "if it makes me uncomfortable…That's what it means to me, if it makes me uncomfortable" (3).

Overall Impressions of the Scale

All participants (100%) described a favorable overall impression of the scale, citing its brevity, clarity, and ease of completion. A majority of respondents explained that the scale, response choices, and directions for completion were already as clear as possible and were unable to offer any suggestions for change.

> "It felt really reasonable compared to a lot of the surveys I fill out. Pretty quick" (7).

> "It was good. It was short" (4).

> "I really don't know of any way you could make it easier" (3).

> "A piece of cake…it's very easy" (20).

> "The directions are very clear, and the selection is in terms of assessing how you feel about each item is pretty spot on" (5).

Three participants (15%) noted that they particularly liked the verbal descriptors for the response options and the 5-point Likert scale, highlighting that more than 5 points would have been too challenging and that qualitative descriptors are preferable to numbers. The exception to this was 1 participant who was interested in knowing how the investigators wanted him to map his symptoms to the qualitative scale.

> "Well, I think this scale is not bad because it spells out things rather than putting them on a scale of 1 to 10 or whatever. I find those not very helpful, because the numbers are relative. How would you know what the end of the scale is if you haven't experienced it, so you don't know really where you are on the scale. But if you say not at all or slightly or moderately, I think that's easier to answer those questions" (13).

> "I think the 4, well 5 columns actually, are really good. I wouldn't do any more…This seems to break it down, so I particularly like that, and I thought the categories were good" (16).

> "I think there's going to be a level of subjectivity in these categories, just because I think some people by nature underestimate the impact of things and some people overestimate maybe… I don't know how you'd do this—but I could see just a sentence or two describing each one of these. You know, if it was more—here's the engineer in me coming out—but if it was more quantitative, you know, if moderately meant, you know, you think about it 3 times a day and extremely meant that it has a significant impact on your daily activity or something like that. That kind of a definition might be in making sure that I was aligned with how these terms are generally thought of" (7).

Participants also appreciated being asked to provide their own perspectives of the cGVHD symptom experience.

> "The subtext for all of this stuff is, well, you're still alive, but when you go into this, you don't know how it's going to be afterward, and it's a pretty big deal" (11).

When asked if there was anything that could be done to make the Lee GVHD Symptom Scale easier to complete, 2 participants (10%) suggested offering the symptom scale online, and 2 participants (10%) suggested adding a section for participants to write in additional free text if they wanted to qualify any of their responses or indicate any unsolicited symptoms.

DISCUSSION

The participants' responses to the Lee cGVHD Symptom Scale were overwhelmingly positive, highlighting the scale's clarity, relevance, and acceptability. Participants expressed interest in and gratitude for being asked to provide their own assessment of their cGVHD symptoms and how those symptoms impact their lives.

Participants were consistent in their understanding of bother when prompted for a definition before seeing the Lee cGVHD Symptom Scale. This suggests a universal, culturally shared understanding of this concept, at least among English speakers. Participants also reported that it was a reasonable measure against which to consider the impact of cGVHD symptoms on their day-to-day lives. However, whether symptoms should be scaled for intensity, frequency, interference, bother, or all of these dimensions, remains a topic of debate in the literature [16-19].

The majority of patients clearly understood the recall period ("within the past week") and said that their answers would not change if asked about their symptoms within the past month instead of within the past week. Although the scale was originally validated with a 1-month recall period, the participants' responses support the acceptability of the shorter recall period.

When the Lee cGVHD Symptom Scale was first developed and validated in the late 1990s, severe manifestations of cGVHD were more common. Two of the symptom scale items, "need to use oxygen" and "receiving nutrition from an intravenous line or feeding tube," reflect that reality. As transplantation regimens and GVHD prophylaxis and treatment continue to evolve, these manifestations are becoming increasingly rare. Two participants noted that these symptoms seemed to be extreme compared with the remainder of the scale, and it is worth evaluating whether or not these symptoms are sufficiently common so that they should continue on the scale moving forward. Although our sample was small, none of our study participants endorsed these problems. Using larger populations, we plan to examine the presence of floor effects for these items and consider the potential impact of removing them from the scale.

Participants generally understood the meaning of the item phrasing in the skin domain; however, several participants found it difficult to report their own skin symptoms using the available items, explaining that their skin symptoms could have been described using various terminologies. It is interesting to note that when prompted to define the items in the skin domain, all participants were able to do so without hesitation. Options that could be explored include providing definitions (eg, "rashes, including bumps, scaling, roughness or other changes in skin texture/feel") or pictures of skin manifestations or expanding the skin domain to include additional specific or nonspecific cutaneous symptoms. Another possibility is to incorporate free text fields into the Lee cGVHD Symptom Scale as single items, thus inviting respondents to list any cGVHD symptoms they feel are not adequately captured by the scale and grade their associated degree of bother. These free text descriptions, using their own words, would then be presented each time they completed the symptom scale, although this could complicate scoring and interpretation of the summary score.

Although all cGVHD symptoms were generally well represented on the Lee cGVHD Symptom Scale, edema/swelling, and vaginal, liver, and fingernail symptoms were each mentioned by 3 participants as cGVHD symptoms that did not appear among the scale items. The symptom of edema and swelling was mentioned by 3 different participants, who described 3 different phenomena that may or may not have been related to cGVHD. This inconsistency in definition requires further exploration before being considered for inclusion in the symptom scale.

Several women mentioned vulvovaginal symptoms without prompting, and several more endorsed vaginal symptoms when asked. In addition, the assessment of vaginal cGVHD symptoms is now included in the NIH consensus scoring and evaluation as an exploratory measure, supporting the addition of such symptoms to the Lee cGVHD Symptom Scale. More qualitative work is needed to assess whether women are able to discern vaginal symptoms from cGVHD as opposed to vaginal symptoms from menopause and hormonal changes, considering that some manifestations may overlap. Vulvovaginal symptoms caused by infections or estrogen deficiency would confound the assessment of cGVHD treatment. No men spontaneously mentioned genital symptoms.

Elevated liver function test results are a common complication of cGVHD. Several participants noted that they had had elevated liver function test values, but participants acknowledged that this did not bother them. Therefore, liver abnormalities are not a good candidate for addition to the Lee symptom scale. Similarly, fingernail symptoms were mentioned by several participants, but were not described as bothersome.

Limitations of this study include the small cohort and a sample drawn from a relatively homogenous population of outpatients at a single center. We enrolled only 20 participants because we reached saturation about interview topics. Racial/ethnic minorities were underrepresented, and participants were very well educated. As with all PRO measures, our instrument was designed for use in patients who are capable of meaningfully self-reporting. The scale was not intended for cognitively impaired patients, children age <12 years, or proxy reporting of symptoms. In-trial guidance has been published to address situations in which study participants have a PRO endpoint due for collection and have cognitive impairment [20].

In summary, the results of this study support the content validity of the Lee cGVHD Symptom Scale, a PRO measure of cGVHD symptom bother. Study participants believed that the scale captured almost all of their symptoms and was generally clear. They also appreciated the brevity, and offered few suggestions for improvement. Future work will focus on additional evaluation of construct validity and responsiveness to change. Although this testing was previously done during development, it has been almost 15 years since this initial work was completed. Evaluation in larger samples and using modern measurement theory will provide additional evidence of the measurement properties of this instrument and support FDA qualification. Because cGVHD has prominent effects on symptom burden and quality of life, and because symptom improvement is part of the criteria for defining therapeutic response, it is crucial that we have available valid, reliable, and responsive measure of cGVHD bother for use in trials of new therapies for cGVHD.

ACKNOWLEDGMENTS

The authors thank Nancy Kline Leidy, PhD, RN, for her contributions to the design of the study, and Tracey Godlewski, PA-C, Carina Moravec, ARNP, Kristen Kindsvogel, ARNP, and Mary Flowers, MD, for their help with enrollment.

Financial disclosure: Funding for this study was provided by the American Cancer Society (ECM) and CA118953 and CA163438 (SJL). The Chronic GVHD Consortium U54 CA163438 is part of the NCATS Rare Diseases Clinical Research Network (RDCRN). RDCRN is an initiative of the Office of Rare Diseases Research (ORDR), NCATS, funded through a collaboration between NCATS and the National Cancer Institute.

Conflict of interest statement: The authors have no conflicts of interest to report.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.bbmt.2015.12.026.

REFERENCES

1. Lee SJ, Vogelsang G, Flowers ME. Chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2003;9:215-233.
2. Socié G, Ritz J. Current issues in chronic graft-versus-host disease. *Blood*. 2014;124:374-384.
3. Martin PJ, Lee SJ, Przepiorka D, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease, VI: 2014 Clinical Trial Design Working Group report. *Biol Blood Marrow Transplant*. 2015;21:1343-1359.
4. Lee SJ, Wolff D, Kitko C, 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: 2014 Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2015;21:984-999.
5. Lee SJ, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8:444-452.
6. Pavletic SZ, Martin P, Lee S, 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;12:252-266.
7. Lee SJ. Have we made progress in the management of chronic graft-vs-host disease? *Best Pract Res Clin Haematol*. 2010;23:529-535.
8. Hayes RP, Blum SI, Gordon MF, et al. The Patient-Reported Outcome (PRO) Consortium: lessons learned along the path to PRO instrument qualification. *Ther Innov Regul Sci*. 2014;49:132-138.
9. Leidy NK, Murray LT, Jones P, Sethi S. Performance of the EXAcerbations of chronic pulmonary disease tool patient-reported outcome measure in three clinical trials of chronic obstructive pulmonary disease. *Annals Am Thorac Soc*. 2014;11:316-325.
10. Coons SJ, Kothari S, Monz BU, Burke LB. The Patient-Reported Outcome (PRO) Consortium: filling measurement gaps for PRO end points to support labeling claims. *Clin Pharmacol Ther*. 2011;90:743-748.
11. Rothman M, Burke L, Erickson P, et al. Use of existing patient-reported outcome (PRO) instruments and their modification. ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force report. *Value Health*. 2009;12:1075-1083.
12. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity: establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation. ISPOR PRO Good Research Practices Task Force report, part 2: assessing respondent understanding. *Value Health*. 2011;14:978-988.
13. Brédart A, Marrel A, Abetz-Webb L, et al. Interviewing to develop Patient-Reported Outcome (PRO) measures for clinical research: eliciting patients' experience. *Health Qual Life Outcomes*. 2014;12:15.
14. Willis GB. *Analysis of the cognitive interview in questionnaire design*. New York: Oxford University Press; 2015.
15. Jagasia MH, Greinix HT, Arora M, et al. National Institutes of Health Consensus Development Project on Criteria for Clinical Trials in Chronic Graft-versus-Host Disease, I: 2014 Diagnosis and Staging Working Group report. *Biol Blood Marrow Transplant*. 2015;21:389-401.e1.
16. Chang CH, Cella D, Clarke S, et al. Should symptoms be scaled for intensity, frequency, or both? *Palliat Support Care*. 2003;1:51-60.
17. Gawlicki MC, McKown SM, Talbert MJ, Brandt BA. Application of bother in patient reported outcomes instruments across cultures. *Health Qual Life Outcomes*. 2014;12:18.
18. Tishelman C, Degner LF, Rudman A, et al. Symptoms in patients with lung carcinoma. *Cancer*. 2005;104:2013-2021.
19. Wu AW, Dave NB, Diener-West M, et al. Measuring validity of self-reported symptoms among people with HIV. *AIDS Care*. 2004;16:876-881.
20. Kyte DG, Draper H, Ives J, et al. Patient reported outcomes (PROs) in clinical trials: is "in-trial" guidance lacking? a systematic review. *PLoS ONE*. 2013;8:e60684.

Biol Blood Marrow Transplant 26 (2020) 562–567

Biology of Blood and Marrow Transplantation

journal homepage: www.bbmt.org

Survivorship

Reliability and Validity of the Modified 7-Day Lee Chronic Graft-versus-Host Disease Symptom Scale

Christopher Teh, Lynn Onstad, Stephanie J. Lee*

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

Article history:
Received 6 September 2019
Accepted 13 November 2019

Keywords:
chronic graft-versus-host disease
symptoms
quality of life
patient-reported outcomes
survey

ABSTRACT

Chronic graft-versus-host disease (cGVHD) adversely affects patient quality of life, functional status, and survival after allogenic hematopoietic cell transplantation. The Lee Symptom Scale is a 30-item scale developed to measure the symptoms of cGVHD. Although the original 30-item scale uses a 1-month recall period, we tested the reliability and validity of a 28-item scale (deleting 2 items based on supportive care needs rather than symptoms) with a 7-day recall period, a format that is more appropriate for use in clinical trials. Results show the modified 7-day scale is reliable and valid in the modern era and may be used to assess the symptom burden of cGVHD in clinical trials. Using the distribution method, a 5- to 6-point difference (half a standard deviation) is considered clinically meaningful.

INTRODUCTION

Chronic graft-versus-host disease (cGVHD) is a serious iatrogenic complication that affects survivors of allogeneic hematopoietic cell transplant (HCT). Twenty percent to 50% of allogeneic transplant survivors develop cGVHD. The disease results in inflammation, scarring, and organ dysfunction. cGVHD is the most common long-term complication of HCT and is associated with a decreased quality of life, impaired functional status, continued need for immunosuppressive medications, and increased nonrelapse mortality [1].

Published in 2002, the Lee cGVHD Symptom Scale (LSS) was developed to measure symptoms in adult outpatients with cGVHD [2]. The scale contains 30 items grouped into 7 subscales (skin, eye, mouth, lung, nutrition, energy, and psychological) and takes 2 minutes to complete. Patients report how "bothered" they feel about each symptom over the previous month using a 5-point Likert scale from "not at all" to "extremely." A 1-month assessment period was chosen for the original scale to capture symptoms over a period of time because most patients with cGVHD are treated as outpatients, and cGVHD symptoms can wax and wane. Subscales range from 0 to 100, with a higher score indicating worse symptoms. Subscales may be scored if at least 50% of items are answered, and subscales are averaged to calculate the summary score. Readers are cautioned to use the correct scoring algorithm (Table 1) because the headers in the survey do not directly correspond to the subscales. In the original publication a 6- to 7-point change in the summary score suggested a clinically meaningful difference in patient symptomatology.

In 2005 and 2014, the National Institutes of Health (NIH) Consensus Development Project on Criteria for Clinical Trials in cGVHD proposed the LSS as a tool to determine the efficacy of cGVHD treatments [3,4]. The relevance of the scale to modern cGVHD patients was confirmed in a 2016 publication about the content validity of the scale [5]. Even though the LSS is now commonly used to evaluate symptoms in cGVHD prevention [6,7] and therapy trials [8-10], the US Food and Drug Administration and pharmaceutical sponsors prefer a shorter recall period of 7 days. Some have also questioned the inclusion of 2 items (ie, "Receiving nutrition from an intravenous line or feeding tube" and "Need to use oxygen") because they reflect use of supportive care measures rather than symptoms. In general, recall is better over shorter periods that are preferred for symptom assessment but may be influenced by fluctuating symptomatology. A 7-day recall period matches the common quality of life instruments.

The aim of the present study was to reassess the instrument's reliability and validity in the modern era with a 7-day recall period to establish internal consistency of items, show that scores are stable if a patient's condition does not change, and demonstrate convergent and divergent validity. These are important features of any scale used to document effectiveness of treatments. We also evaluated the impact of deleting the 2 questions relating to supportive care on the performance of the scale.

Financial disclosure: See Acknowledgments on page 567.

*Correspondence and reprint requests: Stephanie J. Lee, MD, MPH, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, D5-290, Seattle, WA 98109.

E-mail address: sjlee@fredhutch.org (S.J. Lee).

https://doi.org/10.1016/j.bbmt.2019.11.020
1083-8791/© 2019 American Society for Transplantation and Cellular Therapy. Published by Elsevier Inc.

Table 1
Scoring Algorithm for the mLSS

Subscale Name	Number of Items	Items
Skin	5	a. Abnormal skin color b. Rashes c. Thickened skin d. Sores on skin e. Itchy skin
Eye	3	f. Dry eyes g. Need to use eye drops frequently h. Difficulty seeing clearly
Mouth	2	i. Need to avoid certain foods due to mouth pain j. Ulcers in mouth
Lung	4	l. Frequent cough m. Colored sputum o. Shortness of breath at rest ~~p. Need to use oxygen~~ **aa. Fevers**
Nutrition	4	~~**k. Receiving nutrition from an intravenous line or feeding tube**~~ q. Difficulty swallowing solid foods r. Difficulty swallowing liquids s. Vomiting t. Weight loss
Energy	7	**n. Shortness of breath with exercise** **u. Joint and muscle aches** **v. Limited joint movement** **w. Muscle cramps** **x. Weak muscles** y. Loss of energy z. Need to sleep more/take naps
Psych	3	bb. Depression cc. Anxiety dd. Difficulty sleeping

The 7-day mLSS is a 28-item instrument with 7 subscales (skin, eyes, mouth, lung, nutrition, energy, and psych) containing 2 to 7 items that allow calculation of a summary score. Response options for "Please let us know if you have been bothered by any of the following problems in the past 7 days" range from 0 to 4 (Not at all, Slightly, Moderately, Quite a bit, Extremely). A clinically meaningful difference is considered 5-6 points on the summary score. **Bolded** items are scored under a different subscale than where they are located under the headers in the survey. Items p and k were deleted from the original 30 item scale.

Scoring rules:

a. Note that the subscales do not conform exactly to the headers in the patient survey.

b. Items p and k are deleted from the 7-day version.

c. Subscales may be scored if 50% of more of the items in the subscale are completed.

d: Scores are linearly transformed to a 0-100 scale where 0 means all answered items were a "0" and "100" means that all answered items were a "4."

e. Missing items are not included in the scoring.

f. The summary score is the average of the subscale scores, as long as 4 or more subscales are available.

g. Higher scores indicate more severe symptoms.

METHODS

Sixty-eight participants were included in the study if they were adults aged $\geq$ 18 years, able to communicate in English, diagnosed with cGVHD per the 2014 NIH consensus criteria, and symptomatic with active cGVHD with an NIH score $>$ 1 in one or more organs associated with cGVHD. We anticipated enrolling up to 80 subjects to include at least 40 of those with unchanged cGVHD symptoms between the first and second administration of the LSS. Data collection continued until 40 participants indicated no clinically significant change to their symptoms ("about the same") between their baseline and follow-up surveys. The study was approved by the Fred Hutchinson Institutional Review Board and was conducted between 2016 and 2017. The requirement for written documentation of consent was waived given the minimal risk nature of the study because participants were informed of all components of informed consent, including that they could skip over any topic they wished and that participation was voluntary and would not affect their care.

The LSS survey used in this study was identical to the published 1-month recall period version except the recall period used was "the past 7 days." Participants were given a paper survey at enrollment that they completed in the clinic or mailed back. They then were mailed a second survey to complete and return by mail approximately 1 week after completion of the first LSS. The follow-up survey asked whether their cGVHD symptoms had changed since enrollment, using a 7-point scale: very much worse, moderately worse, a little worse, about the same, a little better, moderately better, or very much better. Patients also completed 1 page of sociodemographic questions with their enrollment survey. On both surveys they answered questions about their overall cGVHD status from the NIH patient self-report cGVHD assessment that included reporting whether they considered their cGVHD "mild," "moderate," or "severe."

Descriptive statistics include patient, transplant, and cGVHD characteristics. The survey was scored according to the published recommendations and excluding the 2 supportive care items in question [2,5]; nonresponse was defined as the inability to calculate a score because of missing data. The test-retest correlation was calculated for the summary score and the 7 subscales and reported as the Spearman correlation coefficient. The intraclass correlations are reported as Cronbach's alphas. Generally, test-retest and Cronbach's alpha values of $>$.7 [11] are considered acceptable.

RESULTS

During the study 68 patients enrolled, and 40 (59%) reported that their symptoms were "about the same" on the follow-up survey. The other 28 patients either failed to complete a follow-up survey (n = 12, 18%) or indicated on the follow-up survey that their cGVHD symptoms had improved (n = 16, 24%). No one reported that their symptoms had worsened. Psychometrics are reported based on the 68 enrollment surveys, whereas the test-retest statistics are based on the 40 participants who reported that their cGVHD had not changed since they first completed the LSS. Sample characteristics are shown in Table 2. Ten participants (14.7%) identified themselves as nonwhite: Asian (n = 3), black (n = 3), Hawaiian Native/Pacific Islander (n = 1), and other (n = 3). Forty-three participants (64.2%) had a college or postgraduate degree, and 43 (64.2%) were married or living with a partner. Twenty-two participants (32.8%) were working full-time and 6 (9.0%) part-time.

The median time elapsed from HCT to the diagnosis of cGVHD was 8.4 months. The median time from HCT to enrollment was 34.9 months (interquartile range, 19.4 to 64), and the distribution of global severity of cGVHD using 2014 NIH consensus scoring was mild (n = 10), moderate (n = 30), and severe (n = 28). NIH severity was higher than participant self-reported severity, which was none (n = 5), mild (n = 31), moderate (n = 29), and severe (n = 2). The most common organs with scores of 2 or higher were skin in 31 participants (45.6%) and eye in 23 (33.8%).

Table 3 summarizes the psychometric properties of the survey for both the 30-item and 28-item scales. Cronbach's alpha was $>$.7 for the energy, skin, eye, and mouth subscores and for the summary score but $<$.62 for nutrition, lung, and psychological scales. No participants endorsed the intravenous or feeding tube item, and 2 reported being "slightly" or "moderately" bothered by needing to use oxygen. Removing these items and recalculating the subscale scores minimally improved the Cronbach's alpha of the nutrition and lung subscales to .61 and .43, respectively when compared with inclusion of all items. Importantly, however, Cronbach's alpha of the summary score remained high and was .84 for the 30-item scale and .85 for the 28-item scale in the present study compared to .90 in the original description. Cronbach's alpha remained .76 to .83 when evaluated in the 3 severity groups separately. The standard deviation was 10.5 for the 30-item scale and 10.7 for the 28-item scale; we estimate that a 5- to 6-point difference is clinically meaningful using the distribution method (half a standard deviation) [12,13].

Table 2
Cohort Characteristics (N = 68)

Characteristic	All Participants (N = 68)	Participants Used for Test-Retest (n = 40)
Median age, yr (IQR)	57.5 (42.5-63.5)	58.5 (40.5-63)
Male sex	41 (60.3)	26 (65.0)
Race/ethnicity		
White	58 (85.3)	37 (92.5)
Asian	3 (4.4)	1 (2.5)
Black	3 (4.4)	1 (2.5)
Hawaiian Native/Pacific Islander	1 (1.5)	0 (0)
Other	3 (4.4)[*]	1 (2.5)[†]
Hispanic	0 (0)	0 (0)
Marital status		
Married/living with partner	43 (64.2)	27 (67.5)
Single, never married	14 (20.9)	7 (17.5)
Divorced, separated	8 (11.9)	6 (15.0)
Widowed	1 (1.5)	0 (0)
Married/not living with partner	1 (1.5)	0 (0)
Missing	(n = 1)	
Education		
Less than college	8 (11.9)	5 (13)
Some college	16 (23.9)	9 (22.5)
College graduate	24 (35.8)	14 (35.0)
Post graduate degree	19 (28.4)	12 (30.0)
Missing	(n = 1)	
Work/school status		
Working or school full time	22 (32.8)	15 (37.5)
Working part time	6 (9.0)	3 (7.5)
Retired	16 (23.9)	10 (25.0)
Disabled, unable to work	12 (17.9)	8 (20.0)
Homemaker	7 (10.4)	3 (7.5)
On medical leave	2 (3.0)	0 (0)
Unemployed, looking for work	2 (3.0)	1 (2.5)
Missing	(n = 1)	(n = 0)
NIH severity at enrollment (patient self-report)		
Mild	36 (53.7)	21 (52.5)
Moderate	29 (43.3)	17 (42.5)
Severe	2 (3.0)	2 (5.0)
Missing	(n = 1)	
NIH severity at enrollment (per NIH criteria)		
Mild	10 (14.7)	5 (12.5)
Moderate	30 (44.1)	16 (40.0)
Severe	28 (41.2)	19 (47.5)
Median time from transplant to cGVHD, mo (IQR)	8.4 (5.4-11.8)	10.0 (5.8-12.3)
Median time from cGVHD diagnosis to enrollment, mo (IQR)	34.9 (19.4-64)	34.9 (19.6-60.5)
Score 2-3 organ involvement		
Skin	31 (45.6)	19 (47.5)
Eye	23 (33.8)	15 (37.5)
Mouth	5 (7.4)	2 (5.0)
Gastrointestinal	1 (1.5)	1 (2.5)
Liver (1 missing)	0 (0)	0 (0)
Lung (1 missing)	7 (10.4)	6 (15.4)
Joint	12 (17.6)	8 (20.0)

Values are n (%) unless otherwise defined. IQR indicates interquartile range.

[*] Asian and Indian, Portuguese, and American.

[†] American.

Table 3
Reliability of the 7-Day LSS (N = 68)

			Nutrition		Lung					Summary	
	Energy	Skin	Original	Modified	Original	Modified	Psych	Eye	Mouth	Original	Modified
Items	7	5	5	4	5	4	3	3	2	30	28
Mean	28.0	15.0	5.0	6.3	3.4	3.9	16.5	44.2	19.9	18.9	19.1
Standard deviation	20.2	17.0	8.5	10.6	6.2	7.4	16.6	28.6	25.0	10.5	10.7
Median	25.0	10.0	0	0	0	0	12.5	41.7	12.5	17.6	18.0
Range	0-85.7	0-70	0-45	0-56.3	0-40	0-50	0-75	0-100	0-100	2.4-43.3	2.4-43.5
Cronbach's α	.85	.74	.57	.61	.40	.43	.57	.83	.71	.84	.85
Floor, n (%)	5 (7.4%)	21 (30.9%)	40 (58.8%)	40 (58.8%)	41 (60.3%)	42 (61.8%)	18 (26.5%)	7 (10.3%)	29 (42.6%)	1 (1.5%)	1 (1.5%)
Ceiling, n (%)	1 (1.5%)	1 (1.5%)	1 (1.5%)	1 (1.5%)	1 (1.5%)	1 (1.5%)	1 (1.5%)	4 (5.9%)	1 (1.5%)	1 (1.5%)	1 (1.5%)
Nonresponse	0	0	0	0	0	0	0	0	0	0	0
Test-retest (n = 40)	.89	.88	.76	.76	.71	.70	.82	.85	.79	.79	.81
Intercorrelation											
Energy		.38[*]	.50[†]	.50[†]	.38[*]	.37[*]	.35[*]	.27	.22	.71[†]	.71[†]
Skin			.01	.01	.11	.13	.27	.07	.14	.46[†]	.46[†]
Nutrition					.24		.21	.19	.46[†]	.56[†]	
Nutrition-modified						.28	.21	.19	.46[†]		.57[†]
Lung							.01	.10	−.03	.29	
Lung-modified							.05	.05	.01		.29
Psychological								.15	.22	.55[†]	.55[†]
Eye									.38[*]	.67[†]	.66[†]
Mouth										.62[†]	.62[†]

[*] $P < .01$
[†] $P < .0001$

All test-retest correlations were at least .7 and ranged from .70 to .89. Correlations were ≥.80 for the energy, skin, psychological, and eye subscales and between .70 and .79 for the nutrition, lung, and mouth subscales. The test-retest correlation for the summary score was .79 for the full scale and .81 for the 28-item version. Compared with the original description of the instrument, test-retest scores were higher or the same for energy, skin, lung, psychological, eye, and summary scores and lower for nutrition and mouth but still adequate for all.

Interclass correlations showed that the energy subscale correlated with all subscales except the eye and mouth subscales. The mouth scores correlated with the nutrition and eye subscales, but overall the subscales were fairly independent. All except lung correlated with the summary score.

Table 4 shows that the LSS scores differed for each subscale ($P < .10$) and for the summary score ($P < .001$) between self-reported and NIH mild versus moderate/severe cGVHD except for 2 items. The lung subscale was not correlated with self-reported cGVHD severity, and the psychological scale was not correlated with the NIH cGVHD severity. Removal of the “need to use oxygen” item did not improve the results of the lung subscale analysis. In 2 other prospective multicenter observational studies conducted from 2007 to 2012 [14] and 2013 to 2017 [15], the rates of any endorsement of the “need to use oxygen” (3.2% and 6.1%) and “receiving nutrition from an IV or feeding tube” (1.2% and 2.2%) items were very low.

Table 4
cGVHD Symptoms by Self-Reported and NIH Calculated cGVHD Severity At Enrollment

	Self-Reported cGVHD Severity		
Symptoms	None[*]/Mild (n = 36)	Moderate/Severe[†] (n = 31)	*P*[‡]
Energy	19.0 (13.1)	38.1 (22.0)	<.001
Skin	9.2 (11.9)	21.6 (19.5)	.003
Nutrition	3.2 (8.0)	7.0 (8.6)	.070
Nutrition–modified[‡]	4.1 (10.1)	8.7 (10.8)	.069
Lung	2.6 (4.4)	4.2 (7.7)	.315
Lung-modified[§]	3.0 (5.1)	4.9 (9.4)	.304
Psychological	12.3 (13.7)	21.4 (18.4)	.023
Eye	32.9 (26.3)	57.0 (25.8)	<.001
Mouth	14.6 (17.0)	25.8 (30.9)	.076
Summary	13.4 (7.9)	25.0 (9.6)	<.001
Summary-modified[‡]	13.6 (8.0)	25.3 (9.9)	<.001
	cGVHD Severity per NIH Criteria		
	Mild (n = 10)	Moderate/Severe (n = 58)	*P*[‡]
Energy	11.1 (10.8)	30.9 (20.0)	.003
Skin	3.0 (3.5)	17.1 (17.5)	<.001
Nutrition	1.5 (3.4)	5.6 (8.9)	.014
Nutrition–modified[§]	1.9 (4.2)	7.0 (11.2)	.014
Lung	.0 (.0)	4.0 (6.6)	<.001
Lung–modified[§]	.0 (.0)	4.5 (7.8)	<.001
Psychological	10.0 (10.2)	17.7 (17.3)	.180
Eye	20.0 (16.8)	48.4 (28.2)	.003
Mouth	8.8 (13.2)	21.8 (26.1)	.024
Summary	7.8 (6.0)	20.8 (9.9)	<.001
Summary–modified[§]	7.8 (6.1)	21.0 (10.1)	<.001

Values are mean (standard deviation).

[*] Five patients indicated they had “none” and “0” or “1” severity on a 0-10 scale. They are included in the study because they had mild, moderate, or severe cGVHD per NIH criteria.

[†] One patient did not report cGVHD severity but was NIH severe so was grouped with the self-reported moderate/severe group.

[‡] Based on *t*-test.

[§] Modified from the original by deletion of 2 items (see text).

DISCUSSION

When the LSS was first developed and validated in the late 1990s, severe manifestations of cGVHD were more common. The survey also used a 1-month recall period because the intent was to use the instrument for clinical care and observational studies. This report shows that the items about need for oxygen or intravenous or tube feeding can be removed without adversely affecting test characteristics if a pure symptom scale is desired. The 7-day recall period may be used because the modified instrument retains its overall reliability and validity. Both changes result in a modified LSS (mLSS) that is better suited to clinical trials.

Results from this study and from reanalysis of 2 earlier cohorts show that endorsement of the oxygen and intravenous/feeding tube items was very infrequent. Although these questions were originally conceived to reflect bother due to the severity of cGVHD requiring need for such supportive care, they do not directly reflect cGVHD symptoms because even very symptomatic patients might refuse oxygen or feeding tubes. The low rate of endorsement seen in modern studies may also be due to better recognition and earlier/more effective treatment of cGVHD, although these hypotheses are speculative. Regardless, this study shows that these 2 items may be removed from the scoring algorithm. Although absolute scores will be higher because of removing items that are usually scored as zeros and bring down the average, as long as the enrollment and follow-up surveys are scored using the same formula, change scores are interpretable. Collection of the full 30-item version allows calculation of either the full or modified scale scores.

A previous study asked patients to compare how they would report their symptoms with a 7-day or 1-month time frame, showing that some patients reported the time frame selected would have altered their answers. The primary reason given was that their cGVHD symptoms had changed for better or worse in the past month, which is a legitimate reason for different answers, further justifying the change to a shorter recall period.

Intraclass correlations of 3 subscales (nutrition, lung, and psychological symptoms) were <.7, suggesting the items are not measuring a single construct. Examination of individual questions supports this conclusion; for example, the nutrition subscale includes difficulty swallowing, nausea, and weight loss, all recognized symptoms and signs of gastrointestinal cGVHD that are not always found together.

Limitations of this study include the modest sample size and restriction to outpatients from 1 center. Participants were well educated with 64% being college graduates. Patients with self-reported severe cGVHD were under-represented (3%), whereas there were 41% with severe cGVHD per the NIH criteria. Very few patients had liver and gastrointestinal symptoms, and patients were only stable or improved (none worsened) between the 2 test and retest measurements, which might be explained because the retest survey was administered only 1 week after a clinic visit where symptoms may have been detected and treated.

In summary, our results document the reliability and validity of the 7-day mLSS for evaluating cGVHD symptoms and suggest a 5- to 6-point difference in the summary score is clinically meaningful. The 7-day mLSS may be used in modern clinical trials.

ACKNOWLEDGMENTS

Financial disclosure: Supported by National Cancer Institute (grants CA118953 and CA163438).

Conflict of interest statement: There are no conflicts of interest to report.

REFERENCES

1. Lee SJ. Classification systems for chronic graft-versus-host disease. *Blood*. 2017;129:30–37.
2. Lee S, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8:444–452.
3. Lee SJ, Wolff D, Kitko C, 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. The 2014 Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2015;21:984–999.
4. 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;12:252–266.
5. Merkel EC, Mitchell SA, Lee SJ. Content validity of the Lee chronic graft-versus-host disease symptom scale as assessed by cognitive interviews. *Biol Blood Marrow Transplant*. 2016;22:752–758.
6. Lee SJ, Logan B, Westervelt P, et al. Comparison of patient-reported outcomes in 5-year survivors who received bone marrow vs peripheral blood unrelated donor transplantation: long-term follow-up of a randomized clinical trial. *JAMA Oncol*. 2016;2:1583–1589.
7. Walker I, Panzarella T, Couban S, et al. Pretreatment with anti-thymocyte globulin versus no anti-thymocyte globulin in patients with haematological malignancies undergoing haemopoietic cell transplantation from unrelated donors: a randomised, controlled, open-label, phase 3, multicentre trial. *Lancet Oncol*. 2016;17:164–173.
8. Miklos D, Cutler CS, Arora M, et al. Ibrutinib for chronic graft-versus-host disease after failure of prior therapy. *Blood*. 2017;130:2243–2250.
9. Williams KM, Cheng GS, Pusic I, et al. Fluticasone, azithromycin, and montelukast treatment for new-onset bronchiolitis obliterans syndrome after hematopoietic cell transplantation. *Biol Blood Marrow Transplant*. 2016;22:710–716.
10. Arai S, Pidala J, Pusic I, et al. A randomized phase II crossover study of imatinib or rituximab for cutaneous sclerosis after hematopoietic cell transplantation. *Clin Cancer Res*. 2016;22:319–327.
11. Taber KS. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res Sci Educ*. 2018;48:1273–1296.
12. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific quality of life questionnaire. *J Clin Epidemiol*. 1994;47:81–87.
13. Lydick E, Epstein RS. Interpretation of quality of life changes. *Qual Life Res*. 1993;2:221–226.
14. Chronic GVHD Consortium. Rationale and design of the chronic GVHD cohort study: improving outcomes assessment in chronic GVHD. *Biol Blood Marrow Transplant*. 2011;17:1114–1120.
15. Chronic GVHD Consortium. Design and patient characteristics of the chronic graft-versus-host disease response measures validation study. *Biol Blood Marrow Transplant*. 2018;24:1727–1732.

ORIGINAL ARTICLE

Ruxolitinib for Glucocorticoid-Refractory Chronic Graft-versus-Host Disease

Robert Zeiser, M.D., Nicola Polverelli, M.D., Ph.D., Ron Ram, M.D.,
Shahrukh K. Hashmi, M.D., M.P.H., Ronjon Chakraverty, M.D., Ph.D.,
Jan Moritz Middeke, M.D., Maurizio Musso, M.D., Sebastian Giebel, M.D., Ph.D.,
Ant Uzay, M.D., Peter Langmuir, M.D., Norbert Hollaender, Ph.D.,
Maanasa Gowda, Pharm.D., Tommaso Stefanelli, M.D., Stephanie J. Lee, M.D., M.P.H.,
Takanori Teshima, M.D., Ph.D., and Franco Locatelli, M.D., Ph.D.,
for the REACH3 Investigators*

ABSTRACT

BACKGROUND

Chronic graft-versus-host disease (GVHD), a major complication of allogeneic stem-cell transplantation, becomes glucocorticoid-refractory or glucocorticoid-dependent in approximately 50% of patients. Robust data from phase 3 randomized studies evaluating second-line therapy for chronic GVHD are lacking. In retrospective surveys, ruxolitinib, a Janus kinase (JAK1–JAK2) inhibitor, showed potential efficacy in patients with glucocorticoid-refractory or -dependent chronic GVHD.

METHODS

This phase 3 open-label, randomized trial evaluated the efficacy and safety of ruxolitinib at a dose of 10 mg twice daily, as compared with the investigator's choice of therapy from a list of 10 commonly used options considered best available care (control), in patients 12 years of age or older with moderate or severe glucocorticoid-refractory or -dependent chronic GVHD. The primary end point was overall response (complete or partial response) at week 24; key secondary end points were failure-free survival and improved score on the modified Lee Symptom Scale at week 24.

RESULTS

A total of 329 patients underwent randomization; 165 patients were assigned to receive ruxolitinib and 164 patients to receive control therapy. Overall response at week 24 was greater in the ruxolitinib group than in the control group (49.7% vs. 25.6%; odds ratio, 2.99; P<0.001). Ruxolitinib led to longer median failure-free survival than control (>18.6 months vs. 5.7 months; hazard ratio, 0.37; P<0.001) and higher symptom response (24.2% vs. 11.0%; odds ratio, 2.62; P=0.001). The most common (occurring in ≥10% patients) adverse events of grade 3 or higher up to week 24 were thrombocytopenia (15.2% in the ruxolitinib group and 10.1% in the control group) and anemia (12.7% and 7.6%, respectively). The incidence of cytomegalovirus infections and reactivations was similar in the two groups.

CONCLUSIONS

Among patients with glucocorticoid-refractory or -dependent chronic GVHD, ruxolitinib led to significantly greater overall response, failure-free survival, and symptom response. The incidence of thrombocytopenia and anemia was greater with ruxolitinib. (Funded by Novartis and Incyte; REACH3 ClinicalTrials.gov number, NCT03112603.)

The authors' affiliations are listed in the Appendix. Address reprint requests to Dr. Zeiser at the Department of Medicine I, Faculty of Medicine, Medical Center, University of Freiburg, Freiburg 79085, Germany, or at robert.zeiser@uniklinik-freiburg.de.

*A list of the investigators in the REACH3 Trial Group is provided in the Supplementary Appendix, available at NEJM.org.

Drs. Zeiser and Polverelli and Drs. Teshima and Locatelli contributed equally to this article.

N Engl J Med 2021;385:228-38.
DOI: 10.1056/NEJMoa2033122

Chronic graft-versus-host disease (GVHD) is a serious complication of allogeneic stem-cell transplantation that limits the success of the procedure.[1,2] Chronic GVHD occurs in approximately 30 to 70% of patients who undergo allogeneic stem-cell transplantation[3] and is a leading cause of complications and of nonrelapse-associated death.[2,4-6] Patients with chronic GVHD have impaired physical, social, psychological, and general quality of life, which worsens with disease severity.[7-10]

Standard first-line treatment of chronic GVHD consists of systemic glucocorticoids; however, in approximately 50% of patients the disease becomes glucocorticoid-refractory or glucocorticoid-dependent, greatly increasing the risk of poor outcomes.[11] Second-line treatment of chronic GVHD varies substantially among treatment centers. Although guidelines provide several treatment options, including extracorporeal photopheresis and mycophenolate mofetil, enrolling patients into clinical trials is recommended.[2,12,13] Currently, ibrutinib, a Bruton's tyrosine kinase inhibitor, is the only second-line therapy approved (in the United States and Canada) for treatment of chronic GVHD; it was approved on the basis of a phase 1b–2, open-label, single-group trial (with 42 patients) that showed a best overall response of 67% and alleviation of symptoms.[14] However, since data from large-scale, successful, randomized studies are not available, no standard second-line treatment has been defined.[12]

Preclinical studies showed that Janus kinase 1 and 2 (JAK1–JAK2) signaling is crucial in the steps leading to inflammation and tissue damage in acute GVHD and chronic GVHD[15-19] and that ruxolitinib, a JAK1–JAK2 inhibitor, was an effective treatment in a mouse model of chronic GVHD.[20] In addition, a retrospective survey showed ruxolitinib led to high response and 6-month survival rates in patients with acute or chronic GVHD who were heavily pretreated.[20] After these findings, ruxolitinib was shown to have high response rates in the phase 2 REACH1 trial involving 71 patients, resulting in approval of ruxolitinib in the United States for the treatment of glucocorticoid-refractory acute GVHD in patients 12 years of age or older.[21,22] The phase 3 REACH2 study involving 309 patients with glucocorticoid-refractory acute GVHD showed that ruxolitinib resulted in significant improvements as compared with control therapy.[23] Here we present the primary analysis of REACH3, a phase 3 randomized trial evaluating ruxolitinib as compared with investigator's choice of therapy from a list of 10 commonly used options among patients with glucocorticoid-refractory or -dependent chronic GVHD.

METHODS

PATIENTS

Patients were at least 12 years of age, had undergone allogeneic stem-cell transplantation, and had moderate or severe glucocorticoid-refractory or -dependent chronic GVHD, according to National Institutes of Health (NIH) consensus criteria.[24] (Details on trial design, end points, and statistical analysis are provided in the Supplementary Methods section in the Supplementary Appendix, available with the full text of this article at NEJM.org.) Patients who had been treated previously with JAK inhibitors for acute GVHD were included if treatment had resulted in a complete or partial response and if they had discontinued JAK inhibitor treatment at least 8 weeks before receiving the first dose of ruxolitinib or control therapy. Patients treated previously with 2 or more systemic therapies for chronic GVHD in addition to glucocorticoids with or without calcineurin inhibitors were ineligible. Patients were excluded if they had a relapse of the primary cancer or had graft loss within 6 months before treatment initiation or if they had an active, uncontrolled infection.

TRIAL OVERSIGHT

The study sponsors (Novartis and Incyte), in collaboration with the trial steering committee, designed the trial and analyzed the data. Investigators entered data into the electronic case-report forms. After data analysis, the first two and last two authors developed a draft of the manuscript with writing assistance provided by ArticulateScience and funded by Novartis. All the authors reviewed and approved the manuscript for submission and vouch for the accuracy and completeness of the data and for the fidelity of the trial to the protocol (available at NEJM.org). The trial was designed and conducted in accordance with the guidelines for Good Clinical Practice of the International Council for Harmonisation, applicable local regulations, and the principles of the Declaration of Helsinki. The protocol was approved at each participating center by the relevant institutional review board or ethics committee. An independent data monitoring committee reviewed interim results and safety (a list of the committee members is provided in the Supplementary Appendix). All patients (or their guardians) provided informed consent.

Trial Design

REACH3 was a phase 3 randomized, open-label, multicenter trial (Fig. S1 in the Supplementary Appendix). Patients were randomly assigned in a 1:1 ratio to receive ruxolitinib at a dose of 10 mg twice daily or therapy chosen by the investigators from a list of 10 commonly used options described in the protocol (extracorporeal photopheresis, low-dose methotrexate, mycophenolate mofetil, a mammalian target of rapamycin [mTOR] inhibitor [everolimus or sirolimus], infliximab, rituximab, pentostatin, imatinib, or ibrutinib) and were stratified according to the severity of their chronic GVHD. Control therapy included the most widely used second-line treatments,[25] as outlined by the European Society for Blood and Marrow Transplantation.[12] Patients continued to receive glucocorticoids with or without calcineurin inhibitors. Infection prophylaxis was allowed and administered according to local institutional guidelines.

Patients received assigned treatment for at least 6 cycles (28 days per cycle) unless they had unacceptable side effects or progression of chronic GVHD. Glucocorticoids could be tapered after patients had a complete response or partial response; tapering of calcineurin inhibitors or ruxolitinib was allowed on or after cycle 7 day 1 (week 24) and after patients had a complete or partial response. Addition or initiation of a new control therapy was allowed before week 24 because of lack of response, unacceptable side effects, or a flare of chronic GVHD and was considered treatment failure. For patients who did not have or maintain a complete or partial response, had unacceptable side effects from a control therapy, or had a flare of chronic GVHD, crossover from control therapy to ruxolitinib could occur on or after week 24. Patients in the control group who had a complete or partial response at week 24 could not cross over to ruxolitinib unless they had disease progression, mixed response, or unacceptable side effects from the control therapy.

End Points

The primary end point was overall response (defined as a complete or partial response according to 2014 NIH consensus criteria)[26] at week 24. The two key secondary end points were failure-free survival (defined as time to recurrence of underlying disease, start of new systemic treatment for chronic GVHD, or death, whichever came first) and response on the modified Lee Symptom Scale[27,28] (defined as a ≥7-point reduction from baseline in total symptom score on the scale, which measures the symptoms of chronic GVHD on a scale of 0 to 100, with higher scores indicating worse symptoms) at week 24. Modifications to the Lee Symptom Scale included changing the measure from “bother” to the severity of each symptom and shortening the recall period from the past month to the past 7 days. Secondary and exploratory end points included subgroup analyses of overall response, individual organ responses, best overall response at any time up to week 24, duration of response, change in glucocorticoid dose over time, overall survival, and changes in quality-of-life measures. Safety analyses included patients who received at least 1 dose of treatment; safety data up to week 24 are presented to ensure similar exposure in the two groups. Given that not all patients who crossed over from the control group to the ruxolitinib group had completed 24 weeks of treatment with ruxolitinib at the time of this analysis, the only result presented for crossover patients is the best overall response up to data cutoff.

Statistical Analysis

Sample size calculations were performed to achieve 90% power for overall response rate and failure-free survival; a sample size of 324 patients was considered adequate. The Cochran–Mantel–Haenszel chi-square test, stratified according to severity of chronic GVHD, was used to compare overall responses and responses on the modified Lee Symptom Scale between the two groups; failure-free survival was compared with the use of a stratified log-rank test. Efficacy analyses were performed on the full analysis set according to the intention-to-treat principle. P values, odds ratios, and hazard ratios including 95% confidence limits were derived from the respective stratified analyses. We calculated adjusted risk ratios by fitting a generalized linear model with the treatment group and chronic GVHD severity as covariates.

An overall hierarchical testing procedure[29] (Fig. S2) was applied to test the primary end point and the two key secondary end points in a two-look, group-sequential design at the interim analysis (196 patients; alpha significance level, 0.01176) and at the primary analysis (329 patients; alpha significance level, 0.01858 if not positive at the interim analysis). The testing sequence for key secondary end points differed between the United States (modified Lee Symptom Scale tested before failure-free survival) and other countries (failure-free survival tested before modified Lee Symptom Scale) because regulatory recommendations for demonstrating additional patient benefits differed between countries. The overall hierarchical testing procedure maintained the overall one-sided type I alpha error of 0.025 for the primary and key secondary end points; one-sided tests were applied to allow sequential testing only in cases in which ruxolitinib was superior to control therapy.

RESULTS

PATIENTS

Between July 11, 2017, and November 18, 2019, a total of 329 patients were randomly assigned to receive ruxolitinib (165 patients) or a control therapy (164 patients) at 149 centers across 28 countries (Fig. 1). Patient characteristics were balanced between treatment groups (Table 1 and Table S1). The median age of the patients was 49 years (range, 12 to 76 years; 12 were between 12 and 17 years of age); 61.1% were male. Overall, 42.9% of the patients had moderate chronic GVHD, and 56.5% of patients had severe chronic GVHD; 71.4% had glucocorticoid-refractory chronic GVHD, and 28.6% had glucocorticoid-dependent disease, as reported by the investigator. Control therapy was primarily extracorporeal photopheresis (34.8%), mycophenolate mofetil (22.2%), and ibrutinib (17.1%). Approximately half the patients received calcineurin inhibitors during the trial (Table S2).

At data cutoff (May 8, 2020; median follow-up, 57.3 weeks), 125 patients (38.0%) continued to receive the randomized treatment; 82 patients (49.7%) discontinued ruxolitinib and 122 patients (74.4%) discontinued control therapy (Fig. 1). Reasons for discontinuation included lack of efficacy (14.5% in the ruxolitinib group vs. 42.7% in the control group), adverse events (17.0% vs. 4.9%), and relapse of underlying disease (5.5% vs. 4.3%); 61 patients (37.2%) in the control group crossed over to ruxolitinib. The median exposure to therapy was 41.3 weeks (range, 0.7 to 127.3) in the ruxolitinib group and 24.1 weeks (range, 0.6 to 108.4) in the control group.

EFFICACY

Overall response at week 24 (the primary end point) was higher with ruxolitinib (82 patients, 49.7%) than with control therapy (42 patients, 25.6%) (odds ratio, 2.99 [95% confidence interval {CI}, 1.86 to 4.80]; risk ratio, 1.93 [95% CI, 1.44 to 2.60]; P<0.001) (Fig. 2A and Table S3). A total of 11 patients (6.7%) in the ruxolitinib group and 5 (3.0%) in the control group had a complete response. The efficacy boundary for overall response was crossed at the interim analysis, with the value being higher with ruxolitinib than with control therapy (50.5% [8 with a complete response, 41 with a partial response] vs. 26.3% [3 with a complete response, 23 with a partial response]; P<0.001). A higher overall response was observed with ruxolitinib than with control therapy regardless of the organs involved (Table S4 and Fig. S3). Although patients were not stratified according to organ involvement, odds ratios favored ruxolitinib in all organ subgroups. Response according to the investigator-selected control drug regimen is shown in Figure S4.

Patients receiving ruxolitinib had longer failure-free survival than patients receiving control therapy (median failure-free survival, >18.6 months vs. 5.7 months; hazard ratio, 0.37; 95% CI, 0.27 to 0.51; P<0.001). The median failure-free survival with ruxolitinib was not reached, but the lower boundary of the 95% confidence interval was estimated as 18.6 months, with the efficacy boundary crossed at the interim analysis (Fig. 2B and Fig. S5). The probability of failure-free survival at 6 months, as estimated with the use of the Kaplan–Meier method, was higher with ruxolitinib (74.9%; 95% CI, 67.5 to 80.9) than with control therapy (44.5%; 95% CI, 36.5 to 52.1). The response on the modified Lee Symptom Scale at 24 weeks was also higher with ruxolitinib than with control therapy (24.2% vs. 11.0%; odds ratio, 2.62 [95% CI, 1.42 to 4.82]; risk ratio, 2.19 [95% CI, 1.31 to 3.65]; P=0.001) (Fig. 2C). The dose of glucocorticoids decreased over time in both groups, with a slightly greater decrease with ruxolitinib (Fig. S6).

A best overall response up to week 24 was

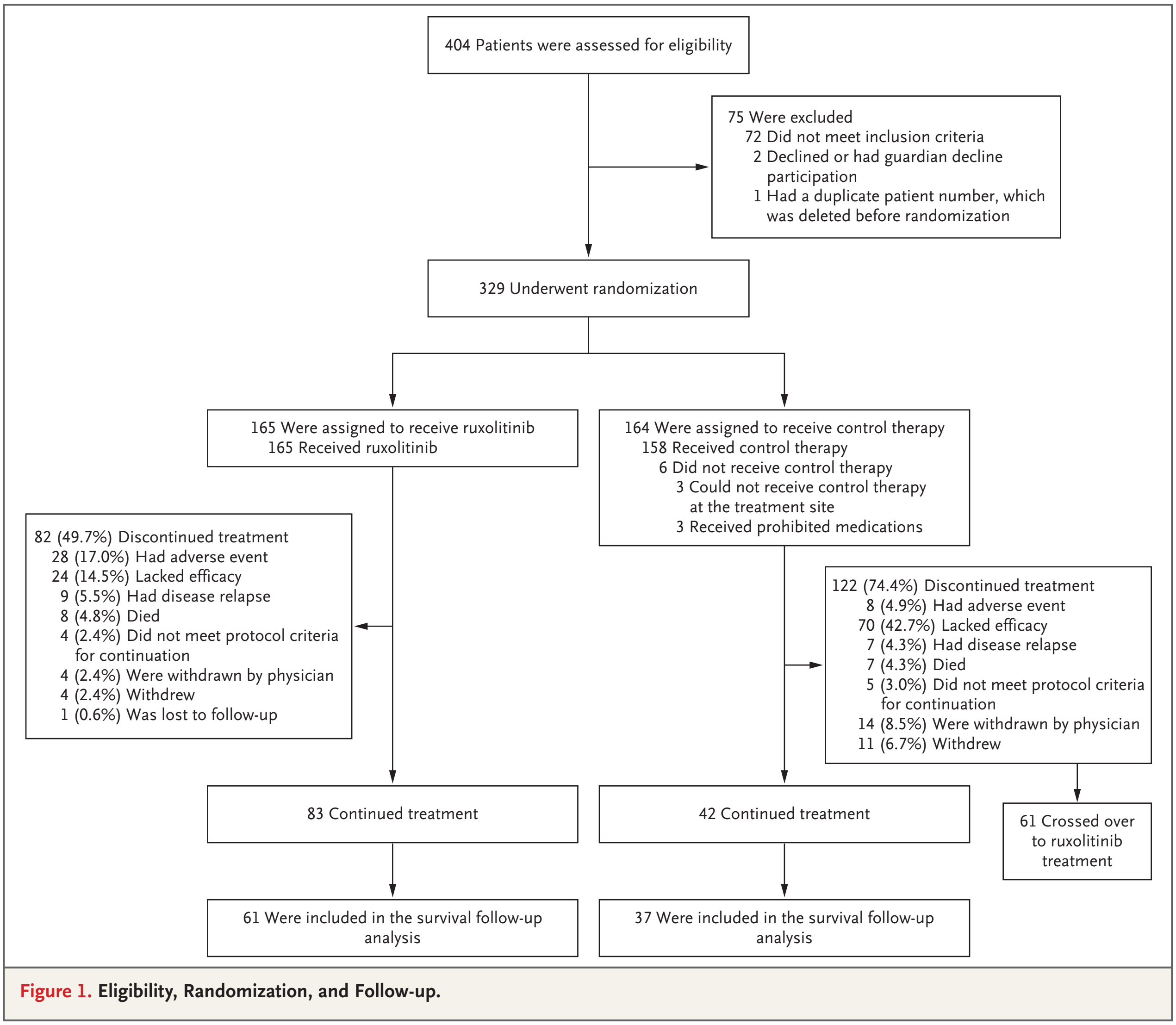
404 Patients were assessed for eligibility
75 Were excluded
72 Did not meet inclusion criteria
2 Declined or had guardian decline participation
1 Had a duplicate patient number, which was deleted before randomization
329 Underwent randomization
165 Were assigned to receive ruxolitinib
165 Received ruxolitinib
164 Were assigned to receive control therapy
158 Received control therapy
6 Did not receive control therapy
3 Could not receive control therapy at the treatment site
3 Received prohibited medications
82 (49.7%) Discontinued treatment
28 (17.0%) Had adverse event
24 (14.5%) Lacked efficacy
9 (5.5%) Had disease relapse
8 (4.8%) Died
4 (2.4%) Did not meet protocol criteria for continuation
4 (2.4%) Were withdrawn by physician
4 (2.4%) Withdrew
1 (0.6%) Was lost to follow-up
122 (74.4%) Discontinued treatment
8 (4.9%) Had adverse event
70 (42.7%) Lacked efficacy
7 (4.3%) Had disease relapse
7 (4.3%) Died
5 (3.0%) Did not meet protocol criteria for continuation
14 (8.5%) Were withdrawn by physician
11 (6.7%) Withdrew
83 Continued treatment
42 Continued treatment
61 Crossed over to ruxolitinib treatment
61 Were included in the survival follow-up analysis
37 Were included in the survival follow-up analysis

Figure 1. Eligibility, Randomization, and Follow-up.

observed in 76.4% of patients in the ruxolitinib group and in 60.4% in the control group (odds ratio, 2.17 [95% CI, 1.34 to 3.52]; risk ratio, 1.24 [95% CI, 1.07 to 1.43]; P=0.001) (Fig. 3A). Among patients with a response at any time, the estimated probability of maintaining a response at 12 months was 68.5% (95% CI, 58.9 to 76.3) in the ruxolitinib group as compared with 40.3% (95% CI, 30.3 to 50.2) in the control group (Fig. 3B). Patients who crossed over from control therapy to ruxolitinib (61 patients) also had a response, with a best overall response at data cutoff in 78.7% (4 with a complete response and 44 with a partial response), a finding consistent with the best overall response with ruxolitinib in the randomized population. Overall survival data were not mature at data cutoff, and median overall survival was not reached in either group (hazard ratio, 1.09; 95% CI, 0.65 to 1.82) (Fig. S7). At 12 months, the estimated probability of survival was 81.4% with ruxolitinib (95% CI, 74.1 to 86.8) and 83.8% with control therapy (95% CI, 76.5 to 89.0).

SAFETY

Safety analyses included 323 patients (165 in the ruxolitinib group and 158 in the control group) who received at least 1 dose of trial treatment up to week 24. Up to day 179, the median duration of exposure to therapy was 25.6 weeks (range, 0.7 to 25.6) in the ruxolitinib group and 24.0 weeks (range, 0.6 to 25.6) in the control group. Adverse events of any grade up to week 24 occurred in 97.6% of the patients (161) who re-

Table 1. Characteristics of the Patients at Baseline.*

Variable	Ruxolitinib (N=165)	Control (N=164)
Age		
Median (range) — yr	49.0 (13.0–73.0)	50.0 (12.0–76.0)
Distribution — no. (%)		
12 to <18 yr	4 (2.4)	8 (4.9)
18 to 65 yr	143 (86.7)	134 (81.7)
>65 yr	18 (10.9)	22 (13.4)
Sex — no. (%)		
Male	109 (66.1)	92 (56.1)
Female	56 (33.9)	72 (43.9)
Previous acute GVHD — no. (%)	92 (55.8)	88 (53.7)
Chronic GVHD severity — no. (%)†		
Mild	1 (0.6)	1 (0.6)
Moderate	67 (40.6)	74 (45.1)
Severe	97 (58.8)	89 (54.3)
Donor type — no. (%)‡		
Related	91 (54.5)	87 (52.1)
Unrelated	76 (45.5)	80 (47.9)
Previous systemic therapy for chronic GVHD or glucocorticoid-refractory or -dependent chronic GVHD — no. (%)§		
Glucocorticoid only	70 (42.4)	81 (49.4)
Glucocorticoid + calcineurin inhibitors	68 (41.2)	69 (42.1)
Glucocorticoid + calcineurin inhibitors + other systemic therapy	10 (6.1)	4 (2.4)
Glucocorticoid + other systemic therapy	14 (8.5)	9 (5.5)
Missing data	3 (1.8)	1 (0.6)

* GVHD denotes graft-versus-host disease.

† Severity was graded according to National Institutes of Health consensus staging criteria[30] at screening. Enrollment of patients with mild glucocorticoid-refractory or glucocorticoid-dependent chronic GVHD was considered a protocol deviation.

‡ Some patients received more than one transplant.

§ Values for previous treatment of chronic GVHD were obtained from documented patient history of medication; topical or local treatments were not counted.

ceived ruxolitinib as compared with 91.8% of the patients (145) who received control therapy (Table 2 and Table S5). Occurrence of adverse events of grade 3 or higher was similar in the two groups (in 57.0% of the patients who received ruxolitinib and in 57.6% of the patients who received control therapy). The most common adverse events of grade 3 or higher were thrombocytopenia (in 15.2% of patients who received ruxolitinib and 10.1% of patients who received control therapy), anemia (in 12.7% and 7.6%), neutropenia (in 8.5% and 3.8%), and pneumonia (in 8.5% and 9.5%). Serious adverse events up to week 24 occurred in 55 patients (33.3%) who received ruxolitinib and in 58 patients (36.7%) who received control therapy (Table S6).

Adverse events led to treatment discontinuation in 27 patients (16.4%) who received ruxolitinib and in 11 (7.0%) who received control therapy. Clinically documented pneumonia was the only adverse event leading to discontinuation by 2% or more of patients in the ruxolitinib group (4.8%, as compared with 1.3% of patients in the control therapy group) (Table S7). Adverse events leading to dose adjustments or interruptions occurred in 62 patients (37.6%) who received ruxolitinib and in 26 patients (16.5%) who received control therapy.

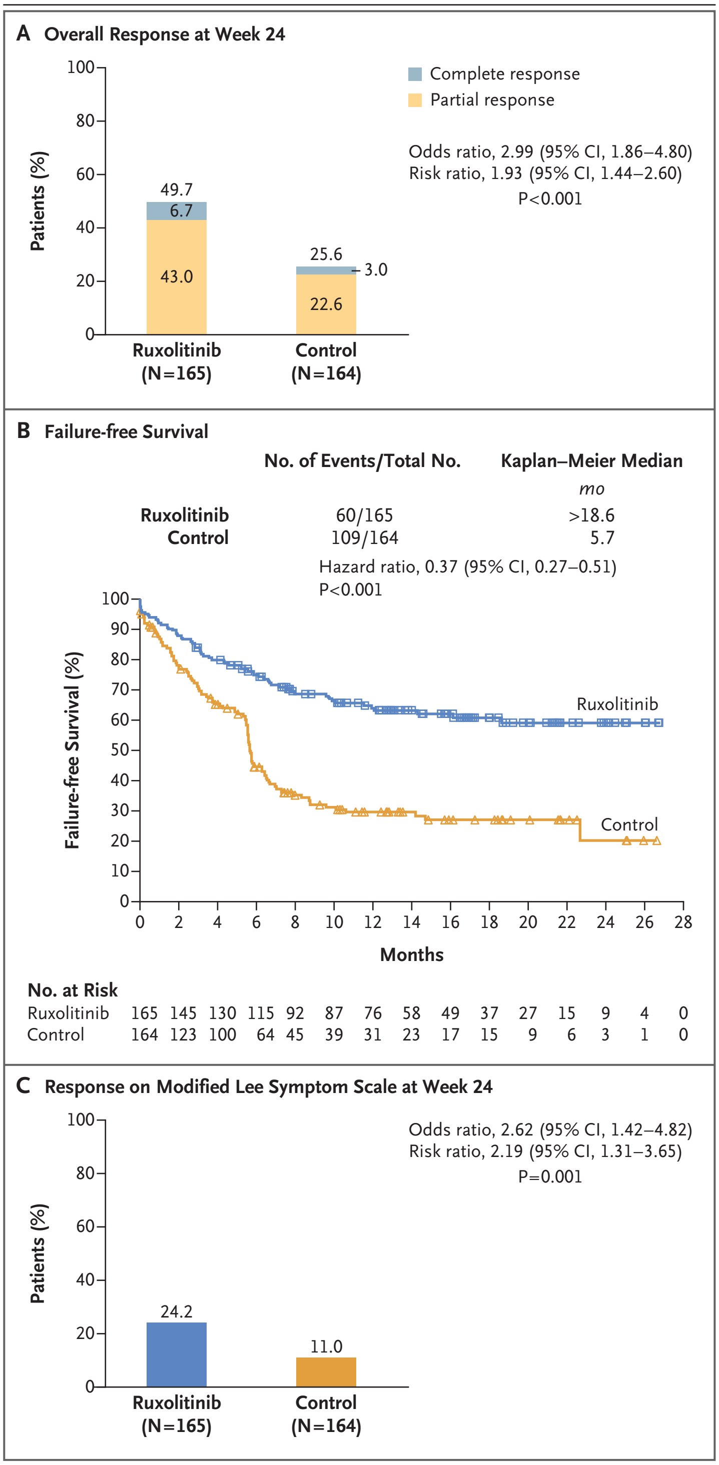

Figure 2. Response at Week 24 and Failure-free Survival.

End points were tested at the interim analysis (196 patients; alpha significance level, 0.01176) and the current primary analysis (329 patients; alpha significance level, 0.01858 if not positive at the interim analysis) according to an overall hierarchical testing procedure to control the one-sided familywise alpha level at 0.025 overall. The test sequence for results among patients outside the United States was overall response, failure-free survival, and score on the modified Lee Symptom Scale; the test sequence for results among patients in the United States was overall response, score on the modified Lee Symptom Scale, and failure-free survival. For the P value for overall response at week 24 (Panel A), the efficacy boundary was crossed at the interim analysis (overall response was 50.5% with ruxolitinib and 26.3% with control therapy; P<0.001). One-sided P value, odds ratio, and 95% confidence interval were calculated with the use of a stratified Cochran–Mantel–Haenszel test, with moderate and severe chronic GVHD as strata. For P values for failure-free survival (Panel B), the efficacy boundary was crossed at the interim analysis for results among patients not in the United States (hazard ratio, 0.32; 95% CI, 0.21 to 0.49; P<0.001). For results among patients in the United States, the hypothesis was retested at the primary analysis according to the overall hierarchical testing procedure (details are provided in the Supplementary Methods section in the Supplementary Appendix). At data cutoff (May 8, 2020), the median failure-free survival was not reached in the ruxolitinib group, but the lower bound of the 95% confidence interval was estimated to be 18.6 months. Patients receiving ruxolitinib had a numerically, but not significantly, higher response (defined as a ≥7-point reduction from baseline in total symptom score) according to the modified Lee Symptom Scale (Panel C) at the interim analysis than those receiving control therapy (19.6% vs. 8.1%; odds ratio, 2.80; P=0.02).

Infections of any type occurred in 63.6% of patients who received ruxolitinib as compared with 56.3% who received control therapy (grade 3 infections, 19.4% vs. 18.4%, according to the grading system described by Cordonnier et al.[31]). Viral infections were the most common (33.9% and 29.1% in the ruxolitinib and control groups, respectively), followed by bacterial (27.9% and 25.9%) and fungal infections (11.5% and 5.7%); infections of unknown type occurred in 21.2% of patients who received ruxolitinib and in 20.3% of patients who received control therapy (Table S8). Cytomegalovirus infection and reactivation were similar in the two groups (5.5% and 8.2%) (Table 2).

As of the data cutoff, 31 patients (18.8%) who received ruxolitinib and 27 patients (16.5%) who received control therapy had died. Deaths were due primarily to complications caused by chronic GVHD disease or treatment (or both) (22 patients [13.3%] who received ruxolitinib vs. 13 patients [7.9%] who received control therapy, including 2 deaths after crossover to ruxolitinib) or infections (2 patients [1.2%] vs. 6 patients

[3.7%]). The incidence of cancer relapse and progression was low in both groups (9 patients [5.8%] and 8 patients [5.0%]). The estimated cumulative incidence of relapse at 6 months was 2.59% (95% CI, 0.85 to 6.08) among patients who received ruxolitinib and 2.65% (95% CI, 0.87 to 6.21) among patients who received control therapy.

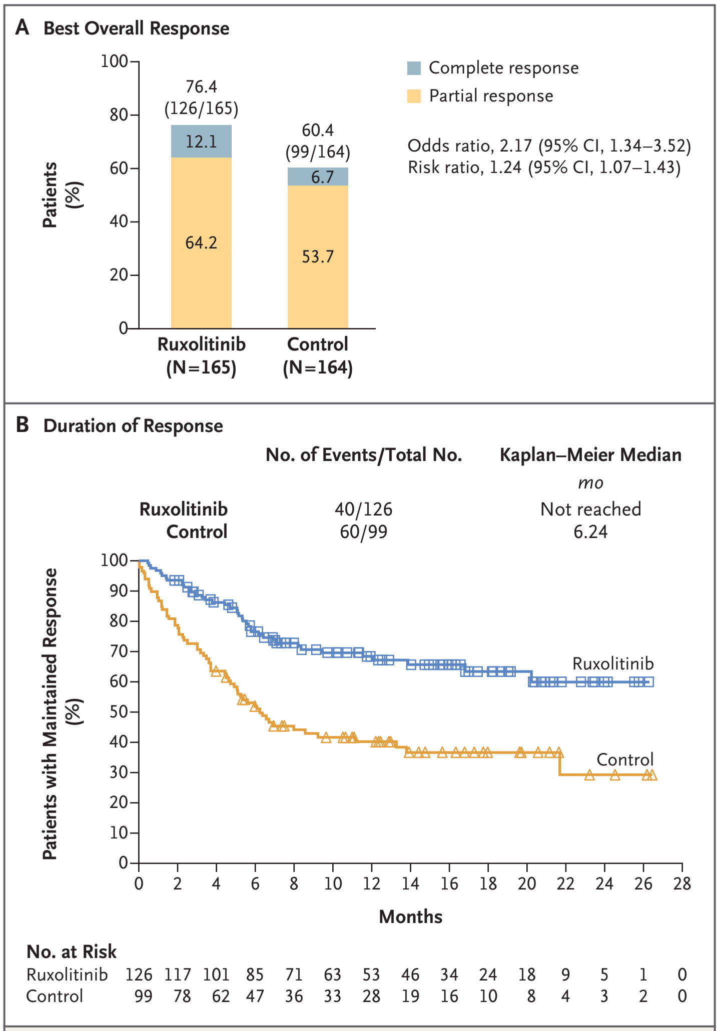

Figure 3. Overall Response and Duration of Response.

The comparisons for overall response (Panel A) and duration of response (Panel B) are based on the subgroup of patients who had a complete or partial response at any time up to week 24. Duration of response to treatment (Panel B) was measured as the time from first documented complete response or partial response.

DISCUSSION

REACH3 is a phase 3 randomized trial that showed the superiority of ruxolitinib over common second-line therapeutic options, including ibrutinib and extracorporeal photopheresis, for treatment of glucocorticoid-refractory or -dependent chronic GVHD. Ruxolitinib led to a higher overall response than control therapy at week 24 (49.7% vs. 25.6%), regardless of the organs involved, and a higher best overall response (76.4% vs. 60.4%), a longer duration of response, and longer failure-free survival. The results in individual organs showed that ruxolitinib led to higher responses in most organs than control therapy. The response in the lungs and liver was low in both treatment groups, which highlights how difficult treatment can be when these organs are affected. However, subgroup analysis of overall response according to organ involvement showed that chronic GVHD in difficult-to-treat organs did not preclude alleviation of chronic GVHD in other organs in patients receiving ruxolitinib, so the overall response was favorable.

In addition, patients treated with ruxolitinib had greater reduction of symptoms than those treated with control therapy, as measured by the modified Lee Symptom Scale, a scale specific to chronic GVHD.[32] Achievement of complete response or partial response, as measured according to NIH criteria and improvements in the modified Lee Symptom Scale score at 6 months, has been associated with better survival.[33,34] Early data do not suggest a difference in survival between treatment groups. Longer follow-up is needed to evaluate the effect of ruxolitinib on survival.

The absence of a strong end point, such as glucocorticoid-free remission, and the presence of confounders, including concomitant treatments, make determination of the effect on glucocorticoid dose over time with ruxolitinib as compared with commonly used therapies difficult. However, patients treated with ruxolitinib had consistent reductions in glucocorticoid dose over time (Fig. S6), suggesting a glucocorticoid-sparing effect, a finding in line with previous observations.[20]

The safety profile of ruxolitinib was consistent with observations in patients with acute GVHD and expectations in patients with glucocorticoid-refractory or -dependent chronic GVHD. The most common adverse event was anemia, which was expected given the mechanism of action and known safety profile of ruxolitinib.[35,36] Thrombocytopenia, another known side effect of ruxolitinib, was also common, but both ane-

Table 2. Adverse Events up to Week 24 in 5% or More of Patients Treated with Ruxolitinib.*

Adverse Event	Ruxolitinib (N=165)		Control (N=158)	
	Any Grade	Grade ≥3	Any Grade	Grade ≥3
	number of patients with event (percent)			
Any	161 (97.6)	94 (57.0)	145 (91.8)	91 (57.6)
Hematologic event				
Anemia	48 (29.1)	21 (12.7)	20 (12.7)	12 (7.6)
Thrombocytopenia†	35 (21.2)	25 (15.2)	23 (14.6)	16 (10.1)
Neutropenia	18 (10.9)	14 (8.5)	8 (5.1)	6 (3.8)
Gastrointestinal event				
Diarrhea	17 (10.3)	1 (0.6)	21 (13.3)	2 (1.3)
Nausea	15 (9.1)	0	16 (10.1)	2 (1.3)
Vomiting	12 (7.3)	0	10 (6.3)	2 (1.3)
Constipation	12 (7.3)	0	8 (5.1)	0
Infection				
Pneumonia	18 (10.9)	14 (8.5)	20 (12.7)	15 (9.5)
Upper respiratory tract infection	14 (8.5)	0	13 (8.2)	2 (1.3)
Urinary tract infection	11 (6.7)	1 (0.6)	5 (3.2)	2 (1.3)
Nasopharyngitis	10 (6.1)	0	6 (3.8)	0
BK virus infection	9 (5.5)	1 (0.6)	2 (1.3)	0
Cytomegalovirus infection or reactivation	9 (5.5)	2 (1.2)	13 (8.2)	0
Laboratory abnormality				
Alanine aminotransferase increased	25 (15.2)	7 (4.2)	7 (4.4)	0
Creatinine increased	23 (13.9)	0	7 (4.4)	1 (0.6)
Aspartate aminotransferase increased	16 (9.7)	3 (1.8)	4 (2.5)	1 (0.6)
Hypertriglyceridemia	16 (9.7)	8 (4.8)	13 (8.2)	6 (3.8)
γ-glutamyltransferase increased	15 (9.1)	11 (6.7)	5 (3.2)	3 (1.9)
Hyperglycemia	13 (7.9)	8 (4.8)	5 (3.2)	3 (1.9)
Hypokalemia	13 (7.9)	3 (1.8)	16 (10.1)	7 (4.4)
Cholesterol increased	12 (7.3)	4 (2.4)	7 (4.4)	3 (1.9)
Amylase increased	11 (6.7)	5 (3.0)	3 (1.9)	0
Lipase increased	10 (6.1)	4 (2.4)	2 (1.3)	1 (0.6)
Hypercholesterolemia	9 (5.5)	2 (1.2)	2 (1.3)	1 (0.6)
Hyperkalemia	9 (5.5)	3 (1.8)	4 (2.5)	1 (0.6)
Other				
Hypertension	26 (15.8)	8 (4.8)	20 (12.7)	11 (7.0)
Pyrexia	26 (15.8)	3 (1.8)	15 (9.5)	2 (1.3)
Cough	17 (10.3)	0	11 (7.0)	0
Fatigue	17 (10.3)	1 (0.6)	12 (7.6)	3 (1.9)
Dyspnea	16 (9.7)	3 (1.8)	10 (6.3)	2 (1.3)
Headache	14 (8.5)	2 (1.2)	12 (7.6)	1 (0.6)
Peripheral edema	12 (7.3)	1 (0.6)	14 (8.9)	0
Back pain	11 (6.7)	1 (0.6)	11 (7.0)	0
Insomnia	11 (6.7)	0	6 (3.8)	0
Myalgia	11 (6.7)	0	5 (3.2)	0
Arthralgia	10 (6.1)	0	8 (5.1)	0

* The safety data include all patients who received at least one dose of study drug.
† Included are events recorded as thrombocytopenia and decreased platelet count.

mia and thrombocytopenia are reversible and can be managed with dose reductions and supportive care.[35,36] Overall, 37.6% of patients had adverse events leading to dose modifications, and 16.4% had events leading to discontinuation of ruxolitinib. A smaller percentage of patients treated with control therapy than with ruxolitinib had adverse events leading to discontinuation (7.0% vs. 16.4%), but this finding may have been confounded or affected by more than 40% of patients discontinuing control therapy early owing to lack of efficacy or by stricter protocol-defined guidance on ruxolitinib dose modifications if adverse events were suspected to be related to a trial drug. A total of 11 deaths were reported as being related to a trial drug (7 deaths [4.2%] with ruxolitinib and 4 [2.5%] with control therapy).

The incidence of grade 3 infection was similar in the two groups (19.4% vs. 18.4%). The incidence of cytomegalovirus infection or reactivation with ruxolitinib was similar to that with control therapy (Table 2) and was lower than that observed in a retrospective analysis (14.6%).[20] A numerically higher incidence of fungal infections (as classified according to the system described by Cordonnier[31]) was observed with ruxolitinib, which suggests the possible occurrence of opportunistic infections during treatment.[37] Given the risk of infections, patients treated with ruxolitinib should receive prophylaxis against infection, and a low threshold for evaluation of new signs and symptoms should be adopted.

In order to accommodate various control-therapy options, an open-label study design was necessary. To minimize potential bias,[38] we assessed response using the latest NIH consensus response criteria. Better adherence to these objective measures in REACH3 than in previous studies may have resulted in lower overall responses with control therapy and ruxolitinib than have been reported previously. Most studies evaluating the most common therapy options, including ibrutinib[14] and extracorporeal photopheresis,[39] in glucocorticoid-refractory chronic GVHD have been uncontrolled, nonrandomized studies, with the few exceptions showing no superiority over control therapy.[40] In addition, many of the previous studies were conducted before NIH response criteria were established, which probably led to higher treatment effects and overestimated responses, as reported in a meta-analysis assessing the effect of deviations from NIH recommendations.[38] Furthermore, many studies, including the ibrutinib study,[14] included best response (at any time) — referred to as best overall response in our trial — whereas our primary end point was overall response at 24 weeks (a single time point). Indeed, the best overall response in the control group (60.4%, as compared with 76.4% in the ruxolitinib group) was closer to what has been reported for other studies.

Our trial showed that among patients with moderate or severe chronic GVHD in whom glucocorticoids produced an inadequate response, ruxolitinib was superior to control therapies, as evidenced by a greater overall response, longer failure-free survival, and greater reduction in symptoms. Patients receiving ruxolitinib had a higher incidence of grade 3 or worse thrombocytopenia and anemia than those receiving control therapy; no new safety signals were observed.

Supported by Novartis and Incyte.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

A data sharing statement provided by the authors is available with the full text of this article at NEJM.org.

We thank Karen Chinchilla, Ph.D., of ArticulateScience, for providing medical writing and editorial support for an earlier version of the manuscript, funded by Novartis in accordance with Good Publication Practice guidelines (http://www.ismpp.org/gpp3).

Appendix

The authors' affiliations are as follows: the Department of Medicine I, Faculty of Medicine, Medical Center, University of Freiburg, Freiburg (R.Z.), and Medizinische Klinik und Poliklinik I, Universitätsklinikum Dresden, Dresden (J.M.M.) — both in Germany; the Unit of Blood Diseases and Stem Cell Transplantation, Department of Clinical and Experimental Sciences, ASST Spedali Civili di Brescia, University of Brescia, Brescia (N.P.), UOC di Oncoematologia e TMO, Dipartimento Oncologico "la Maddalena," Palermo (M.M.), and Dipartimento di Oncoematologia Pediatrica, IRCCS, Ospedale Pediatrico Bambino Gesu', Sapienza, Università di Roma, Rome (F.L.) — all in Italy; the BMT Unit, Tel Aviv (Sourasky) Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel (R.R.); the Oncology Center, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia (S.K.H.); the Department of Medicine, Sheikh Shakhbout Medical City, Mayo Clinic, Abu Dhabi, United Arab Emirates (S.K.H.); UCL Cancer Institute, Institute of Immunity and Transplantation, London (R.C.); the Department of Bone Marrow Transplantation and Onco-Hematology, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice, Poland (S.G.); Acibadem University Hospital, Hematology Department, Istanbul, Turkey (A.U.); Incyte, Wilmington, DE (P.L.); Novartis Pharma, Basel, Switzerland (N.H., T.S.); Novartis Pharmaceuticals, East Hanover, NJ (M.G.); the Fred Hutchinson Cancer Research Center, Seattle (S.J.L.); and the Department of Hematology, Hokkaido University Faculty of Medicine, Sapporo, Japan (T.T.).

REFERENCES

1. Blazar BR, Murphy WJ, Abedi M. Advances in graft-versus-host disease biology and therapy. Nat Rev Immunol 2012; 12:443-58.

2. Zeiser R, Blazar BR. Pathophysiology of chronic graft-versus-host disease and therapeutic targets. N Engl J Med 2017; 377:2565-79.

3. Arora M, Cutler CS, Jagasia MH, et al. Late acute and chronic graft-versus-host disease after allogeneic hematopoietic cell transplantation. Biol Blood Marrow Transplant 2016;22:449-55.

4. Socié G, Stone JV, Wingard JR, et al. Long-term survival and late deaths after allogeneic bone marrow transplantation. N Engl J Med 1999;341:14-21.

5. Wingard JR, Majhail NS, Brazauskas R, et al. Long-term survival and late deaths after allogeneic hematopoietic cell transplantation. J Clin Oncol 2011;29:2230-9.

6. Lee SJ, Zahrieh D, Alyea EP, et al. Comparison of T-cell-depleted and non-T-cell-depleted unrelated donor transplantation for hematologic diseases: clinical outcomes, quality of life, and costs. Blood 2002;100:2697-702.

7. Pidala J, Kurland B, Chai X, et al. Patient-reported quality of life is associated with severity of chronic graft-versus-host disease as measured by NIH criteria: report on baseline data from the Chronic GVHD Consortium. Blood 2011;117:4651-7.

8. Kurosawa S, Oshima K, Yamaguchi T, et al. Quality of life after allogeneic hematopoietic cell transplantation according to affected organ and severity of chronic graft-versus-host disease. Biol Blood Marrow Transplant 2017;23:1749-58.

9. Jacobs JM, Fishman S, Sommer R, et al. Coping and modifiable psychosocial factors are associated with mood and quality of life in patients with chronic graft-versus-host disease. Biol Blood Marrow Transplant 2019;25:2234-42.

10. Griffith S, Fenech AL, Nelson A, Geer JA, Temel JS, El-Jawahri A. Post-traumatic stress symptoms in hematopoietic stem cell transplant (HCT) recipients. J Clin Oncol 2020;38:Suppl:7505. abstract.

11. Axt L, Naumann A, Toennies J, et al. Retrospective single center analysis of outcome, risk factors and therapy in steroid refractory graft-versus-host disease after allogeneic hematopoietic cell transplantation. Bone Marrow Transplant 2019;54: 1805-14.

12. Penack O, Marchetti M, Ruutu T, et al. Prophylaxis and management of graft versus host disease after stem-cell transplantation for haematological malignancies: updated consensus recommendations of the European Society for Blood and Marrow Transplantation. Lancet Haematol 2020;7(2):e157-e167.

13. Dignan FL, Amrolia P, Clark A, et al. Diagnosis and management of chronic graft-versus-host disease. Br J Haematol 2012;158:46-61.

14. Miklos D, Cutler CS, Arora M, et al. Ibrutinib for chronic graft-versus-host disease after failure of prior therapy. Blood 2017;130:2243-50.

15. Hechinger A-K, Smith BAH, Flynn R, et al. Therapeutic activity of multiple common γ-chain cytokine inhibition in acute and chronic GVHD. Blood 2015; 125:570-80.

16. Choi J, Ziga ED, Ritchey J, et al. IFNγR signaling mediates alloreactive T-cell trafficking and GVHD. Blood 2012;120:4093-103.

17. Nicholson SE, Oates AC, Harpur AG, Ziemiecki A, Wilks AF, Layton JE. Tyrosine kinase JAK1 is associated with the granulocyte-colony-stimulating factor receptor and both become tyrosine-phosphorylated after receptor activation. Proc Natl Acad Sci U S A 1994;91:2985-8.

18. Schwab L, Goroncy L, Palaniyandi S, et al. Neutrophil granulocytes recruited upon translocation of intestinal bacteria enhance graft-versus-host disease via tissue damage. Nat Med 2014;20:648-54.

19. Spoerl S, Mathew NR, Bscheider M, et al. Activity of therapeutic JAK 1/2 blockade in graft-versus-host disease. Blood 2014;123:3832-42.

20. Zeiser R, Burchert A, Lengerke C, et al. Ruxolitinib in corticosteroid-refractory graft-versus-host disease after allogeneic stem cell transplantation: a multicenter survey. Leukemia 2015;29:2062-8.

21. Jagasia M, Perales M-A, Schroeder MA, et al. Ruxolitinib for the treatment of steroid-refractory acute GVHD (REACH1): a multicenter, open-label phase 2 trial. Blood 2020;135:1739-49.

22. Jakafi (ruxolitinib). Wilmington, DE: Incyte, 2020 (prescribing information) (https://www.jakafi.com/pdf/prescribing-information.pdf).

23. Zeiser R, von Bubnoff N, Butler J, et al. Ruxolitinib for glucocorticoid-refractory acute graft-versus-host disease. N Engl J Med 2020;382:1800-10.

24. Martin PJ, Lee SJ, Przepiorka D, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease. VI. The 2014 Clinical Trial Design Working Group report. Biol Blood Marrow Transplant 2015;21:1343-59.

25. Sarantopoulos S, Cardones AR, Sullivan KM. How I treat refractory chronic graft-versus-host disease. Blood 2019;133: 1191-200.

26. Lee SJ, Wolff D, Kitko C, 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. The 2014 Response Criteria Working Group report. Biol Blood Marrow Transplant 2015;21:984-99.

27. Lee SK, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. Biol Blood Marrow Transplant 2002;8:444-52.

28. Teh C, Onstad L, Lee SJ. Reliability and validity of the modified 7-Day Lee Chronic Graft-versus-Host Disease Symptom Scale. Biol Blood Marrow Transplant 2020;26:562-7.

29. Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. Stat Med 2010;29: 219-28.

30. Jagasia MH, Greinix HT, Arora M, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease. I. The 2014 Diagnosis and Staging Working Group report. Biol Blood Marrow Transplant 2015;21(3):389.e1-401.e1.

31. Cordonnier C, Maury S, Ribaud P, et al. A grading system based on severity of infection to predict mortality in allogeneic stem cell transplant recipients. Transplantation 2006;82:86-92.

32. Merkel EC, Mitchell SA, Lee SJ. Content validity of the Lee Chronic Graft-versus-Host Disease Symptom Scale as assessed by cognitive interviews. Biol Blood Marrow Transplant 2016;22:752-8.

33. Palmer J, Chai X, Pidala J, et al. Predictors of survival, nonrelapse mortality, and failure-free survival in patients treated for chronic graft-versus-host disease. Blood 2016;127:160-6.

34. Martin PJ, Storer BE, Inamoto Y, et al. An endpoint associated with clinical benefit after initial treatment of chronic graft-versus-host disease. Blood 2017;130:360-7.

35. Verstovsek S, Mesa RA, Gotlib J, et al. A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis. N Engl J Med 2012;366:799-807.

36. Harrison C, Kiladjian J-J, Al-Ali HK, et al. JAK inhibition with ruxolitinib versus best available therapy for myelofibrosis. N Engl J Med 2012;366:787-98.

37. Polverelli N, Palumbo GA, Binotto G, et al. Epidemiology, outcome, and risk factors for infectious complications in myelofibrosis patients receiving ruxolitinib: a multicenter study on 446 patients. Hematol Oncol 2018 April 6 (Epub ahead of print).

38. Olivieri J, Manfredi L, Postacchini L, et al. Consensus recommendations for improvement of unmet clinical needs — the example of chronic graft-versus-host disease: a systematic review and meta-analysis. Lancet Haematol 2015;2(7):e297-e305.

39. Drexler B, Buser A, Infanti L, Stehle G, Halter J, Holbro A. Extracorporeal photopheresis in graft-versus-host disease. Transfus Med Hemother 2020;47:214-25.

40. Flowers MED, 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;112:2667-74.