Biol Blood Marrow Transplant 25 (2019) 1869–1874

Biology of Blood and Marrow Transplantation

journal homepage: www.bbmt.org

Analysis

Organ Changes Associated with Provider-Assessed Responses in Patients with Chronic Graft-versus-Host Disease

Paul J. Martin[1,2,*], Barry E. Storer[1], Jeanne Palmer[3], Madan H. Jagasia[4], George L. Chen[5], Raewyn Broady[6], Mukta Arora[7], Joseph A. Pidala[8], Betty K. Hamilton[9], Stephanie J. Lee[1,2]

[1] *Clinical Research Division, Fred Hutchinson Cancer Center, Seattle, Washington*
[2] *Department of Medicine, University of Washington, Seattle, Washington*
[3] *Mayo Clinic, Phoenix, Arizona*
[4] *Division of Hematology/Oncology, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee*
[5] *Department of Medicine, Roswell Park Cancer Institute, Buffalo, New York*
[6] *Leukemia/Bone Marrow Transplant Program of British Columbia, BC Cancer Agency, Vancouver, British Columbia, Canada*
[7] *Hematology, Oncology and Transplantation, University of Minnesota, Minneapolis, Minnesota*
[8] *Blood and Marrow Transplantation, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida*
[9] *Blood and Marrow Transplantation, Department of Hematology and Medical Oncology, Taussig Cancer Institute, Cleveland Clinic, Cleveland, Ohio*

Article history:
Received 1 April 2019
Accepted 6 May 2019

Key Words:
Chronic graft-versus-host disease
Hematopoietic cell transplantation
Treatment response

A B S T R A C T

Assessments of overall improvement and worsening of chronic graft-versus-host disease (GVHD) manifestations by the algorithm recommended by National Institutes of Health (NIH) response criteria do not align closely with those reported by providers, particularly when patients have mixed responses with improvement in some manifestations but worsening in others. To elucidate the changes that influence provider assessment of response, we used logistic regression to generate an overall change index based on specific manifestations of chronic GVHD measured at baseline and 6 months later. We hypothesized that this overall change index would correlate strongly with overall improvement as determined by providers. The analysis included 488 patients from 2 prospective observational studies who were randomly assigned in a 3:2 ratio to discovery and replication cohorts. Changes in bilirubin and scores of the lower gastrointestinal tract, mouth, joint/fascia, lung, and skin were correlated with provider-assessed improvement, suggesting that the main NIH response measures capture relevant information. Conversely, changes in the eye, esophagus, and upper gastrointestinal tract did not correlate with provider-assessed response, suggesting that these scales could be modified or dropped from the NIH response assessment. The area under the receiver operator characteristic curve in the replication cohort was 0.72, indicating that the scoring algorithm for overall change based on NIH response measures is not well calibrated with provider-assessed response.

INTRODUCTION

The clinical management of patients with chronic graft-versus-host disease (GVHD) entails an evaluation of whether manifestations have improved or worsened since the previous evaluation [1]. Overall improvement would typically prompt a reduction in the intensity of immunosuppressive treatment, especially in decreasing glucocorticoid doses as much as possible to limit toxicity. Overall worsening would prompt providers to consider increasing the intensity of immunosuppressive treatment, either by raising the glucocorticoid dose or by adding a new agent or replacing an existing agent with an alternative agent that could be more effective [2]. Thus, clinician perception of patient response is a critical variable when managing patients, and ideally, it should correlate with the endpoints used in clinical trials.

The National Institutes of Health (NIH) Consensus Conference on Clinical Trials for Treatment of Chronic GVHD has provided scales that can be used to evaluate whether changes in each of the various manifestations of chronic GVHD represent improvement or worsening [3,4]. The NIH Consensus Conference has also recommended an algorithm that can be used to determine overall improvement or worsening [5]. Although the field has reached good agreement with respect to the scales used to assess individual manifestations of chronic GVHD based on empirical data reflecting meaningful changes [6,7], additional work is needed to assess the validity of the

Financial disclosure: See Acknowledgments on page 1873.

* Correspondence and reprint requests: Paul J. Martin, MD, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, D2-100, PO Box 19024, Seattle, WA 98109-1024.
E-mail address: pmartin@fredhutch.org (P.J. Martin).

https://doi.org/10.1016/j.bbmt.2019.05.008
1083-8791/

algorithm used to determine overall improvement or worsening. In this respect, it is troubling that assessments of overall improvement and worsening by the algorithm recommended by NIH response criteria do not align closely with those reported by providers, particularly when patients have mixed responses with improvement in some manifestations but worsening in others, a situation considered by the NIH criteria as lack of response [8].

The goal of this study was to develop and evaluate a simple algorithm based on objective measures that could be used reliably to assess improvement versus lack of improvement at 6 months after the baseline in patients with chronic GVHD, using provider assessments of overall response as the gold standard. We used logistic regression to generate an overall change index based on the sum of individually weighted change scores from objective scales measuring specific manifestations of chronic GVHD at baseline and 6 months later. We hypothesized that this overall change index would correlate strongly with overall improvement as actually reported by providers as external validation that the NIH response criteria are well calibrated and clinically meaningful.

METHODS

Patients

Patients were enrolled in 1 of 2 prospective, multicenter observational studies of patients with chronic GVHD. The "Chronic GVHD Consortium Improving Outcomes Assessment in Chronic GVHD" (Outcomes Assessment) study enrolled 601 patients with chronic GVHD between 2007 and 2012 [9]. Patients in this study enrolled at any time after starting systemic treatment for chronic GVHD. At enrollment and every 6 months thereafter, providers and patients recorded standardized information regarding current chronic GVHD organ involvement and symptoms using forms developed according to the 2005 NIH chronic GVHD consensus criteria [3]. For incident cases, providers and patients also recorded the same information at 3 months after enrollment. The "Chronic GVHD Consortium Response Measures Validation Study" (Response Measures Validation study) enrolled 383 patients with chronic GVHD between 2013 and 2017 [10]. Patients in this study enrolled within 4 weeks before or after starting a new systemic treatment for chronic GVHD. At enrollment and at 3, 6, and 18 months thereafter, providers and patients recorded standardized information according to the 2014 NIH consensus conference criteria [4].

Exclusion criteria in both studies included primary disease relapse and an inability to comply with study procedures. In both studies, patients rated the severity of GVHD symptoms according to the Lee Chronic GVHD Symptom Scale [11]. At each assessment, providers and patients rated overall changes in GVHD manifestations from baseline according to an 8-point scale with categories of "completely gone," "very much better," "moderately better," "a little better," "about the same," "a little worse," "moderately worse," or "a lot worse." The protocols were approved by the institutional review board at each site, and all patients provided informed consent in accordance with the Declaration of Helsinki.

Statistical Analysis

The analysis included patients for whom baseline and 6-month evaluations of chronic GVHD manifestations were available and excluded patients who had primary disease relapse between enrollment and the 6-month assessment. The 488 eligible patients were randomly assigned in a 3:2 ratio to a discovery cohort (n = 296) and a replication cohort (n = 192). The study was designed to identify changes in objective measures associated with overall improvement assessed by providers as chronic GVHD manifestations that were "completely gone," "very much better," or "moderately better" versus no overall improvement ("a little better," "about the same," "a little worse," "moderately worse," or "a lot worse") because this categorization was felt to correlate with the recognition that a complete or partial response was clinically meaningful. A stepwise regression analysis was used to select significant change variables (ie, factors) for building multivariate models of improvement, using data from the discovery cohort and a P value threshold of $<.05$. If significant factors were closely related (eg, changes in serum alanine aminotransferase [ALT] concentration or the ALT/upper limit of normal ratio), only the most strongly associated variable was considered for inclusion in the multivariate analysis. Final factors in the model were based on a forward selection procedure.

Model parameters (log-odds ratio coefficients) were used to define weights per unit change for each factor in the model, yielding a sum representing an overall chronic GVHD activity change index for each patient. Correlation of the chronic GVHD activity change index with binary improvement versus no improvement outcomes was analyzed as the area under the receiver operator characteristic curve (AUC) in the discovery cohort, and results were further evaluated by testing the same model and weights in the replication cohort. We used a similar approach to assess the extent to which abnormalities present at 6 months were associated with the provider overall assessment without considering change from baseline.

RESULTS

Baseline Characteristics

Tables 1 and 2 summarize demographic and transplant characteristics of patients and the baseline manifestations of chronic GVHD [12] in the discovery and verification cohorts, and Supplemental Table S1 lists the centers that enrolled patients in the 2 studies that contributed to the analysis. As expected by the randomization of patients between the discovery and verification cohorts, the demographic characteristics were similar in the 2 groups (Table 1). Likewise, the proportions of patients with abnormalities at baseline and the mean values of measures among patients with abnormal values were similar in the 2 groups (Table 2).

Changes Associated with Overall Improvement as Assessed by Providers

In univariate analysis, changes in 11 objective factors were associated with overall improvement when providers assessed chronic GVHD manifestations at 6 months compared with baseline (Table 3). These included decreases in the serum alkaline phosphatase, ALT, and total bilirubin concentrations; lower gastrointestinal tract score; oral lichenoid score; oral summary score; joint range-of-motion summary score; joint/fascia score; lung score; mouth score; and skin score. Six of

Table 1
Demographic Characteristics of Patients in the Discovery and Replication Cohorts

Characteristic	Discovery	Replication
Patient age at initial treatment, median (range), yr	54 (12-78)	55 (2-79)
Patient sex, n (%)		
Male	180 (61)	115 (60)
Female	116 (39)	77 (40)
Donor-patient sex combination, n (%)		
Female to male	78 (27)	50 (27)
Other	216 (73)	138 (73)
Diagnosis, n (%)		
Myeloid malignancy	162 (55)	108 (56)
Lymphoid malignancy	103 (35)	62 (32)
Other/nonmalignant	31 (10)	22 (11)
Conditioning regimen, n (%)		
High dose with or without total body irradiation	145 (49)	84 (44)
Reduced intensity or nonmyeloablative	150 (51)	107 (56)
Graft source, n (%)		
Bone marrow	19 (6)	11 (6)
Mobilized blood cells	270 (91)	174 (91)
Cord blood	7 (2)	6 (3)
Donor and HLA type, n (%)		
HLA-matched related	117 (40)	58 (30)
HLA-matched unrelated	141 (48)	102 (53)
HLA antigen or allele-mismatched related	5 (2)	5 (3)
HLA antigen or allele-mismatched unrelated	33 (11)	26 (14)

Table 2
Baseline Chronic GVHD Measures in the Discovery and Replication Cohorts

	Discovery (n = 296)			Replication (n = 192)		
Measure (Scale Range*)	n	% Abnormal	Mean†	n	% Abnormal	Mean†
Alkaline phosphatase, U/dL	286	32	276.8	184	35	236.5
Alkaline phosphatase/upper limit of normal ratio	286	32	3.5	184	35	3.2
Alanine aminotransferase, U/dL	287	40	209.2	187	44	200.5
Alanine aminotransferase/upper limit of normal ratio	287	40	4.0	187	44	3.9
Total serum bilirubin, mg/dL	287	8	3.9	187	9	2.7
Total serum bilirubin/upper limit of normal ratio	287	8	3.0	187	9	2.1
Esophagus score (0-3)	296	14	1.4	192	22	1.4
Lower gastrointestinal tract score (0-3)	294	14	1.4	192	14	1.5
Oral erythema score (0-3)	295	41	1.4	192	46	1.4
Oral lichenoid score (0-3)	296	61	1.6	191	62	1.6
Oral sum score (0-12)	296	69	3.2	192	68	3.6
Oral ulcer score (0-6)	295	18	3.0	190	19	3.5
Upper gastrointestinal tract score (0-3)	296	17	1.5	192	21	1.6
Ankle range-of-motion score (1-4)	265	26	2.8	163	26	2.8
Elbow range-of-motion score (1-7)	269	19	5.5	165	20	5.5
Shoulder range-of-motion score (1-7)	269	24	5.5	167	26	5.4
Wrist and finger range-of-motion score (1-7)	270	31	5.1	166	36	5.0
Joint range-of-motion sum score (0-25)	263	44	21.6	162	49	21.5
Eye score (0-3)	296	53	1.5	191	59	1.4
Genital score (0-3)	207	15	1.4	135	13	1.7
Gastrointestinal tract score (0-3)	296	34	1.4	191	36	1.3
Joints/fascia score (0-3)	296	43	1.4	191	43	1.5
Lung score (0-3)	296	26	1.2	192	24	1.4
Mouth score (0-3)	296	59	1.3	192	63	1.3
Skin score (0-3)	296	70	2.2	192	74	2.2
Forced expiratory volume in first second, % of predicted	149	21	62.7	85	29	62.9
Summary symptom score reported by patient (0-100)	251	100	24.0	167	99	23.6

* In all scales, values of 0 indicate no abnormality attributed to GVHD.
† Among patients with abnormal values.

these change factors remained significant in the multivariate logistic regression model: total serum bilirubin concentration, lower gastrointestinal tract score, oral lichenoid score, joint/fascia score, lung score, and skin score (Table 4). With these 6 factors, the chronic GVHD change index had an AUC of 0.79 for its association with improvement versus no improvement in the discovery cohort and 0.72 for this association in the replication cohort. Results were similar in models that identified changes associated with provider assessments of overall complete or partial response versus stable disease or progression (data not shown). Notably, 6 of 9 organs used in the NIH response assessment were independently associated with provider-assessed response. Of the 3 that were not, 2 were less common manifestations (esophagus, 14% to 22%; upper gastrointestinal tract, 17% to 21%). Changes in the eye score were not statistically correlated with provider-assessed response.

Abnormalities That Have Causes Other Than Chronic GVHD

When asked to assess response to treatment for GVHD, providers might instinctively distinguish between manifestations caused by chronic GVHD and abnormalities that have other causes [13]. The 2014 NIH response criteria capture whether organ dysfunction is solely attributed to non-GVHD causes and recommended exclusion of these organs in the global calculation. To test whether this distinction affected provider assessments, we analyzed whether covariate associations with overall improvement assessed by providers change when abnormalities not attributed to chronic GVHD were excluded from consideration. Providers were asked to make such distinctions in the Response Measures Validation study but not in the earlier Outcomes Assessment study. Therefore, the analysis of provider assessments included only the patients enrolled in the Response Measures Validation study (n = 318). The results showed no evidence that omission of abnormalities attributed to causes other than chronic GVHD affected the association of covariates with overall improvement assessed by providers, suggesting that it is not necessary to exclude these organs when calculating responses per NIH criteria (Supplemental Table S1).

Abnormalities Present at 6 Months without Consideration of Change from Baseline

We considered the possibility that the overall provider assessment was influenced as much by abnormalities at 6 months as by changes from baseline. In univariate analysis, 9 objective factors present at 6 months were associated with overall improvement when providers assessed chronic GVHD manifestations at 6 months (Table 5). These included decreases in serum alkaline phosphatase, esophagus score, oral lichenoid score, oral sum score, range-of-motion sum score, eye score, joint/fascia score, lung score, and skin score. Five factors remained significant in the multivariate logistic regression model: serum alkaline phosphatase concentration, oral lichenoid score, joint range-of-motion sum score, lung score, and skin score. With these 5 factors, the chronic GVHD change index had an AUC of 0.84 for association with improvement versus no improvement in the discovery cohort and 0.73 for this association in the replication cohort. The oral lichenoid

Table 3
Changes in Measures Associated with Overall Improvement as Assessed by Providers*

Measure	Number Evaluated	χ^2	*P*
Alkaline phosphatase, U/dL	268	8.55	.004
Alkaline phosphatase/upper limit of normal ratio	268	1.75	.19
Alanine aminotransferase, U/dL	268	8.56	.003
Alanine aminotransferase/upper limit of normal ratio	268	6.91	.009
Total serum bilirubin, mg/dL	269	11.25	.001
Total serum bilirubin/upper limit of normal ratio	269	11.16	.001
Esophagus score	283	1.97	.16
Lower gastrointestinal tract score	281	5.33	.02
Oral erythema score	279	3.94	.05
Oral lichenoid score	281	20.49	<.0001
Oral sum score	283	7.74	.005
Oral ulcer score	282	0.30	.58
Upper gastrointestinal tract score	283	0.01	.91
Ankle range-of-motion score	240	3.02	.08
Elbow range-of-motion score	247	2.69	.10
Shoulder range-of-motion score	248	0.14	.71
Wrist and finger range-of-motion score	247	0.53	.47
Range-of-motion sum score	235	4.43	.04
Eye score	283	3.03	.08
Genital score	154	0.26	.61
Gastrointestinal tract score	282	1.62	.20
Joints/fascia score	282	6.40	.01
Lung score	283	10.65	.001
Mouth score	283	4.19	.04
Skin score	283	26.52	<.0001
Forced expiratory volume in first second, % of predicted	86	1.25	.26
Summary symptom score reported by patient	214	3.74	.05

* Univariate analysis in the discovery cohort.

Table 4
Weights per Unit Change in Measures Associated with the Provider Assessment of Overall Improvement*

Measure†	Weight per Unit Change‡
Bilirubin	−0.763
Lower gastrointestinal tract score	−0.687
Oral lichenoid score	−0.829
Joint/fascia score	−0.713
Lung score	−0.550
Skin score	−0.409

* Discovery cohort.
† In each of these scales, higher values reflect more severe manifestations of chronic GVHD.
‡ Weights are multiplied by the difference between the 6-month assessment value and the baseline value. For example, the value of a change from a score of 3 at baseline to 1 at 6 months is −2. For the lower gastrointestinal score, −2 is multiplied by −0.687, yielding a weighted improvement score of 1.382. The change index is the sum of weighted scores in all 6 measures.

score, lung score, and skin score were the identical factors identified by the change score. Two additional variables were similar, although the parameter selected by the model differed: liver involvement and joint/fascia scores.

Table 5
Measures at 6 Months Associated with Overall Improvement as Assessed by Providers*

Measure	Number Evaluated	χ^2	*P*
Alkaline phosphatase, U/dL	276	6.62	.01
Alkaline phosphatase/upper limit of normal ratio	276	4.23	.04
Alanine aminotransferase, U/dL	275	1.40	.24
Alanine aminotransferase/upper limit of normal ratio	275	2.42	.12
Total serum bilirubin, mg/dL	276	0.32	.57
Total serum bilirubin/upper limit of normal ratio	276	0.07	.80
Esophagus score	283	4.53	.03
Lower gastrointestinal tract score	283	2.10	.15
Oral erythema score	280	7.86	.005
Oral lichenoid score	281	8.82	.003
Oral sum score	283	8.15	.004
Oral ulcer score	283	1.74	.19
Upper gastrointestinal tract score	283	1.28	.26
Ankle range-of-motion score	253	14.04	.0002
Elbow range-of-motion score	256	25.34	<.0001
Shoulder range-of-motion score	257	29.44	<.0001
Wrist and finger range-of-motion score	250	38.39	<.0001
Range-of-motion sum score	255	16.80	<.0001
Eye score	283	6.42	.01
Genital score	185	2.11	.15
Gastrointestinal tract score	282	2.22	.14
Joints/fascia score	282	42.89	<.0001
Lung score	283	21.44	<.0001
Mouth score	283	3.23	.07
Skin score	283	70.04	<.0001
Forced expiratory volume in first second, % of predicted	110	5.46	.02
Summary symptom score reported by patient	214	3.74	.05

* Univariate analysis in the discovery cohort.

Association of Provider Responses with Survival after the 6-Month Assessment

Neither the provider response derived from the 8-point scale nor the overall NIH response is predictive of subsequent overall survival. For the provider "completely gone," "very much better" or "moderately better" versus "a little better," "about the same," "a little worse," "moderately worse" or "a lot worse" categories, the hazard ratio of mortality was 0.90 (95% confidence interval, 0.60 to 1.36; $P = .62$) (92 events in 459 patients). For the NIH complete or partial response versus stable or progressive disease categories, the hazard ratio was 0.88 (95% confidence interval, 0.58 to 1.33; $P = .54$) (94 events in 480 patients).

DISCUSSION

Results of this study have identified a set of objective change measures associated with overall response assessments by providers caring for patients with chronic GVHD, validating the components of the NIH response assessment, but further work is needed to improve the sensitivity and specificity parameters of this association. We did not find evidence that the association of change measures with overall response was appreciably confounded by abnormalities unrelated to chronic GVHD.

Several explanations might account for the finding that AUC values for the change index in the replication cohort did not exceed 0.75. An AUC value of 0.5 indicates no correlation, and a value of 1.0 indicates perfect correlation. An AUC value of 0.75 sits halfway between no correlation and perfect correlation. First, chronic GVHD manifestations can contribute to the logistic regression model only if they occur at an appreciable frequency and have a high probability of change between the baseline and the assessment. For example, transaminase elevations occur frequently and are readily reversible, whereas esophageal abnormalities occur much less frequently and are less likely to improve. Manifestations that are infrequent and measures that are unlikely to change within 6 months lack power to contribute to the model, even though they may determine provider or patient response assessments in some individual cases. The scale definitions also may be an issue, because an esophageal score of 3 is assigned if dilation is required, and eye scores may reflect punctual plugging and use of corneal protective devices. Strict application of these definitions means that anyone who has ever required esophageal dilation cannot have a score of 2 or less, and once punctual plugs are placed, an eye score of 1 or less is impossible, even if symptoms resolve. Also, the esophageal and upper gastrointestinal tract severity scores have never been empirically validated according to provider and patient assessments, as has been done for skin, mouth, eye, gastrointestinal, liver, and joint scores [14–19].

Second, accurate response assessment requires equivalent awareness of current and baseline organ measures. For providers and patients, awareness of current measures far exceeds the awareness of baseline measures. Because of memory bias, it is likely that response assessments are driven more by changes between the most recent visit and the current visit than by changes between the baseline and the current visits. This memory bias might be especially applicable when systemic treatment changed between the baseline and the 6-month assessment. We attempted to analyze this possibility by excluding patients with a change of systemic treatment before the 6-month assessment. The results were uninformative because of insufficient statistical power (data not shown).

Third, mental algorithms used by providers to assess overall response are likely to be inconsistent. Providers are likely to differ in the weights assigned to current symptoms caused, for example, by oral ulcers, versus asymptomatic evidence of progressive fibrotic changes that could lead to future disability. If oral ulcers are resolving but asymptomatic fibrotic manifestations are progressing, some providers might decrease the intensity of immunosuppressive treatment, whereas others might increase immunosuppressive treatment. The only way to test this possibility would be to sort patients by provider to determine whether individual providers apply distinctly different weights from each other. Our study did not have enough patients consistently treated by any 2 providers to compare the weights between them.

The assessment of overall response in patients with chronic GVHD is unambiguous when at least 1 manifestation has demonstrable improvement and no other manifestations have demonstrable worsening. The situation is far more complex when some manifestations have demonstrable improvement and others have demonstrable worsening, and it may indeed be correct that cases with major improvement in some manifestations and minor worsening in others should be categorized as overall improvement when deciding whether changes in treatment are indicated. The NIH criteria would consider these cases in the same category as those who have major worsening in all organs. The logistic regression model used to derive weights for unit changes in each scale in the current analysis incorporates an assumption that a change from 0 to 1 (change from no abnormality to abnormality with no functional effect) carries the same weight as a change from 2 to 3 (change from abnormality with moderate functional effect to abnormality with severe functional effect). The 2014 NIH consensus conference on measurement of response acknowledged that this equivalence is not appropriate in recognizing that many score changes from 0 to 1 had little clinical significance [4]. More flexible modeling approaches might be able to recognize and adjust for scale differences among organs.

The current results show some correlation between recorded changes in the objective clinical and subjective symptom data, but the overall assessment of improvement versus no improvement cannot be fully explained by measures that we considered in this study. Whether this result reflects measurement error in the explanatory or outcome variables or cognitive biases cannot be addressed by these data. Results of a previous analysis showed that the NIH response criteria do not closely agree with overall assessments by providers [8]. If the provider assessment resulted in a change of treatment, most clinical trials would categorize the outcome as failure, even if the NIH response had not shown evidence of progression. These considerations highlight the need for better understanding of the inputs that drive provider assessments of overall response. The current results may be helpful in clinical practice by identifying a limited number of chronic GVHD manifestations associated with provider assessments, but it would be premature to use scores derived from the scales and weights in the current study for making decisions in the clinical management of patients with chronic GVHD.

Future efforts toward this goal would benefit from methods to ensure that providers carefully compare serial values to minimize recall bias before reaching conclusions about overall improvement or worsening. Future efforts would also benefit from more complex analytic methods such as machine learning that could accommodate different weights assigned to unit changes from baseline anchors across the range of each scale [20]. The success of such efforts will certainly be limited by the inherent variability of provider judgment even when evaluating the same clinical scenario. Nonetheless, it may yet be possible to derive an algorithm of objective change measures that reflects the clinical judgment of overall improvement versus no improvement by most providers. In summary, our results support the current NIH response measures as relevant and reasonably scaled, but we also identify areas for improvement.

ACKNOWLEDGMENTS

The authors thank Lynn Onstad for data management and Mary Flowers, Paul Carpenter, Corey Cutler, Sally Arai, Stefanie Sarantopoulos, and Amin Alousi for contributing data.

Financial disclosure: This research was supported by National Institutes of Health, National Cancer Institute grant CA118953 (S.J.L.).

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

Authorship statement: S.J.L. designed the study. B.E.S. performed statistical analysis. P.J.M., B.E.S., and S.J.L. analyzed results and wrote the manuscript. J.P., M.H.J., G.L.C., R.B., M.A., J.A.P., B.K.H., and S.J.L. contributed data. All authors reviewed the manuscript for critical content.

SUPPLEMENTARY MATERIALS

Supplementary data related to this article can be found online at doi:10.1016/j.bbmt.2019.05.008.

REFERENCES

1. Cutler CS, Koreth J, Ritz J. Mechanistic approaches for the prevention and treatment of chronic GVHD. *Blood*. 2017;129:22–29.
2. MacDonald KP, Hill GR, Blazar BR. Chronic graft-versus-host disease: biological insights from preclinical and clinical studies. *Blood*. 2017;129:13–21.
3. Pavletic SZ, Martin P, Lee SJ, et al. Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2006;12:252–266.
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. The 2014 Response Criteria Working Group report. *Biol Blood Marrow Transplant*. 2015;21:984–999.
5. Kerep AZ, Broome J, Pirsl F, et al. Impact of the 2014 NIH chronic graft-versus-host disease scoring criteria modifications assessed in a large cohort of severely affected patients. *Bone Marrow Transplantation*. 2019;54:76–84.
6. Schoemans HM, Goris K, Van Durm R, et al. The eGVHD app has the potential to improve the accuracy of graft-versus-host disease assessment: a multicenter randomized controlled trial. *Haematologica*. 2018;103:1698–1707.
7. Curtis LM, Pirsl F, Steinberg SM, et al. Predictors for permanent discontinuation of systemic immunosuppression in severely affected chronic graft-versus-host disease patients. *Biol Blood Marrow Transplant*. 2017;23:1980–1988.
8. Palmer JM, Lee SJ, Chai X, et al. Poor agreement between clinician response ratings and calculated response measures in patients with chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2012;18:1649–1655.
9. 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.
10. 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.
11. 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.
12. Lee SJ. Classification systems for chronic graft-versus-host disease. *Blood*. 2017;129:30–37.
13. Aki SZ, Inamoto Y, Carpenter PA, et al. Confounding factors affecting the National Institutes of Health (NIH) chronic graft-versus-host disease organ-specific score and global severity. *Bone Marrow Transplant*. 2016;51:1350–1353.
14. Jacobsohn DA, Kurland BF, Pidala J, et al. Correlation between NIH composite skin score, patient-reported skin score, and outcome: results from the Chronic GVHD Consortium. *Blood*. 2012;120:2545–2552. quiz 2774.
15. Treister N, Chai X, Kurland B, et al. Measurement of oral chronic GVHD: results from the Chronic GVHD Consortium. *Bone Marrow Transplant*. 2013;48:1123–1128.
16. Inamoto Y, Chai X, Kurland BF, et al. Validation of measurement scales in ocular graft-versus-host disease. *Ophthalmology*. 2012;119:487–493.
17. Pidala J, Chai X, Kurland BF, et al. Analysis of gastrointestinal and hepatic chronic grant-versus-host disease manifestations on major outcomes: a chronic grant-versus-host disease consortium study. *Biol Blood Marrow Transplant*. 2013;19:784–791.
18. Palmer J, Williams K, Inamoto Y, et al. Pulmonary symptoms measured by the National Institutes of Health lung score predict overall survival, nonrelapse mortality, and patient-reported outcomes in chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2014;20:337–344.
19. Inamoto Y, Pidala J, Chai X, et al. Assessment of joint and fascia manifestations in chronic graft-versus-host disease. *Arthritis Rheumatol*. 2014;66(4):1044–1052.
20. Gandelman JS, Byrne MT, Mistry AM, et al. Machine learning reveals chronic graft-versus-host disease phenotypes and stratifies survival after stem cell transplant for hematologic malignancies. *Haematologica*. 2019;104:189–196.