TRANSPLANTATION

Predictors of survival, nonrelapse mortality, and failure-free survival in patients treated for chronic graft-versus-host disease

Jeanne Palmer,[1] Xiaoyu Chai,[2] Joseph Pidala,[3] Yoshihiro Inamoto,[2,4] Paul J. Martin,[2] Barry Storer,[2] Iskra Pusic,[5] Mary E. D. Flowers,[2] Mukta Arora,[6] Steven Z. Pavletic,[7] and Stephanie J. Lee[2]

[1]Division of Hematology/Oncology, Mayo Clinic, Phoenix, AZ; [2]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [3]Blood and Marrow Transplantation, Moffitt Cancer Center, Tampa, FL; [4]Division of Hematopoietic Stem Cell Transplantation, National Cancer Center Hospital, Tokyo, Japan; [5]Division of Hematology/Oncology, Washington University, St. Louis, MO; [6]Blood and Marrow Transplant Program, University of Minnesota, Minneapolis, MN; and [7]National Cancer Institute, National Institutes of Health, Bethesda, MD

Key Points

- Survival of chronic GVHD patients was predicted by clinician-assessed response and changes in patient-reported outcomes.
- FFS was predicted by clinician-assessed response, changes in patient-reported outcomes, and the 2014 NIH response criteria.

Chronic graft-versus-host disease (GVHD) is a pleotropic syndrome that lacks validated methods of measuring response in clinical trials, although several end points have been proposed. To investigate the prognostic significance of these proposed end points, such as the 2005 National Institutes of Health (NIH) response measures, 2014 NIH response measures, clinician-reported response, and patient-reported response, we tested their ability to predict subsequent overall survival (OS), nonrelapse mortality (NRM), and failure-free survival (FFS). Patients (n = 575) were enrolled on a prospective chronic GVHD observational trial. At 6 months, clinician-reported response (P = .004) and 2014 NIH-calculated response (P = .001) correlated with subsequent FFS, and clinician-reported response predicted OS (P = .007). Multivariate models were used to identify changes in organ involvement, laboratory values, and patient-reported outcomes that were associated with long-term outcomes. At 6 months, a change in the 2005 NIH 0 to 3 clinician-reported skin score and 0 to 10 patient-reported itching score predicted subsequent FFS. Change in the Lee skin symptom score and Functional Assessment of Cancer Therapy–Bone Marrow Transplant score predicted subsequent OS. Change in the Lee skin symptom score predicted subsequent NRM. This study provides evidence that clinician-reported response and patient-reported outcomes are predictive of long-term survival. The trial was registered at www.clinicaltrials.gov as #NCT00637689. (*Blood*. 2016;127(1):160-166)

Introduction

Chronic graft-versus-host disease (GVHD) is a significant cause of morbidity and mortality in survivors of allogeneic hematopoietic cell transplantation (HCT),[1-8] and more effective treatments are needed. Although many clinical trials have been conducted, interpretation of results has been difficult because documentation of response in chronic GVHD has been particularly challenging. Because of the long time course of chronic GVHD, standard end points such as overall survival (OS) and nonrelapse mortality (NRM) require longer-term follow-up than might be desired in most early phase chronic GVHD trials. Therefore, efforts have been made to identify an interim response measure, such as failure-free survival (FFS). Thus far, there are no validated response measures; therefore, subjective clinical judgment is often used to determine response.[9-16]

In 2005, the National Institutes of Health (NIH) Chronic GVHD Consensus Conference recommended response measures based on serial organ assessments.[17,18] Response is determined by comparing baseline and follow-up scores for each organ system to calculate overall response. In analysis of individual organ systems, this scoring system has been predictive of meaningful end points such as OS and NRM.[19-24] However, when imputed into a calculated composite response, measures have had variable correlation with clinician-reported response and survival outcomes. Although 1 study demonstrated good correlation between response calculated per the 2005 NIH consensus criteria and clinician-reported response, and a survival advantage to those patients who had a favorable response (partial response [PR] or better) at 6 months, this study had a small number of patients evaluated.[25] In other studies, these response criteria have not been associated with clinician-reported response,[26] subsequent survival, or improved quality of life.[27,28]

In 2014, a second NIH consensus conference was held.[29] Several changes to the NIH response algorithm were made. Notably, skin, mouth, and eye measurements were simplified; new joint measures were introduced; new mild symptoms in gastrointestinal (GI) and liver were not considered progression; and attribution of clinical manifestations (symptoms or signs) to causes other than chronic GVHD was captured and incorporated into scoring. The 2014 NIH response criteria have not been used in clinical trials yet.

FFS is a proposed intermediate end point of treatment success, defined as continued disease-free survival without addition of a

Submitted August 5, 2015; accepted October 2, 2015. Prepublished online as *Blood* First Edition paper, November 2, 2015; DOI 10.1182/blood-2015-08-662874.

The online version of this article contains a data supplement.

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

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

new systemic immunosuppressive medication. FFS is appealing because it is easy to document and change of therapy has been associated with a higher mortality rate,[30] but FFS also has the disadvantage of relying on the clinician's treatment approach, which is subject to bias and variation in management styles. OS and NRM are attractive end points because they are definitive and acceptable to the US Food and Drug Administration, but survival is typically not the primary end point in trials of chronic illnesses such as chronic GVHD.[29]

We tested whether different aggregate measurements of response, including the 2005 and 2014 calculated NIH response, clinician-reported response, and patient-reported response measures were predictive of FFS, OS, and NRM. We also tested whether changes in individual organ assessments, laboratories, or patient-reported symptoms were predictive of FFS, OS, and NRM.

Methods

Chronic GVHD Consortium

A cohort of HCT recipients affected by chronic GVHD was enrolled in a multicenter observational study (#NCT00637689).[31] Chronic GVHD was diagnosed according to 2005 NIH consensus criteria.[32] The protocol was approved by the institutional review board at each site, and all subjects provided written informed consent. Patients enrolled in the cohort were allogeneic HCT recipients at least 2 years of age with chronic GVHD requiring systemic therapy, including both classic and overlap subtypes. For this analysis, only adults were included because of the small number of pediatric patients. Cases were classified as incident (enrollment <3 months after chronic GVHD diagnosis) or prevalent (enrollment 3 or more months after chronic GVHD diagnosis but <3 years after transplantation). Primary disease relapse, inability to comply with study procedures, and anticipated survival of <6 months were exclusion criteria. At enrollment and every 6 months thereafter, clinicians and patients reported standardized information summarizing chronic GVHD organ involvement and symptoms. Incident cases had an additional assessment time point at 3 months after enrollment.[15] Objective medical data including ancillary testing, laboratory results and medical complications, and medication profiles were abstracted through standardized chart review after each visit.

End point definitions

For the FFS end point, failure was defined as malignancy relapse, death, or addition of a new immunosuppressive medication (eg, sirolimus, rituximab) or treatment (eg, extracorporeal photopheresis) intended for systemic treatment of chronic GVHD.[33] Determination of failure was made by 2 separate reviewers (J. Palmer and S.J.L.) independently, and discrepancies were resolved by discussion. Treatment with pulse high-dose solumedrol (ie, 500 mg or 1000 mg for several days) was considered as a treatment change; steroid dose increases within the standard range were not considered a treatment change, consistent with other reports.[34] Addition of topical therapies (eg, topical steroids, ophthalmic cyclosporine, topical GI steroids), supportive care treatments (eg, ursodeoxycholic acid), or systemic nonimmunosuppressive medications for management of GVHD involving specific organs (eg, montelukast or azithromycin for obstructive lung disease) was not considered a failure.

OS was defined as the time from enrollment to death from any cause. NRM was defined as time from enrollment to death with relapse as a competing risk.

Potential predictors

Overall response was determined in 3 ways: (1) NIH-calculated response was according to both the 2005[35] and 2014[29] consensus criteria algorithms that use changes in skin, mouth, eye, lungs, joints, GI, and liver measures to assign patients to the categories of complete response (CR), PR, stable disease (SD), and progressive disease (PD). Although the 2014 response criteria were not available when the study started, the relevant measures were collected in the study and available to calculate response using the 2014 algorithm. (2) Clinician-reported response was CR, PR, SD, and PD as reported on clinician-completed surveys. (3) Patients also reported whether their chronic GVHD was improving, stable, or worsening on a 7-point Likert scale. Supplemental Table 1 (see the *Blood* Web site) shows the cross tabulation of the 2014 calculated responses with the 2005 calculated responses, clinician-reported responses, and patient-reported responses.

To identify individual organ assessments, laboratory values, and patient-reported variables associated with FFS, OS, and NRM, univariable analyses used all available information. The complete list of variables may be found in supplemental Table 2. Briefly, all recommended measures from the 2005 NIH Consensus Conference on Clinical Trials in Chronic GVHD were included.[17,32,36] In addition, other scales used in prior clinical trials were also collected and analyzed (eg, Vienna skin score[37] and Hopkins scales[38]). Data regarding comorbidities, disease, and transplant characteristics were considered as potential predictors.

Statistical analysis

Cumulative incidence estimates of relapse, NRM, and addition of a new therapy as causes of failure were derived, treating each event as a competing risk for the other two.[39]

NIH-calculated, clinician-reported, and patient-reported changes in chronic GVHD disease activity were tested in landmark analyses for their ability to predict subsequent FFS, NRM, and survival after 3 and 6 months with $P < .01$ considered significant because of multiple testing. Only incident cases were included in the 3-month landmark analysis because data at 3 months were not collected for prevalent cases. Both incident and prevalent cases were included in the 6-month primary landmark analysis, but they were also tested separately. The FFS analysis was limited to patients who had not had a treatment change or recurrent malignancy before the landmark. Because FFS is a composite end point, we also tested whether response measures correlated with treatment change, considering relapse and death as competing risks. OS and NRM analyses included all patients who were alive without recurrent malignancy at the landmark, regardless of prior treatment change.

The analysis of organ assessments, laboratory values- and patient-reported variables was complicated because of the large number of variables. Supplemental Figure 1 shows how the large number of potential predictive covariates was reduced to the variables whose change scores independently predicted the outcomes of interest. Cox regression models were used to identify risk factors for various types of failure (FFS, NRM, mortality), using sequential selection processes within each organ because of the number of potential predictors (supplemental Table 2), many of which are correlated. First, the change in each variable was fit into a univariable Cox regression model while adjusting for the baseline value (supplemental Table 3). Second, within each organ system, factors associated with failure with a univariable likelihood ratio test P value $\leq .05$ were then fit into a multivariable model for each organ using a backward elimination procedure (supplemental Table 4). Lastly, variables still significant on multivariate analysis within each organ system were considered in building a multivariable Cox regression model including all organ systems. A backward elimination procedure was used to exclude risk factors until the P value of the likelihood ratio test for all remaining risk factors was $\leq .05$. The final multivariable model included all statistically significant variables (enrollment plus change scores) after backward elimination, as well as any statistically significant transplant characteristics (Table 3). We tested for interaction among significant variables and found none.

Agreement between overall responses at 3- or 6-month visit was tested by weighted κ statistic for ordinal measures with Fleiss-Cohen weights.[40] Empirical interpretation was used for the κ coefficient (0, no agreement; 0-0.2, slight agreement; 0.2-0.4 fair agreement; 0.4-0.6, moderate agreement; 0.6-0.8, substantial agreement; 0.8-1.0, almost perfect agreement). SAS/STAT version 9.3 (SAS Institute Inc., Cary, NC) and R version 2.15.2 (R Foundation for Statistical Computing, Vienna, Austria) were used for statistical analyses.

Table 1. Patient characteristics and outcomes

Characteristic/category	n	Count (%)
Study site	575	
Fred Hutchinson Cancer Research Center		253 (44%)
University of Minnesota		61 (11%)
Dana-Farber Cancer institute		65 (11%)
Stanford University Medical Center		74 (13%)
Vanderbilt University Medical Center		48 (8%)
Medical College of Wisconsin		23 (4%)
Washington University Medical Center		4 (1%)
Moffitt Cancer Center		39 (7%)
Memorial Sloan Kettering Cancer Center		8 (1%)
Case type	575	
Incident		342 (59%)
Prevalent		233 (41%)
Patient age at registration (y), median (range)	575	52 (19-79)
Patient gender	575	
Female		242 (42%)
Male		333 (58%)
Diagnosis	575	
AML		193 (34%)
ALL		62 (11%)
CML		30 (5%)
CLL		46 (8%)
MDS		89 (15%)
NHL		85 (15%)
HL		17 (3%)
MM		29 (5%)
AA		6 (1%)
Other		18 (3%)
Disease status	571	
Early		184 (32%)
Intermediate		248 (43%)
Advanced		139 (24%)
Transplant source	575	
Bone marrow		40 (7%)
Cord blood		28 (5%)
Peripheral blood		507 (88%)
Transplant type	571	
Myeloablative		297 (52%)
Nonmyeloablative		274 (48%)
Donor-patient CMV status	569	
Patient and donor CMV both negative		192 (34%)
Patient or donor CMV positive		377 (66%)
Donor-patient gender combination	569	
Female into male		167 (29%)
Others		402 (71%)
Donor match	573	
Matched related		238 (42%)
Matched unrelated		244 (43%)
Mismatched		91 (16%)
Prior grade 2-4 acute GVHD	575	
Yes		311 (54%)
No		264 (46%)
2005 NIH chronic GVHD global severity score	575	
Mild or less		53 (9%)
Moderate		302 (53%)
Severe		220 (38%)

	Estimate (95% CI)
FFS	
At 1 y after enrollment	45% (41%-49%)
At 2 y after enrollment	29% (25%-33%)
At 4 y after enrollment	11% (4%-22%)
Relapse	
At 1 y after enrollment	6% (4%-8%)
At 2 y after enrollment	10% (8%-13%)
At 4 y after enrollment	13% (10%-17%)
NRM	
At 1 y after enrollment	4% (3%-6%)
At 2 y after enrollment	6% (4%-8%)
At 4 y after enrollment	14% (7%-27%)
Survival	
At 1 y after enrollment	89% (86%-91%)
At 2 y after enrollment	81% (78%-85%)
At 4 y after enrollment	71% (66%-76%)

AA, aplastic anemia; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CI, confidence interval; CLL, chronic lymphocytic leukemia; CML, chronic myeloid leukemia; CMV, cytomegalovirus; HL, Hodgkin lymphoma; MDS, myelodysplastic syndrome; MM, multiple myeloma; NHL, non-Hodgkin lymphoma.

Results

Patients

Between August 2007 and January 2013, 575 patients were enrolled in this chronic GVHD prospective study (Table 1). Four hundred and fifty-one patients had evaluations at 6 months, and 307 of those patients were alive without recurrent malignancy or prior treatment change. There were 1856 follow-up visits, for a total of 2431 visits. The cohort included 342 (59%) incident cases and 233 (41%) prevalent cases. The median time to enrollment was 11.9 months after transplant (range 2.9-294), median follow-up after enrollment for survivors was 44 months (range 0.9-76), and 149 (26%) have died.

Landmark analysis

We first performed a landmark analysis evaluating how different response measurements at 3 months (incident cases only) or 6 months (both incident and prevalent cases) correlated with subsequent FFS, OS, and NRM (Table 2). At 3 months, clinician-reported (CR + PR vs SD + PD: HR = 0.34; 95% CI, 0.22-0.52; $P < .001$), patient-reported (improvement vs stable vs worsening: overall $P < .001$), and 2014 NIH-calculated (CR + PR vs SD + PD: HR = 0.60; 95% CI, 0.41-0.89; $P = .01$) response correlated with longer subsequent FFS but not with NRM or OS. Results were similar for treatment change. At 6 months, clinician-reported (CR + PR vs SD + PD: $P = .004$) and 2014 NIH-calculated (CR+PR vs SD + PD: HR = 0.58; 95% CI, 0.42-0.80; $P = .001$) response correlated with higher subsequent FFS but not NRM. Clinician-reported response was correlated with improved OS (HR = 0.55; 95% CI, 0.36-0.85; $P = .007$). The 2005 NIH-calculated response did not predict FFS, NRM, or OS but was predictive of treatment change (Table 2). The κ statistic between the 2005 and 2014 NIH-calculated responses was 0.32 suggesting poor to fair correlation. Kaplan-Meier plots for FFS and OS according to clinician-reported response and 2014 NIH-calculated response are shown in Figures 1 and 2, which also illustrate that the first FFS event occurred at a median of 16.3 months (95% CI, 13.5-19.4) after the 6-month assessment. Results were similar when the 6-month analysis was limited to incident cases (data not shown).

Predictors

In order to identify the changes in individual variables at 6 months that are most predictive of subsequent outcomes, we performed a multivariate analysis to compare the performance of all the collected measures against one another.

Table 2. Landmark analysis after 3 months and 6 months

Three-month landmark	Treatment change after 3 mo	FFS after 3 mo	NRM after 3 mo	OS after 3 mo
NIH calculated (2005) CR + PR vs SD + PD	$P = .37$	$P = .49$	$P = .55$	$P = .51$
NIH calculated (2014) CR + PR vs SD + PD	HR 0.50 (0.30-0.80) $P = .003$	HR 0.60 (0.41-0.89) $P = .01$	$P = .32$	$P = .59$
Clinician reported CR + PR vs SD + PD	HR 0.28 (0.18-0.47) $P < .001$	HR 0.34 (0.22-0.52) $P < .001$	$P = .84$	$P = .47$
Patient-reported improvement (I) vs stable (S) vs worsening (W)	Overall $P < .001$	Overall $P < .001$	$P = .62$	$P = .36$
I vs S	HR 0.41 (0.24-0.74), $P = .002$	HR 0.43 (0.27-0.70), $P < .001$		
S vs W	HR 0.24 (0.10-0.69), $P = .003$	HR 0.25 (0.11-0.63), $P = .001$		
Six-month landmark	**Treatment change after 6 mo**	**FFS after 6 mo**	**NRM after 6 mo**	**OS after 6 mo**
NIH calculated (2005) CR + PR vs SD + PD	HR 0.61 (0.40-0.90) $P = .01$	$P = .06$	$P = .24$	$P = .06$
NIH calculated (2014) CR + PR vs SD + PD	HR 0.56 (0.37-0.83) $P = .003$	HR 0.58 (0.42-0.80) $P = .001$	$P = .28$	$P = .20$
Clinician reported CR + PR vs SD + PD	HR 0.53 (0.36-0.81) $P = .004$	HR 0.61 (0.44-0.85) $P = .004$	$P = .06$	HR 0.55 (0.36-0.85) $P = .007$
Patient-reported improvement vs stable vs worsening	$P = .13$	$P = .08$	$P = .33$	$P = .44$

HR, hazard ratio.

All tested variables and the results of the organ-specific univariate and multivariate models are included in supplemental Tables 1-4. Overall, multivariate results are reported in Table 3. Improvements in the NIH 0 to 3 clinician-reported skin score and 0 to 10 patient-reported itching score at 6 months predicted longer subsequent FFS. Improvements in the Lee skin symptom score predicted longer subsequent OS and NRM, and the Functional Assessment of Cancer Therapy–Bone Marrow Transplant (FACT-BMT) trial outcome index score predicted longer subsequent OS.

We added aggregate measures, such as clinician-reported overall response, patient-reported response, the 2014 NIH response measure, and the NIH global severity score, to the models to determine whether they could replace the organ-specific or patient-reported measures. None of the aggregate measures nor the NIH global severity score showed a statistically significant association with any outcome (all $P > .09$) in these models, whereas the individual measures remained statistically significant ($P < .01$), except for the FACT-BMT total score in the survival model, which had a P value of .08. These results suggest that changes in the individual measures have independent prognostic significance and cannot be replaced by knowledge of the summary response measures.

Discussion

Our primary goal in this analysis was to identify measurements at 3 or 6 months predictive of subsequent long-term outcomes. Many of the intermediate or long-term outcomes, such as FFS, NRM, and OS require longer follow-up times and are not practical for clinical trial design. Therefore, it is critical to identify 6-month surrogate end points that are able to predict long-term outcomes. Our analysis identified several surrogate end points that were predictive of the more critical end points such as FFS, NRM, and OS. At 6 months, the calculated 2014 NIH response and clinician-reported response predict subsequent FFS, whereas clinician-reported response predicts subsequent OS. These results suggest that the 2014 NIH response measures do reflect changes in chronic GVHD disease activity because patients who do not achieve a CR or PR are more likely to have their treatment regimen changed. Although the 2014 NIH response criteria were not associated with subsequent OS or NRM, it is important to remember that the NIH response measures were never designed to predict survival. They were instead designed to capture relevant changes in chronic GVHD disease activity as a result of chronic GVHD-directed therapy.

We proceeded to analyze which specific measured variables were most predictive of FFS, OS, and NRM by multivariate analysis. Surprisingly, we found that FFS, OS, and NRM were primarily predicted by changes in patient-reported measures. Patient-reported symptoms and quality of life may be more sensitive to overall health than clinician-reported chronic GVHD measures. Another possibility is that clinicians may aggressively immunosuppress more symptomatic patients leading to shorter FFS and worse survival.

The ability of the clinician-reported and patient-reported responses to predict FFS is not surprising, because therapy is likely to be changed if the clinician or patient concludes that the response to current therapy is not adequate. Interestingly, treatment changes occurred at a median of 16 months after the landmark, suggesting that even if the patient has not achieved a CR or PR by 6 months, it may still be some time before an actual treatment change is made. The ability of clinician-reported response at 6 months to predict OS is encouraging, because some chronic GVHD trials currently use clinician-reported end points. An earlier analysis of our cohort did not show an association between clinician-reported response and survival; however, this was likely because of shorter follow-up time[27] because, in the current study, the survival benefit associated with clinician-reported response at 6 months did not appear until after 2 years. These findings demonstrate the prolonged disease course in patients with chronic GVHD and highlight the importance of identifying appropriate 6-month surrogate end points.

The 2014 NIH response criteria[29] performed better than the 2005 criteria in our analysis. The improved performance of the 2014 NIH response measures is likely because of modifications based on data from studies done between the 2 consensus conferences. First, skin body surface area measurements were removed, and the NIH 0 to 3 skin score is now used to measure response. Second, change from 0 to 1 in GI and liver score is no longer considered progression. Third, Schirmer's test has been replaced by change in 0 to 3 NIH eye score. Fourth, assessment of response in the lung is based on the forced expiratory volume in 1 second only and no longer includes diffusing

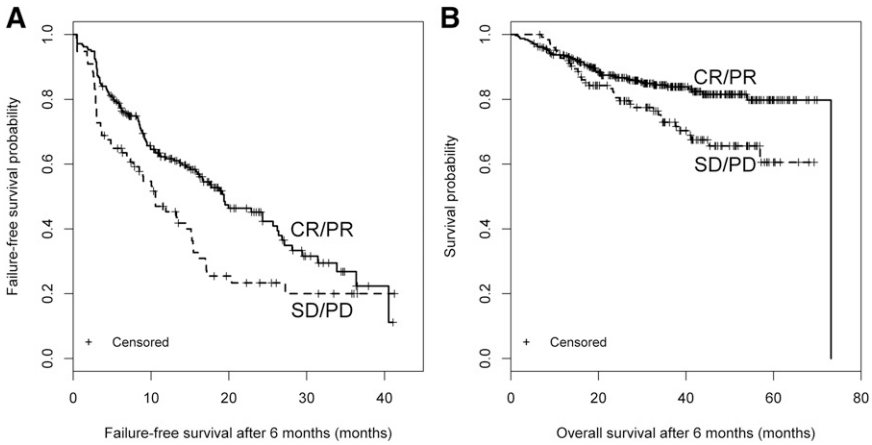

Figure 1. Clinician-reported response. Response at 6 months and subsequent (A) FFS and (B) OS.

capacity of the lungs for carbon monoxide. Finally, the NIH 0 to 3 joint score has been incorporated into the response criteria. As a result, overall response assignments derived from the 2014 algorithm show only poor to fair correlation with response assignments derived from the 2005 algorithm. Although the 2014 NIH response criteria also exclude an organ from the calculated response if the manifestation is entirely because of nonchronic GVHD causes, we could not ascertain in our data set whether signs or symptoms were related to chronic GVHD. Despite this limitation, the 2014 NIH response was predictive of subsequent FFS in our study population.

We also analyzed individual variables to understand which factors were most associated with long-term outcomes. We were able to identify a few individual measures whose change predicted subsequent FFS, OS, and NRM. Notably, of the 5 identified variables, 1 was a clinician-reported skin measure and 4 were patient-reported measures. A change in the NIH 0 to 3 skin score and patient reported 0 to 10 itching score predicted subsequent FFS. Skin manifestations are bothersome to patients, easily noted on exam, and likely drive treatment changes. We did not expect that patient-reported measures would predict OS and NRM and were surprised to find that these were the only identified predictors. Specifically, change in the Lee skin symptom score and the FACT-BMT score predicted OS. Worsening of the Lee skin symptom score predicted NRM, as has been previously reported in an earlier analysis of this patient cohort.[23] These associations may be because of the increased immunosuppression given to patients who have advanced and symptomatic chronic GVHD. Alternatively, worsening symptoms and quality of life may simply reflect declining health with its associated higher mortality rates.

Several findings were unexpected. First, baseline disease risk, which usually predicts relapse, did not predict FFS, OS, or NRM. FFS is largely determined by the addition of other systemic treatment and not by relapse or death. Also, because the median time to enrollment was 11.9 months, patients who relapsed early after transplant were not enrolled in our cohort. Second, factors that have historically predicted survival in chronic GVHD, such as platelet count,[41] hyperbilirubinemia,[22] overlap syndrome,[42] and lower GI involvement,[22] were not associated with survival in multivariate analysis. We previously reported that FFS was associated with enrollment NIH scores for skin and GI tract, range of motion, forced vital capacity, bronchiolitis obliterans syndrome, hepatic dysfunction, female donor into male recipient, prior grade 2-4 acute GVHD, and quality of life, but in the current analysis, only change in the NIH skin score was found to be associated with FFS. Although several of these variables were significant in organ-specific univariate analysis, they were not significant in the multivariate analysis. These apparent discrepancies may be partially explained by differences in the analytic approaches. In the current analysis, patients who experienced death or treatment change before the landmark were excluded from the analysis, potentially eliminating the statistical associations previously observed when the models started at enrollment. Another possibility is that what matters for prognosis is whether an organ is involved, not how it

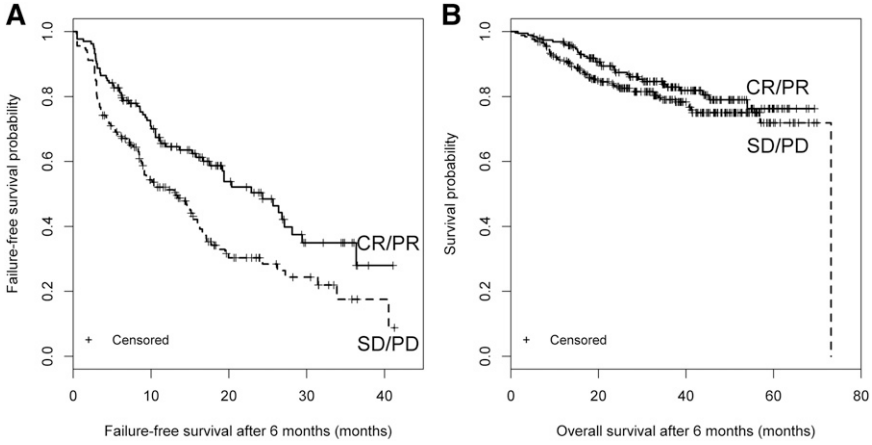

Figure 2. The 2014 NIH-calculated response. Response at 6 months and subsequent (A) FFS and (B) OS.

Table 3. Multivariate landmark analyses at 6 mo for subsequent FFS, OS, and NRM

Outcome	Parameter*	No. events/no. at risk after excluding missing	*P*	HR (95% CI)
FFS	Change in 2005 NIH 0 to 3 skin score	112/211	.001	1.53 (1.19-1.96)
	Change in patient 0 to 10 skin itching		.002	1.15 (1.06-1.24)
OS	Change in Lee skin symptom score	64/308	.005	1.02 (1.01-1.04)
	FACT-BMT total score		.04	0.98 (0.97-0.99)
NRM	Change in Lee skin symptom score	48/326	.001	1.03 (1.01-1.04)

*Models include the enrollment value for the significant change scores and adjustment for baseline characteristics.

responds over time. Finally, it is possible that not enough change occurred in the chronic GVHD activity of the organ by 6 months to demonstrate an association with FFS or OS.

Our study has several limitations. First, this analysis was conducted as a discovery exercise. Although the results are informative, they will need to be validated in a separate independent cohort prior to drawing definitive conclusions, and such a study is ongoing. The timing of assessments in the study was calendar driven and not influenced by the patient's clinical status or changes in therapy. Therefore, the measured and reported responses may not accurately recapitulate the circumstances of a clinical trial. Additionally, sequential assessments might have been done by different providers, causing inconsistency, especially because forms from the previous assessment were not routinely made available for reference. No direct instructions regarding subjective response assessments were provided to clinicians in assigning a clinical CR, PR, SD, or PD. Finally, some patient-reported outcome measures were missing. Despite these limitations, our data derive strength from the prospective collection of data with the use of standardized forms, the detailed chronic GVHD assessments that were performed, and the large number of patients from multiple centers.

In summary, our data show that the 2014 NIH response measures and clinician-reported response at 3 and 6 months correlate with subsequent FFS. Patient-reported response at 3 months predicted subsequent FFS. Clinician-reported response at 6 months predicted OS. Additionally, this study demonstrates the importance of specific patient-reported measures such as the Lee skin symptom score, for which changes predict OS and NRM, and the FACT BMT, for which changes predict OS. These results lend credence to the 2014 NIH response measures as reflective of disease activity, although not predictive of OS. They also emphasize the critical contribution of patient-reported measures to the assessment of patients with chronic GVHD. Based on these data, we recommend that for now, the 2014 NIH response measures, clinician-reported responses, and patient-reported outcomes be collected in therapeutic trials of chronic GVHD to ensure that relevant data are available once the best algorithm to capture a meaningful objective response is determined.[43]

Acknowledgments

This work was supported by a grant from the National Institutes of Health, National Cancer Institute (CA118953). The Chronic GVHD Consortium (grant U54 CA163438) is a part of the National Institutes of Health Rare Disease Clinical Research Network, supported through collaboration between the Office of Rare Diseases Research, the National Center for Advancing Translational Sciences, and the National Cancer Institute.

Authorship

Contribution: J. Palmer and S.J.L. designed research and drafted the manuscript; all authors contributed to analysis and interpretation of data and critical review of the manuscript; J. Palmer, P.J.M., Y.I., J. Pidala, S.Z.P., M.A., I.P., M.E.D.F., and S.J.L. contributed patients; and X.C. and B.S. performed statistical analysis.

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

Correspondence: Jeanne Palmer, Mayo Clinic Phoenix, 5777 East Mayo Blvd, Phoenix, AZ 85054; e-mail: palmer.jeanne@mayo.edu.

References

1. Socie G, Ritz J, Martin PJ. Current challenges in chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2010;16(1, suppl): S146-S151.
2. Socié G, Salooja N, Cohen A, et al; Late Effects Working Party of the European Study Group for Blood and Marrow Transplantation. Nonmalignant late effects after allogeneic stem cell transplantation. *Blood.* 2003;101(9):3373-3385.
3. Socié G, Stone JV, Wingard JR, et al; Late Effects Working Committee of the International Bone Marrow Transplant Registry. Long-term survival and late deaths after allogeneic bone marrow transplantation. *N Engl J Med.* 1999;341(1):14-21.
4. Lee SJ, Kim HT, Ho VT, et al. Quality of life associated with acute and chronic graft-versus-host disease. *Bone Marrow Transplant.* 2006; 38(4):305-310.
5. Lee SJ, Vogelsang G, Flowers MED. Chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2003;9(4):215-233.
6. Fraser CJ, Bhatia S, Ness K, et al. Impact of chronic graft-versus-host disease on the health status of hematopoietic cell transplantation survivors: a report from the Bone Marrow Transplant Survivor Study. *Blood.* 2006;108(8): 2867-2873.
7. Wingard JR, Majhail NS, Brazauskas R, et al. Long-term survival and late deaths after allogeneic hematopoietic cell transplantation. *J Clin Oncol.* 2011;29(16):2230-2239.
8. Socié G, Schmoor C, Bethge WA, et al; ATG-Fresenius Trial Group. Chronic graft-versus-host disease: long-term results from a randomized trial on graft-versus-host disease prophylaxis with or without anti-T-cell globulin ATG-Fresenius. *Blood.* 2011;117(23):6375-6382.
9. Arora M, Wagner JE, Davies SM, et al. Randomized clinical trial of thalidomide, cyclosporine, and prednisone versus cyclosporine and prednisone as initial therapy for chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2001;7(5):265-273.
10. Akpek G, Lee SM, Anders V, Vogelsang GB. A high-dose pulse steroid regimen for controlling active chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2001;7(9):495-502.
11. Chen GL, Arai S, Flowers MED, et al. A phase 1 study of imatinib for corticosteroid-dependent/ refractory chronic graft-versus-host disease: response does not correlate with anti-PDGFRA antibodies. *Blood.* 2011;118(15):4070-4078.
12. Jedlickova Z, Burlakova I, Bug G, Baurmann H, Schwerdtfeger R, Schleuning M. Therapy of sclerodermatous chronic graft-versus-host disease with mammalian target of rapamycin inhibitors. *Biol Blood Marrow Transplant.* 2011; 17(5):657-663.
13. Couriel DR, Hosing C, Saliba R, et al. Extracorporeal photochemotherapy for the treatment of steroid-resistant chronic GVHD. *Blood.* 2006;107(8):3074-3080.
14. 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(7):2667-2674.

15. Cutler C, Miklos D, Kim HT, et al. Rituximab for steroid-refractory chronic graft-versus-host disease. *Blood.* 2006;108(2):756-762.
16. Johnston LJ, Brown J, Shizuru JA, et al. Rapamycin (sirolimus) for treatment of chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2005;11(1):47-55.
17. Pavletic SZ, Martin P, Lee SJ, et al; Response Criteria Working Group. Measuring therapeutic response in chronic graft-versus-host disease: National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. Response Criteria Working Group report. *Biol Blood Marrow Transplant.* 2006;12(3):252-266.
18. Pavletic SZ, Lee SJ, Socie G, Vogelsang G. Chronic graft-versus-host disease: implications of the National Institutes of Health consensus development project on criteria for clinical trials. *Bone Marrow Transplant.* 2006;38(10):645-651.
19. 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(3): 337-344.
20. Inamoto Y, Pidala J, Chai X, et al; Chronic GVHD Consortium. Assessment of joint and fascia manifestations in chronic graft-versus-host disease. *Arthritis Rheumatol.* 2014;66(4): 1044-1052.
21. Pidala J, Chai X, Martin P, et al. Hand grip strength and 2-minute walk test in chronic graft-versus-host disease assessment: analysis from the Chronic GVHD Consortium. *Biol Blood Marrow Transplant.* 2013;19(6):967-972.
22. 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 [published correction appears in *Biol Blood Marrow Transplant.* 2014;20(2):290]. *Biol Blood Marrow Transplant.* 2013;19(5):784-791.
23. 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(13):2545-2552, quiz 2774.
24. Inamoto Y, Chai X, Kurland BF, et al; Chronic GVHD Consortium. Validation of measurement scales in ocular graft-versus-host disease. *Ophthalmology.* 2012;119(3):487-493.
25. Olivieri A, Cimminiello M, Corradini P, et al. Long-term outcome and prospective validation of NIH response criteria in 39 patients receiving imatinib for steroid-refractory chronic GVHD. *Blood.* 2013; 122(25):4111-4118.
26. 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(11):1649-1655.
27. Inamoto Y, Martin PJ, Chai X, et al; Chronic GVHD Consortium. Clinical benefit of response in chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2012;18(10):1517-1524.
28. Martin PJ, Storer BE, Carpenter PA, et al. Comparison of short-term response and long-term outcomes after initial systemic treatment of chronic graft-versus-host disease. *Biol Blood Marrow Transplant.* 2011;17(1):124-132.
29. 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(6):984-999.
30. Palmer J, Chai X, Martin PJ, et al. Failure-free survival in a prospective cohort of patients with chronic graft-versus-host disease. *Haematologica.* 2015;100(5):690-695.
31. Chronic GVHD Consortium. Rationale and design of the chronic GVHD cohort study: improving outcomes assessment in chronic GVHD. *Biol Blood Marrow Transplant.* 2011;17(8):1114-1120.
32. Filipovich AH, Weisdorf D, Pavletic S, et al. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: I. Diagnosis and Staging Working Group report. *Biol Blood Marrow Transplant.* 2005;11(12):945-956.
33. Inamoto Y, Storer BE, Lee SJ, et al. Failure-free survival after second-line systemic treatment of chronic graft-versus-host disease. *Blood.* 2013; 121(12):2340-2346.
34. Martin PJ, Storer BE, Rowley SD, et al. Evaluation of mycophenolate mofetil for initial treatment of chronic graft-versus-host disease. *Blood.* 2009; 113(21):5074-5082.
35. Pavletic SZ, Martin P, Lee SJ, et al. Measuring therapeutic response in chronic graft-versus-host disease: national institutes of health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: IV. Response Criteria Working Group Report. *Biol Blood Marrow Transplant.* 2006;12(3):252-266.
36. Martin PJ, Weisdorf D, Przepiorka D, et al; Design of Clinical Trials Working Group. National Institutes of Health consensus development project on criteria for clinical trials in chronic graft-versus-host disease: VI. Design of Clinical Trials Working Group report. *Biol Blood Marrow Transplant.* 2006;12(5):491-505.
37. Greinix HT, Pohlreich D, Maalouf J, et al. A single-center pilot validation study of a new chronic GVHD skin scoring system. *Biol Blood Marrow Transplant.* 2007;13(6):715-723.
38. Jacobsohn DA, Chen AR, Zahurak M, et al. Phase II study of pentostatin in patients with corticosteroid-refractory chronic graft-versus-host disease. *J Clin Oncol.* 2007;25(27): 4255-4261.
39. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Stat Med.* 1999;18(6):695-706.
40. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas.* 1973; 33(3):613-619.
41. Arora M, Klein JP, Weisdorf DJ, et al. Chronic GVHD risk score: a Center for International Blood and Marrow Transplant Research analysis. *Blood.* 2011;117(24):6714-6720.
42. Pidala J, Vogelsang G, Martin P, et al. Overlap subtype of chronic graft vs. host disease is associated with adverse prognosis, functional impairment, and inferior patient reported outcomes: a chronic graft vs. host disease Consortium study. *Haematologica.* 2011;96(11): 1678-1684.
43. 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(8):1343-1359.