Regular Article

blood

TRANSPLANTATION

An endpoint associated with clinical benefit after initial treatment of chronic graft-versus-host disease

Paul J. Martin,[1,2] Barry E. Storer,[1] Yoshihiro Inamoto,[1] Mary E. D. Flowers,[1,2] Paul A. Carpenter,[1,3] Joseph Pidala,[4] Jeanne Palmer,[5] Mukta Arora,[6] Madan Jagasia,[7] Sally Arai,[8] Corey S. Cutler,[9] and Stephanie J. Lee[1,2]

[1]Division of Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, WA; [2]Department of Medicine and [3]Department of Pediatrics, University of Washington, Seattle, WA; [4]Blood and Marrow Transplantation, Moffitt Cancer Center, Tampa, FL; [5]Division of Hematology/Oncology, Mayo Clinic, Phoenix, AZ; [6]Department of Medicine, University of Minnesota, Minneapolis, MN; [7]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN; [8]Department of Medicine, Stanford University Medical Center, Stanford, CA; and [9]Department of Medicine, Harvard Medical School, Dana-Farber Cancer Institute, Boston, MA

Key Points

- Complete or partial response at 1 year without secondary systemic treatment provides clinical benefit in patients with chronic GVHD.
- Success defined by this endpoint is currently observed in fewer than 20% of patients after initial systemic treatment of chronic GVHD.

No gold standard has been established as a primary endpoint in trials of initial treatment of chronic graft-versus-host disease (GVHD), and evidence showing the association of any proposed primary endpoint with clinical benefit has not been conclusively demonstrated. To address this gap, we analyzed outcomes in a cohort of 328 patients enrolled in a prospective, multicenter, observational study within 3 months after diagnosis of chronic GVHD. Complete and partial response, stable disease, and progressive disease were defined according to the 2014 National Institutes of Health Consensus Development Conference on Criteria for Clinical Trials in Chronic Graft-Versus-Host Disease. Success was defined as complete or partial response with no secondary systemic treatment or recurrent malignancy at 1 year after enrollment. Success was observed in fewer than 20% of the patients. The burden of disease manifestations at 1 year was lower for patients in this category than for those with stable or progressive disease. Systemic treatment ended earlier, and subsequent mortality was lower among patients with complete or partial response than among those with stable or progressive disease and those who had received secondary systemic treatment. We conclude that survival with a complete or partial response and no previous secondary systemic treatment or recurrent malignancy at 1 year after initial systemic therapy is associated with clinical benefit, a critical characteristic for consideration as a primary endpoint in a pivotal clinical trial. This prospective observational study was registered at www.clinicaltrials.gov as #NCT00637689. (*Blood*. 2017;130(3):360-367)

Introduction

The long-term success of allogeneic hematopoietic cell transplantation (HCT) is limited by the morbidity and mortality caused by chronic graft-versus-host disease (GVHD), a complication that develops in 30% to 40% of recipients.[1,2] With current approaches, the median duration of systemic treatment is between 1.0 and 3.5 years for the 50% of patients whose chronic GVHD is eventually controlled.[3,4] Approximately 10% of patients require systemic treatment beyond 7 years, and the remaining 40% die or develop recurrent malignancy during systemic treatment within 7 years after diagnosis. Many patients have irreversible impairment caused by skin and connective tissue sclerosis and damage to lacrimal and salivary glands, which greatly compromise quality of life.[5-7] Development of more effective treatments for chronic GVHD represents an urgent unmet clinical need.

The identification of robust clinical endpoints for this complex, multiorgan syndrome poses a major challenge.[8] No gold standard has been established as a primary endpoint in trials of treatment of chronic GVHD, and evidence showing the association of any proposed primary endpoint with clinical benefit has not been conclusively demonstrated. In previous retrospective studies, we suggested that the absence of secondary systemic treatment, nonrelapse mortality, and recurrent or progressive malignancy could be incorporated into a composite, failure-free survival (FFS) endpoint to evaluate results of treatment.[9-11]

The premise underpinning this endpoint was that chronic GVHD was adequately controlled when no secondary systemic treatment had been given before the end assessment and that GVHD was not adequately controlled when secondary systemic treatment had been given before the end assessment. Because this endpoint does not provide any direct information about changes in GVHD-related symptoms, activity, damage, functional impairment, or disability, however, we recommended that measures of these outcomes should be included in any study that relied on FFS as the primary endpoint.[10]

To address this problem, we incorporated response definitions from the 2014 National Institutes of Health (NIH) Consensus Development Conference on Criteria for Clinical Trials in Chronic Graft-Versus-Host Disease as an additional component in the FFS composite endpoint.[12] Success was defined as FFS with complete or partial response at 1 year after enrollment. Results were analyzed in a cohort of 328 patients

Submitted 24 March 2017; accepted 8 May 2017. Prepublished online as *Blood* First Edition paper, 11 May 2017; DOI 10.1182/blood-2017-03-775767.

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.

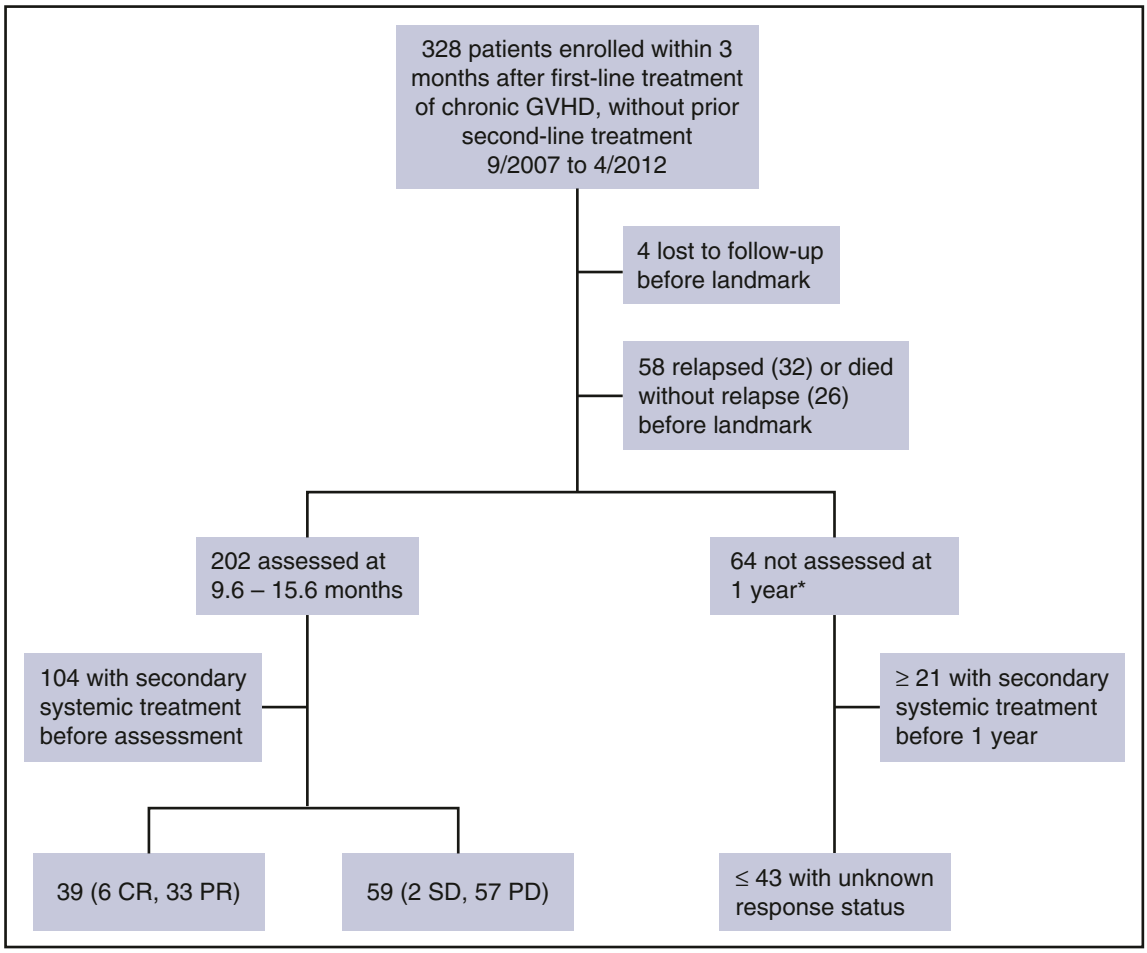

Figure 1. Flow of outcomes among patients enrolled in the study within 3 months after beginning systemic treatment of chronic GVHD. Patients who received second-line systemic treatment before enrollment in the study were excluded from this analysis. Fifteen of the 32 patients with relapse died within the first year after enrollment. CR/PR indicates complete or partial response at the time of assessment; SD/PD indicates stable or progressive disease at the time of assessment. *This category includes patients who were alive without recurrent or progressive malignancy at 1 year but did not have the intended assessment.

enrolled in a prospective, multicenter, observational study sponsored by the Chronic GVHD Consortium.[11,13] Contrary to expectations, the results did not support the premise that chronic GVHD was adequately controlled when no secondary systemic treatment had been given before the end assessment. Results of the current study show that FFS with complete or partial response at 1 year after initial treatment is associated with clinical benefit.

Methods

Study cohort

Details of enrollment and follow-up in this prospective multicenter observational study have been reported previously[13] and are summarized in the supplemental Methods (available on the *Blood* Web site). The institutional review board at each site approved the protocol, and all subjects provided written informed consent.

Endpoint definitions

For the assessment of FFS, failure was defined as malignancy relapse, death, or addition of a secondary immunosuppressive medication (eg, sirolimus, rituximab) or treatment (eg, extracorporeal photopheresis) intended for systemic treatment of chronic GVHD. Determination of failure was made by 2 separate reviewers (J. Palmer and S.J.L.) independently, and discrepancies were resolved by discussion.[11]

Overall response was determined according to the 2014 NIH consensus criteria algorithms, using changes in skin, mouth, eye, lungs, joints, gastrointestinal (GI), and liver measures to assign outcomes as complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD).[12] Although the 2014 response criteria were not available when the study started, the relevant measures were collected in the study and available to calculate responses with the 2014 algorithm. Providers and patients also rated the global severity of chronic GVHD according to 4-point (0-3) and 11-point (0-10) scales. Quality of life was measured according to the Lee Symptom Scale[14] and Functional Assessment of Cancer Therapy—Bone Marrow Transplant (FACT-BMT).[15] Except for the FACT-BMT, higher numerical values in all scales indicated greater severity of disease manifestations. Clinically significant improvement was defined for each component scale according to the 2014 NIH Consensus Conference.

Statistical analysis

A two-sample Student *t* test with Satterthwaite correction was used to compare the mean change in measure according to complete or partial response (CR/PR) vs stable disease or progressive disease (SD/PD). Fisher's exact test was used to compare the proportions of improved patients according to CR/PR vs SD/PD. Cumulative incidence estimates were used to analyze the end of systemic treatment, and Kaplan-Meier estimates were used to analyze survival. The end of systemic treatment was defined as the permanent discontinuation of all systemic agents, including prednisone at low physiologic replacement doses. Follow-up for at least 3 months was required to confirm that systemic treatment had ended. Discontinuation of systemic treatment was not considered the end of systemic

treatment if systemic treatment was subsequently resumed. Cox proportional hazard models were used to compare end of systemic treatment and survival between groups defined according to FFS with CR/PR or SD/PD. Patients who had received secondary systemic treatment before the end assessment were analyzed as a separate group for additional comparison. Recurrent malignancy was treated as a competing risk in analyzing end of systemic treatment, because this event does not indicate resolution of chronic GVHD.

Results

Characteristics of the study cohort

Between August 2007 and January 2013, 328 patients were enrolled in this study. Figure 1 summarizes the flow of outcomes in the study. Patient characteristics are summarized in Table 1. The median age of patients was 52 years (range, 19-79 years). Of the 328 patients, 168 (51%) were prepared with high-dose conditioning regimens, 273 (84%) had HLA-matched related or unrelated donors, and 286 (87%) received mobilized blood cell grafts.

GVHD characteristics at enrollment are summarized in Table 2. The median time from HCT to enrollment was 8.3 months (range, 2.9-60.7 months), and the median number of days from chronic GVHD diagnosis to enrollment was 10 (range, 0-95 days). The sites most frequently involved with chronic GVHD were the mouth, skin, and liver. Pulmonary abnormalities defined according to the 2005 organ-scoring algorithm[16] were present in 158 patients (48%) and were not necessarily related to chronic GVHD. By 2005 NIH criteria, 42% of the patients had severe chronic GVHD, 49% had moderate severity, and 9% had mild severity. Chronic GVHD severity at enrollment was slightly lower among patients who enrolled in the study more than 30 days after diagnosis in comparison with those who enrolled within 30 days after diagnosis (supplemental Table 1).

Initial treatment of chronic GVHD included prednisone with or without a calcineurin inhibitor in 189 patients (58%), prednisone with or without a calcineurin inhibitor and other agents in 95 patients (29%), and other agents without prednisone in 44 patients (13%).

Table 1. Patient characteristics (n = 328)

Characteristic	N (%)
Patient sex	
Male	180 (55)
Female	148 (45)
Donor–patient sex combination	
Female to male	90 (28)
Other	237 (72)
Diagnosis	
Myeloid malignancy	174 (53)
Lymphoid malignancy	124 (38)
Other/nonmalignant	30 (9)
Disease risk at transplant	
Low	112 (34)
Intermediate	146 (45)
High	69 (21)
Conditioning regimen	
Myeloablative without TBI	79 (24)
Myeloablative with TBI	89 (27)
Nonmyeloablative	157 (48)
Graft source	
Bone marrow	22 (7)
PBSC	286 (87)
Cord blood	20 (6)
Donor and HLA match	
HLA-matched related	137 (42)
HLA-matched unrelated	136 (42)
HLA-mismatched related	10 (3)
HLA-mismatched unrelated	43 (13)
Transplant center	
Fred Hutchinson	154 (47)
University of Minnesota	43 (13)
Dana-Farber	28 (9)
Stanford University	42 (13)
Vanderbilt University	25 (8)
Medical College of Wisconsin	14 (4)
Moffitt	20 (6)
Memorial Sloan Kettering	2 (1)

For some categories, totals are less than 328 because of missing data.
PBSC, peripheral blood stem cells; TBI, total body irradiation.

Outcomes during first-line treatment

Figure 1 shows a flow diagram of outcomes during first-line treatment. Four patients (1%) were lost to follow-up, and 58 (18%) died or had recurrent malignancy during the first year and did not have the intended assessment at approximately 1 year after enrollment. Sixty-four patients (20%) survived without recurrent or progressive malignancy to 1 year but did not have the intended assessment of first-line treatment at 1 year. Twenty-one of these patients are known to have received secondary systemic treatment during the first year, and data were missing for the other 43. At 9 to 16 months after enrollment, 202 patients (62%) were assessed. Of these, 104 patients had received secondary systemic treatment before the assessment, and 98 had not. According to 2014 NIH overall response criteria (supplemental Table 2), 6 of the 98 had a CR, 33 had a PR, 2 had SD, and 57 had PD. The distribution of outcomes did not differ among patients who enrolled in the study within 30 days after diagnosis of chronic GVHD in comparison with those who enrolled more than 30 days after diagnosis (supplemental Table 1), suggesting that delayed enrollment in the study did not bias the results.

Of the 57 patients with PD by NIH criteria, at least 26 (46%) had worsening of the joint score or decreased range of motion, 17 (30%) had greater than 10% loss of forced expiratory volume (FEV1), 11 (19%) had worsening of oral manifestations, 9 (16%) had worsening of ocular manifestations, and 7 (12%) had worsening of cutaneous manifestations (supplemental Table 3). Nine of the 17 patients with PD because of worsened FEV1 had no other reason for PD, and only 1 of the 9 had bronchiolitis obliterans or cryptogenic organizing pneumonia documented before or at the 1-year assessment. Outcomes would have been categorized as PR in the 8 remaining patients if FEV1 had not been considered in the response algorithm.

According to the provider 0 to 3 score, 17 of the 57 patients (30%) with PD by NIH criteria were judged to have clinically significant improvement (CR or PR), 32 (56%) were judged to have SD, and only 8 (14%) were judged to have worsening of manifestations (PD). Results were similar for the provider 0 to 10 score. Similarly, 8 of the 37 patients (22%) with patient 0 to 3 scores available were judged to have clinically significant improvement, 23 (62%) were judged to have SD, and only 6 (16%) were judged to have worsening of manifestations. Results were similar for the patient 0 to 10 score.

Change in GVHD manifestations according to CR or PR vs SD or PD

Change scores for individual chronic GVHD manifestations were compared for patients according to the presence or absence of CR/PR. In this analysis, the absence of change in a given manifestation could indicate either that it was absent at both the baseline and the 1-year assessment or that it was present without improvement or worsening at the

Table 2. Chronic GVHD characteristics at enrollment (n = 328)

Characteristic	N (%)
Sites involved	
Skin	206 (63)
Eyes	144 (44)
Mouth	209 (64)
Liver (1 unknown)	188 (57)
Gastrointestinal	120 (37)
Lung symptoms or abnormal lung function	158 (48)
Bronchiolitis obliterans or cryptogenic organizing pneumonia	16 (5)
Joints	80 (24)
Genital tract (26 unknown)	30 (10)
Number of sites involved	
1 or 2	86 (26)
3	94 (29)
≥4	148 (45)
NIH global severity	
Mild	30 (9)
Moderate	160 (49)
Severe	138 (42)
Subcategory of GVHD	
Classic	37 (11)
Overlap	291 (89)
Karnofsky score	
80-100	222 (68)
<80	106 (32)
Platelet count*	
<100 000 per μL	60 (18)
≥100 000 per μL	267 (82)
Serum total bilirubin	
>2 mg/dL	22 (7)
≤2 mg/dL	306 (93)
Progressive onset*	
No	238 (73)
Yes	89 (27)
Prior grade II–IV acute GVHD	
No	146 (45)
Yes	182 (55)
Onset during treatment with prednisone	
None	227 (69)
<0.5 mg/kg/d	65 (20)
≥0.5 and <1.0 mg/kg/d	22 (7)
≥1.0 mg/kg/d	13 (4)
Dose unknown	1 (< 1)
Initial treatment of chronic GVHD	
Prednisone ± calcineurin inhibitor alone	189 (58)
Prednisone ± calcineurin inhibitor + other agents	95 (29)
Mycophenolate mofetil included	32 (10)
Sirolimus included	44 (13)
Other agents included	28 (9)
No prednisone	44 (13)
Calcineurin inhibitor included	34 (10)
Mycophenolate mofetil included	9 (3)
Sirolimus included	5 (2)
Other agents included	7 (2)
Prednisone dose for initial treatment	
None	44 (13)
<0.5 mg/kg/d	68 (21)
≥0.5 and <1.0 mg/kg/d	107 (33)
1.0 mg/kg/d	55 (17)
>1.0 mg/kg/d	36 (11)
Dose unknown	18 (5)
Number of agents for initial treatment	
1	85 (26)
2	187 (57)
≥3	56 (17)

*One is unknown.

1-year assessment. Eye scores, joint and fascia scores, range of motion scores, lung symptom scores, and lower GI scores showed statistically significant changes in opposite directions between patients with CR/PR and those with SD/PD (Table 3), as was expected, given that changes in these scores are used to define response categories. Liver function tests, the FEV1, patient-assessed grading of itching, oral sensitivity, the chief eye complaint, and the Lee Symptom Scale showed no statistically significant differences in changes between patients with CR/PR and those with SD/PD. On all other scales, patients with CR/PR showed statistically significant greater improvement than did those with SD/PD.

Clinically significant categorical improvement in GVHD manifestations according to overall CR or PR vs SD or PD

In further analysis, individual chronic GVHD manifestations were scored categorically as a clinically significant improvement or not, according to 2014 NIH Response Criteria (supplemental Table 2). Patients who were unaffected by a given manifestation at both the baseline and the 1-year assessment were not included in this analysis. In comparison with patients who had SD or progression, those with CR/PR had statistically significant categorical improvements in the skin, mouth, joints, lungs, and lower GI tract, as assessed by providers (Table 4). Liver function tests showed no differences in propensity toward categorical improvement vs no improvement in patients with CR/PR vs SD/PD. Patient assessments of itching, oral sensitivity, and the chief eye complaint also did not show any statistically significant differences in propensity toward categorical improvement vs no improvement in patients with CR/PR vs SD/PD.

In comparison with patients who had SD or progression, those with CR/PR had statistically significant categorical improvements in provider global ratings, the patient 0 to 10 global rating, and the FACT-BMT score (Table 4). The Lee Symptom Scale did not show statistically significant differences in propensity to categorical improvement vs no improvement in patients with CR/PR vs SD/PD.

Outcomes associated with end of systemic treatment and survival after the 1-year landmark

At the time of analysis, survivors among the 202 patients who were assessed at 9 to 16 months after enrollment (Figure 1) had a subsequent median follow-up of 53 (range, 5-88) months. Systemic treatment ended earlier in the CR/PR group (n = 39) than in the PR/SD group (n = 59) and secondary treatment group (n = 104) (hazard ratio [HR], 2.61; 95% confidence interval [CI], 1.41-4.83; $P = .002$; and HR, 2.83; 95% CI, 1.64-4.87; $P = .0002$, respectively; global $P = .0009$). Rates of discontinued systemic treatment were similar in the SD/PD and secondary treatment groups (Figure 2A). Systemic treatment ended earlier in the CR/PR group than in the combined group with SD/PD or secondary treatment (HR, 2.75; 95% CI, 1.67-4.51; $P < .0001$).

The mortality rate was lower in the CR/PR group than in the SD/PD and secondary treatment groups (HR, 0.19; 95% CI, 0.04-0.85; $P = .03$; and HR, 0.18; 95% CI, 0.04-0.77; $P = .02$, respectively; global $P = .01$). Mortality rates after the 1-year landmark were similar in the SD/PD and secondary treatment groups (Figure 2B). The mortality rate was lower in the CR/PR group than in the combined group with SD/PD or secondary systemic treatment (HR, 0.19; 95% CI, 0.05-0.77; $P = .02$). The extremely low hazard ratios and wide confidence intervals reflect the low number of deaths in the CR/PR group (n = 2). Provider overall assessment scores did not show a statistically significant association with survival after the 1-year landmark. Patients with CR/PR defined according to either the provider 3-point or 10-point scale did not have better survival than did those without CR/PR (HR = 1.13, $P = .70$, and HR = 0.89, $P = .70$, respectively).

Table 3. Change measures between enrollment and 1 year, according to CR or PR vs SD or PD among patients with failure-free survival

	Overall CR or PR (n = 39)		Overall SD or PD (n = 59)		
Assessment	**N**	**Mean (SD)**	**N**	**Mean (SD)**	***P****
Provider grading of specific measures					
NIH Skin Score (0-3)	39	−1.05 (1.12)	59	−0.29 (1.07)	.001
NIH Eye Score (0-3)	39	−0.18 (0.51)	59	0.10 (0.69)	.02
Modified Oral Mucosa Rating Scale (0-12)	39	−2.21 (2.62)	59	−0.51 (2.78)	.003
Total of serum bilirubin	39	−0.59 (2.61)	57	−0.14 (0.42)	.29
Alanine aminotransferase	39	−78.8 (144.2)	57	−75.0 (196.4)	.91
Alkaline phosphatase	39	−43.3 (107.3)	57	−24.2 (158.1)	.48
Percentage of predicted FEV1	17	1.76 (8.79)	20	−4.10 (16.1)	.17
NIH Joint and Fascia Score (0-3)	39	−0.18 (0.45)	59	0.12 (0.77)	.02
Photographic range of motion (4-25)	22	0.32 (1.49)	39	−0.49 (1.17)	.04
Provider grading of specific symptoms					
NIH Lung Symptom Score (0-3)	39	−0.21 (0.52)	59	0.19 (0.75)	.003
Upper GI Score (0-3)	39	−0.38 (0.88)	59	−0.05 (0.63)	.04
Lower GI Score (0-3)	39	−0.28 (0.72)	59	0.07 (0.55)	.01
Esophagus Score (0-3)	39	−0.26 (0.72)	59	0.00 (0.32)	.04
Patient grading of specific symptoms					
Skin itching (0-10)	25	−0.80 (2.20)	38	−1.00 (3.05)	.76
Oral sensitivity (0-10)	25	−1.60 (2.16)	40	−0.63 (3.51)	.17
Chief eye complaint (0-10)	25	−0.24 (2.45)	39	0.21 (3.67)	.56
Global rating scales					
Provider 0-3	39	−0.87 (0.80)	59	−0.22 (0.74)	.0001
Provider 0-10	39	−3.03 (1.99)	59	−1.02 (1.91)	<.0001
Patient 0-3	26	−0.62 (0.75)	37	−0.05 (0.62)	.003
Patient 0-10	23	−2.83 (2.19)	37	−0.43 (2.22)	.0002
Lee Symptom Scale (0-100)†	27	−5.51 (7.41)	39	−1.59 (12.3)	.11
FACT-BMT	25	11.9 (14.7)	38	1.20 (14.4)	.007

*Two-sample Student *t* test with Satterthwaite correction comparing mean change in measure according to CR or PR vs SD or PD by NIH criteria. Patients unaffected by specific manifestations are included. Some patient-reported data are missing.

†None of the subscales showed statistically significant differences in mean change between patients with overall CR or PR in comparison with those who had SD or PD.

Estimated proportions of enrolled patients with FFS and overall CR or PR at 1 year

At least 125 patients received secondary treatment before the 1-year landmark, and 98 patients had FFS at the landmark (Figure 1), representing 44% of the 223 patients across these 2 groups. An additional 43 patients survived to at least 1 year without recurrent malignancy, but records were not available to determine whether secondary systemic treatment had been given before the 1-year landmark. If a similar 44% FFS rate were assumed for the 43 patients with missing data, we would estimate that 19 additional patients had FFS. Under this assumption, 117 patients representing 36% of the 324 patients who were not lost to follow-up had FFS at approximately 1 year after enrollment.

Of the 98 patients with FFS who were assessed at approximately 1 year, 39 (40%) had a CR or PR, representing 12% of the 324 patients who were not lost to follow-up. If a similar 40% CR/PR rate is assumed for the estimated 19 additional patients with FFS at 1 year, we would estimate that 8 additional patients had FFS with CR or PR. Under this assumption, 47 patients representing 15% of the 324 patients who were not lost to follow-up had FFS and a CR or PR at approximately 1 year after enrollment.

Outcomes at 6 months

A remaining question is whether the efficacy of initial systemic treatment of chronic GVHD could be assessed at an earlier time point after study enrollment. At 6 months after enrollment in the current study, 98 patients had already received secondary systemic treatment, and 71 patients had FFS with CR/PR. At least 27 patients received secondary systemic treatment during the interval from 6 months to 1 year after enrollment. As is indicated above, we estimated that between 39 and 47 patients had CR/PR without secondary treatment at 1 year. Eight of these patients did not have CR/PR at 6 months, indicating that 44% to 55% of the 71 patients who had CR/PR at 6 months also had CR/PR at 1 year. Differences in the rates of ending systemic treatment and mortality between groups defined according to CR/PR, SD/PD, and prior secondary systemic treatment at 6 months (supplemental Figure 2) are much less striking than are the differences between groups defined according to outcomes at 1 year (Figure 1), as might be expected with the 6 additional months of observation.

Discussion

At least 4 major findings emerged from this prospective observational study to evaluate outcomes after initial systemic treatment of chronic GVHD according to the 2014 NIH response criteria. First, the proportion of patients with FFS and CR/PR at 1 year was less than 20%, much lower than might have been expected. Second, PD as defined by the 2014 NIH response criteria showed poor correspondence with global assessments by providers or patients. Third, FFS with overall SD at 1 year is a distinctly unusual outcome. Accordingly, the assessment of FFS with CR or PR yields unambiguous, nearly binary outcomes of improvement or worsening, leaving very few remaining indeterminate SD outcomes at 1 year. Finally, FFS with a CR or PR at 1 year is associated with clinical benefit, a critical characteristic for consideration as a primary endpoint in a pivotal clinical trial. Patients in this category had a lower burden of disease manifestation at 1 year, a shorter time to the end of systemic treatment, and better survival than did those without CR or PR.

Several explanations could account for the low proportion of patients surviving at FFS and CR/PR at 1 year. First, the 36% estimated proportion of patients with FFS at 1 year in the current study was lower than the 54% proportion in a previous retrospective study.[10] In the

Table 4. Proportions of patients with clinically significant improvement, according to CR or PR vs SD or PD with failure-free survival at 1 year after enrollment

	Overall CR or PR (n = 39)		Overall SD or PD (n = 59)		
Assessment	**N**	**Number improved (%)**	**N**	**Number improved (%)**	***P****
Provider grading of specific measures					
NIH Skin Score (0-3)	27	24 (89)	37	20 (54)	.003
NIH Eye Score (0-3)	17	7 (41)	41	10 (24)	.22
Modified Oral Mucosa Rating Scale (0-12)	29	20 (69)	48	16 (33)	.004
Total serum bilirubin	5	5 (100)	5	5 (100)	NA
Alanine aminotransferase	22	15 (68)	27	25 (85)	.19
Alkaline phosphatase	18	11 (61)	27	17 (63)	>.99
Percentage of predicted FEV1	0	NA	4	1 (25)	NA
NIH Joint and Fascia Score (0-3)	9	6 (67)	26	5 (19)	.01
Photographic range of motion (4-25)	4	1 (25)	24	4 (17)	>.99
Provider grading of specific symptoms					
NIH Lung Symptom Score (0-3)	6	6 (100)	26	8 (31)	.003
Upper GI Score (0-3)	8	8 (100)	13	8 (62)	.11
Lower GI Score (0-3)	9	8 (89)	8	2 (25)	.02
Esophagus Score (0-3)	6	5 (83)	8	3 (38)	.14
Patient grading of specific symptoms					
Skin itching (0-10)	21	8 (38)	30	13 (43)	.78
Oral sensitivity (0-10)	18	10 (56)	30	13 (43)	.55
Chief eye complaint (0-10)	15	5 (33)	37	13 (35)	>.99
Global rating scales					
Provider 0-3	39	25 (64)	59	18 (31)	.002
Provider 0-10	39	29 (74)	59	23 (39)	.0009
Patient 0-3	26	12 (46)	37	8 (22)	.06
Patient 0-10	22	17 (77)	36	13 (36)	.003
Lee Symptom Scale (0-100)	27	11 (41)	39	9 (23)	.17
FACT-BMT	25	14 (56)	38	9 (24)	.02

*Fisher's exact test comparing the proportions of improved patients according to CR or PR vs SD or PD by NIH criteria. Patients unaffected by specific manifestations are not included. Some patient-reported data are missing. See supplemental Table 2 for definitions of clinically significant improvement.

NA, not applicable.

previous study, multivariate analysis identified 4 risk factors associated with treatment failure: time interval less than 1 year from transplantation to initial treatment; patient age that was ≥60 years; severe involvement of the gastrointestinal tract, liver, or lungs; and Karnofsky score of <80% at initial treatment. The proportions of patients with these risk factors in the current study were similar to those in the previous study. On the other hand, the proportion of patients with mild global severity was lower in the present study (9%) than in the previous study (30%).

In the previous study,[10] the causes of failure at 1 year were secondary systemic treatment (30%), nonrelapse mortality (9%), and recurrent malignancy (7%). In the current study, nonrelapse mortality and recurrent malignancy, respectively, accounted for failure in 8% and 10% of the 324 patients who were not lost to follow-up, similar to the results of the previous study. Secondary systemic treatment accounted for failure in 39% to 58% of all patients, depending on whether or not only 21 patients or all 64 patients who were not assessed at 1 year had secondary systemic treatment (Figure 1). It is possible that the previous retrospective study did not capture all secondary systemic treatments that patients received during the first year.

Several explanations could account for the poor correspondence of PD with global assessments by providers and patients. First, the 2014 response criteria[12] might benefit from further refinement. Many patients with overall PD could have had improvement in some manifestations with worsening in others, such that the improved manifestations were more apparent or had greater clinical effect than did those that had worsened in the overall clinical assessment. Second, recall bias could unconsciously highlight comparisons with the most recent evaluation, whereas the overall response evaluation should be based on comparison with the baseline assessment. Third, clinically insignificant changes in the joints, lung, oral mucosa, eyes, or skin can cross boundaries from scores of 1 to 2 or from 2 to 3, thereby meeting a technical definition of PD without sufficient justification for secondary systemic treatment. Isolated worsening of ocular or oral manifestations would not usually prompt secondary systemic treatment, which might explain why some patients with PD did not receive secondary treatment before the assessment at 1 year. Fourth, exacerbation of certain manifestations could have causes other than chronic GVHD.[17] In particular, dyspnea and pulmonary function abnormalities were reported frequently but were not necessarily caused by chronic GVHD. Finally, inattention to detail, response shift, and desire to claim improvement by both providers and patients could contribute to the poor correspondence of PD with global assessments by providers and patients.

Most patients with FFS at 1 year had CR/PR or PD, though very few had SD. Patients with SD/PD ended systemic treatment later and had higher mortality than did those with CR/PR, similar to patients with secondary treatment before the assessment at 1 year. Accordingly, it would appear that most patients classified as PD according to the 2014 response criteria actually had PD. If so, the results would emphasize the importance of detailed, real-time collection and cleaning of baseline and endpoint assessment data and the reliance on well-defined, objective algorithms to establish the diagnosis in each organ and to assess response in clinical trials of treatment of chronic GVHD.

Conversely, results of this study suggest that CR or PR defined by the 2014 NIH response criteria is associated with clinical benefit. Improvements in many measures were greater in patients with overall CR/PR than in those with overall SD/PD, although improvement rates in liver function tests and scores for skin itching and the chief eye complaint were comparable in the two groups. Scores for the Lee Symptom Scale improved in both the CR/PR and SD/PD groups, although more so in the CR/PR group than in the SD/PD group, as might be expected. In

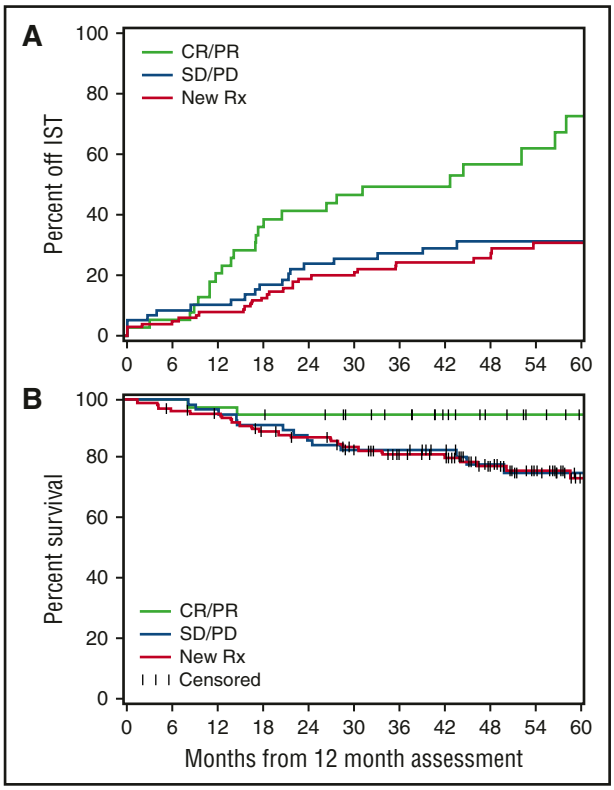

Figure 2. Patients with a CR or PR at the 1-year landmark and no secondary systemic treatment before the 1-year landmark have a shorter time to end of systemic treatment and better subsequent survival than do those with stable disease or progression at the landmark or secondary systemic treatment before the landmark. (A) Cumulative incidence of ended systemic treatment. (B) Survival after the 1-year landmark. CR/PR indicates complete or partial response at the time of assessment (n = 39). SD/PD indicates stable or progressive disease at the time of assessment (n = 59). New Rx indicates patients who received secondary systemic treatment of chronic GVHD before the landmark (n = 104). A few patients ended systemic treatment before the landmark. IST, immunosuppressive treatment.

this study, however, the effect size was too small to be statistically significant, given the numbers of patients in each group. On the other hand, improvements in quality of life measured by FACT-BMT scores were greater in the CR/PR group than in the SD/PD group. The earlier end of systemic treatment and higher survival among patients with CR/PR in comparison with those who had SD/PD and those with secondary systemic treatment offer additional evidence for the clinical benefit of FFS with CR/PR at 1 year as a primary endpoint for future clinical trials. A previous study showed that patients with CR/PR as assessed at 6 months by providers had better subsequent survival than did those with SD/PD.[18] Unlike the present study, this comparison included patients who received secondary treatment before the 6-month landmark.

Trials using FFS with CR/PR as an endpoint must prespecify a time point for assessment of response, because responses are not stable across time.[6] The loss of CR/PR or administration of secondary systemic treatment between 6 months and 1 year in 45% to 56% of patients and the lack of strong association between outcomes at 6 months and rates of ending systemic treatment and subsequent mortality suggest that assessment of outcomes at 6 months after enrollment would be premature in a pivotal trial testing initial treatment. Further analysis is needed to determine the utility of measuring outcomes at 6 months in earlier phase studies.

Strengths of this study include the prospective design, the participation of multiple centers, the collection and cleaning of data in real time with the use of the case-report forms, the use of the 2014 NIH response criteria, and the application of these criteria by algorithms that compared assessments at 1 year with those at baseline. Very few patients were lost to follow-up. As one limitation, this study did not include pediatric patients. The possibility of bias due to delayed enrollment cannot be completely excluded. As many as 147 patients had FFS at 1 year, but 43 of the 147 (29%) of these patients did not have response assessments. These results leave some uncertainty regarding the true rate of FFS with CR/PR at 1 year. Reasons for secondary systemic treatment were not recorded, and in some cases, changes could have been prompted by problems other than inadequately controlled chronic GVHD. Finally, some patients could have been misclassified as having PD if exacerbations had causes other than chronic GVHD. In particular, infections and muscle weakness can decrease FEV1; whether decreased FEV1 in the absence of documented bronchiolitis obliterans or cryptogenic organizing pneumonia should be classified as PD is a question still to be decided.

Despite this uncertainty, we suggest that results of this study could provide benchmarks for future studies testing new interventions for initial systemic treatment of chronic GVHD. Further studies will be needed to identify baseline risk factors associated with outcomes for FFS with CR/PR at 6 months and 1 year. Confidence in the estimated proportion of patients with FFS and CR/PR at 1 year could be strengthened by results from a similar prospective study of a different cohort such as the BMT-CTN 0801 trial, which enrolled 151 patients. A much larger prospective study would be needed to verify whether this endpoint is associated with shorter time to the end of systemic treatment and improved survival. Nonetheless, the estimated 12% to 15% of patients meeting this endpoint in the current study leaves much room for improvement of initial therapy and highlights a major unmet need in the field.

Acknowledgments

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

Authorship

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

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

ORCID profiles: P.J.M., 0000-0001-9051-1215.

Correspondence: Paul J. Martin, Fred Hutchinson Cancer Research Center, PO Box 19024, Seattle, WA 98109-1024; e-mail: pmartin@fredhutch.org.

References

1. Flowers ME, Inamoto Y, Carpenter PA, et al. Comparative analysis of risk factors for acute graft-versus-host disease and for chronic graft-versus-host disease according to National Institutes of Health consensus criteria. *Blood*. 2011;117(11):3214-3219.
2. Socié G, Ritz J. Current issues in chronic graft-versus-host disease. *Blood*. 2014;124(3):374-384.
3. Vigorito AC, Campregher PV, Storer BE, et al; National Institutes of Health. Evaluation of NIH consensus criteria for classification of late acute and chronic GVHD. *Blood*. 2009;114(3):702-708.
4. Newell LF, Flowers ME, Gooley TA, et al. Characteristics of chronic GVHD after cord blood transplantation. *Bone Marrow Transplant*. 2013;48(10):1285-1290.
5. Vogelsang GB. How I treat chronic graft-versus-host disease. *Blood*. 2001;97(5):1196-1201.
6. Flowers ME, Martin PJ. How we treat chronic graft-versus-host disease. *Blood*. 2015;125(4):606-615.
7. Lee SJ. Classification systems for chronic graft-versus-host disease. *Blood*. 2017;129(1):30-37.
8. 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.
9. 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.
10. Inamoto Y, Flowers ME, Sandmaier BM, et al. Failure-free survival after initial systemic treatment of chronic graft-versus-host disease. *Blood*. 2014;124(8):1363-1371.
11. 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.
12. 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.
13. 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.
14. Lee S, Cook EF, Soiffer R, Antin JH. Development and validation of a scale to measure symptoms of chronic graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8(8):444-452.
15. 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(4):357-368.
16. 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.
17. 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(10):1350-1353.
18. 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(1):160-166.