REGULAR ARTICLE

blood advances

The MAGIC algorithm probability is a validated response biomarker of treatment of acute graft-versus-host disease

Hrishikesh K. Srinagesh,[1] Umut Özbek,[2] Urvi Kapoor,[1] Francis Ayuk,[3] Mina Aziz,[1] Kaitlyn Ben-David,[1] Hannah K. Choe,[4] Zachariah DeFilipp,[5] Aaron Etra,[1] Stephan A. Grupp,[6] Matthew J. Hartwell,[1] Elizabeth O. Hexner,[7] William J. Hogan,[8] Alexander B. Karol,[1] Stelios Kasikis,[1] Carrie L. Kitko,[9] Steven Kowalyk,[1] Jung-Yi Lin,[2] Hannah Major-Monfried,[1] Stephan Mielke,[10,11] Pietro Merli,[12] George Morales,[1] Rainer Ordemann,[13] Michael A. Pulsipher,[14] Muna Qayed,[15] Pavan Reddy,[16] Ran Reshef,[17] Wolf Rösler,[18] Karamjeet S. Sandhu,[19] Tal Schechter,[20] Jay Shah,[1] Keith Sigel,[1] Daniela Weber,[21] Matthias Wölfl,[22] Kitsada Wudhikarn,[23] Rachel Young,[1] John E. Levine,[1,*] and James L. M. Ferrara[1,*]

[1]Tisch Cancer Institute and [2]Biostatistics Shared Resource Facility, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY; [3]Department of Stem Cell Transplantation, University Medical Center, Hamburg-Eppendorf, Germany; [4]Blood and Marrow Transplantation Program, The Ohio State University Comprehensive Cancer Center, Columbus, OH; [5]Blood and Marrow Transplant Program, Massachusetts General Hospital, Boston, MA; [6]Division of Oncology, Department of Pediatrics, Center for Childhood Cancer Research, Children's Hospital of Philadelphia and Perelman School of Medicine, and [7]Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA; [8]Blood and Marrow Transplant Program, Division of Hematology, Mayo Clinic, Rochester, MN; [9]Pediatric Blood and Marrow Transplantation Program, Vanderbilt University Medical Center, Nashville, TN; [10]Department of Medicine II, Würzburg University Medical Center, Würzburg, Germany; [11]Cellterapi och Allogen Stamcellstransplantation, Department of Laboratory Medicine, Karolinska University Hospital and Institutet, Stockholm, Sweden; [12]Department of Pediatric Hematology/Oncology, Istituto di Ricovero e Cura a Carattere Scientifico Ospedale Pediatrico Bambino Gesù, Rome, Italy; [13]Medical Department 1, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany; [14]Blood and Marrow Transplantation Program, Children's Hospital Los Angeles, Los Angeles, CA; [15]Pediatric Blood and Marrow Transplantation Program, Aflac Cancer and Blood Disorders Center, Emory University and Children's Healthcare of Atlanta, Atlanta, GA; [16]Blood and Marrow Transplantation Program, University of Michigan, Ann Arbor, MI; [17]Blood and Marrow Transplantation Program, Columbia University Irving Medical Center, New York, NY; [18]Department of Internal Medicine 5, Hematology/Oncology, University Hospital Erlangen, Erlangen, Germany; [19]Hematology and Hematopoietic Cell Transplant, City of Hope Medical Center, Duarte, CA; [20]Division of Haematology/Oncology, The Hospital for Sick Children, University of Toronto, Toronto, Canada; [21]Blood and Marrow Transplantation Program, University of Regensburg, Regensburg, Germany; [22]Pediatric Blood and Marrow Transplantation Program, Children's Hospital, University of Würzburg, Würzburg, Germany; and [23]Blood and Marrow Transplantation Program, Chulalongkorn University, Bangkok, Thailand

Key Points

- The MAGIC algorithm probability, computed from 2 serum biomarkers, predicts mortality in all GVHD grades after 4 weeks of treatment.
- Dynamic changes in the MAGIC algorithm probability occur within all biomarker risk groups and can guide therapy.

The Mount Sinai Acute GVHD International Consortium (MAGIC) algorithm probability (MAP), derived from 2 serum biomarkers, measures damage to crypts in the gastrointestinal tract during graft-versus-host disease (GVHD). We hypothesized that changes in MAP after treatment could validate it as a response biomarker. We prospectively collected serum samples and clinical stages of acute GVHD from 615 patients receiving hematopoietic cell transplantation in 20 centers at initiation of first-line systemic treatment and 4 weeks later. We computed MAPs and clinical responses and compared their abilities to predict 6-month nonrelapse mortality (NRM) in the validation cohort (n = 367). After 4 weeks of treatment, MAPs predicted NRM better than the change in clinical symptoms in all patients and identified 2 groups with significantly different NRM in both clinical responders (40% vs 12%, $P < .0001$) and nonresponders (65% vs 25%, $P < .0001$). MAPs successfully reclassified patients for NRM risk within every clinical grade of acute GVHD after 4 weeks of treatment. At the beginning of treatment, patients with a low MAP that rose above the threshold of 0.290 after 4 weeks of treatment had a significant increase in NRM, whereas patients with a high MAP at onset that fell below that threshold after treatment had a striking decrease in NRM that translated into clear differences in overall survival. We conclude that a MAP measured before and after treatment of acute GVHD is a response biomarker that predicts long-term outcomes more accurately than change in clinical symptoms. MAPs have the potential to guide therapy for acute GVHD and may function as a useful end point in clinical trials.

Submitted 2 August 2019; accepted 19 September 2019. DOI 10.1182/bloodadvances.2019000791.

*J.E.L. and J.L.M.F. contributed equally to this work.

Presented in abstract form at the 61st annual meeting of the American Society of Hematology, Orlando, FL, 8 December 2019.

For original data, please contact james.ferrara@mssm.edu.

The full-text version of this article contains a data supplement.

Introduction

Hematologic malignancies can be cured by hematopoietic cell transplantation (HCT) through a donor lymphocyte-mediated eradication of malignant cells, known as the graft-versus-leukemia effect.[1] Unfortunately, graft-versus-leukemia is closely linked to the toxicity of graft-versus-host disease (GVHD), the leading cause of nonrelapse mortality (NRM) after HCT. Acute GVHD, which typically occurs in 40% to 50% of HCT patients, can be lethal when severe and is graded on a clinical scale of 1 to 4 based on symptoms in the skin, liver, and gastrointestinal (GI) tract.[2,3] Systemic corticosteroids are the primary treatment of significant (grade 2-4) acute GVHD and induce clinical responses in a majority of patients.[4-6] Patients who do not respond to primary therapy within 4 weeks experience long-term NRM from 40% to 70%.[6-8] Thus the change in GVHD clinical staging, or clinical response after 4 weeks of systemic treatment, has served as the primary end point in acute GVHD treatment trials for at least a decade.[5,9]

GVHD in the small and large bowel is the principal driver of NRM, and patients with persistent lower GI GVHD experience an overall survival at 2 years of 25%.[10] In the past decade, 2 validated serum biomarkers have been shown to accurately measure the severity of GI GVHD.[11,12] Regenerating islet-derived 3α (REG3α), a peptide that has antimicrobial and regenerative properties, is released into the systemic circulation from Paneth cells in the intestinal crypt that are damaged during GVHD.[13] Suppressor of tumorigenesis 2 (ST2), the soluble receptor for the alarmin interleukin-33 (IL-33), is shed from multiple cell types when the gastrointestinal crypt is damaged.[14] The 2 biomarkers are combined into a single algorithm developed by the Mount Sinai Acute GVHD International Consortium (MAGIC) to generate an individual patient's estimated probability of 6-month NRM, known as the MAGIC algorithm probability (MAP).[15] Thus measurement in serum of REG3α and ST2 can be considered a "liquid biopsy" of the degree of damage to the lower GI tract caused by GVHD.[6,15,16] We have previously validated MAP as a prognostic biomarker of acute GVHD as defined by the US Food and Drug Administration and the National Institutes of Health.[17] In this study, we measured MAP before and after 4 weeks of treatment of acute GVHD to determine whether a change in MAP could serve as a response biomarker showing that a biological response had occurred after a medical intervention.[17] We also evaluated the MAP in patients of all risk groups before and after treatment defined by either clinical or biomarker parameters.

Methods

Study design and oversight

MAGIC comprises 20 international centers that monitor the clinical status of HCT patients and collect longitudinal serum samples for analysis and storage (supplemental Table 1). Patients from MAGIC centers were enrolled at the time of HCT and all patients were monitored for 6 months for signs and symptoms of acute GVHD. All patients consented to participation in an institutional review board–approved protocol. Patients who received a first allogeneic HCT between 1 January 2008, and 28 February 2018, and subsequently received first-line therapy of acute GVHD that included systemic corticosteroids were consecutively enrolled in this study (supplemental Figure 1). Patients were excluded from the analysis if they did not have serum samples available (n = 531) or if they relapsed and died within 4 weeks of GVHD treatment (n = 3). Patients were divided into sequential training and validation cohorts with roughly equal numbers of NRM events (supplemental Figure 1). All key clinical parameters of acute GVHD and its long-term outcomes were similar between included and excluded patients with the exception of a higher percentage of maximum grade III/IV GVHD among included patients (38% vs 32%) (supplemental Figure 2). We used the training cohort to develop a model to predict 6-month NRM using both biomarker concentrations and clinical responses, and we used the validation cohort to test the results of the original MAP model, multivariable models including the MAP, and the new combined model. Fifty-nine patients of this training cohort had been included in the previously published training cohort that generated the MAP algorithm, and no patients in the validation cohort contributed to the development of that algorithm.[15] Patients in the training cohort underwent first HCT before 31 December 2015 (n = 248), and in the validation cohort after 31 December 2015 (n = 367), so that the validation cohort reflected recent transplant practices such as the increased use of haploidentical donors or posttransplant cyclophosphamide-based GVHD prophylaxis (Table 1).

GVHD clinical criteria

The severity of clinical GVHD was staged using published guidelines.[18,19] The clinical response to treatment was determined at weekly time points during the first month of therapy according to published criteria.[18] All clinical grades of III and IV are reported as combined III/IV, where no differences exist with other clinical grading systems. All MAGIC data coordinators received training in GVHD data extraction from primary source documents and passed a detailed examination before entering data into the database. All data were reviewed centrally by computer logic checks and aberrant or unusual scenarios were queried. Deidentified data were discussed during monthly webinars with senior investigators when appropriate. First- and second-line systemic treatments of acute GVHD are listed in supplemental Table 2. Patients were classified as nonresponders if GVHD symptoms did not improve or progressed, if additional systemic immunosuppression to treat GVHD was prescribed, or if the patient died within the first 4 weeks of treatment. Complete response was defined as complete resolution of GVHD symptoms in all 3 target organs. Partial response was defined as an improvement in stage of all organs with GVHD involvement without complete resolution of symptoms, as previously published.[5,6,19] All causes of death for patients are listed in supplemental Table 3. For 15 patients of the training cohort and 12 patients of the validation cohort who died of acute GVHD before 4 weeks of treatment, both the clinical grade and target organ stage were imputed by carrying forward the last measurement before death.

Biomarker determination

Samples were shipped to a central laboratory, where ST2 and REG3α were analyzed by enzyme-linked immunosorbent assay in batches, as previously described.[6,15] The concentrations of ST2 are reported as picogram per milliliter and of REG3α as nanogram per milliliter. The MAP is calculated as a single value between 0.001 and 0.999 according to the formula: $\log[-\log(1 - \text{MAP})] = -11.263 + 1.844(\log_{10}\text{ST2}) + 0.577(\log_{10}\text{REG3}\alpha)$.[15] At the start of GVHD treatment, 2 thresholds divide MAPs into 3 separate groups with different NRMs, termed the Ann Arbor score.[15] Ann Arbor 1 is

Table 1. Patient characteristics (n = 618)

Characteristic	Training cohort (n = 248)	Validation cohort (n = 367)
Median age (range)	53 y (3 mo-74 y)	54 y (9 mo-77 y)
Pediatric patients (<18 y), n (%)	23 (9)	45 (12)
Indication for HCT, n (%)		
Acute leukemia	120 (48)	197 (54)
MDS/MPN	53 (21)	103 (28)
Lymphoma	32 (13)	32 (9)
Other malignant	31 (12)	18 (5)
Nonmalignant	12 (5)	17 (5)
Donor type, n (%)		
Related	59 (24)	75 (20)
Unrelated	183 (74)	261 (71)
Haploidentical	6 (2)	31 (8)
Cell source, n (%)		
Bone marrow	39 (16)	72 (20)
Umbilical cord blood	26 (10)	18 (5)
Peripheral blood stem cells	183 (74)	277 (75)
HLA match, n (%)		
Matched	167 (67)	266 (72)
Mismatched	75 (30)	70 (19)
Haploidentical	6 (2)	31 (8)
Conditioning regimen intensity, n (%)		
Full	192 (77)	222 (60)
Reduced	56 (23)	145 (40)
ATG, n (%)		
Yes	77 (31)	168 (46)
No	171 (69)	199 (54)
GVHD prophylaxis, n (%)		
CNI/MTX ± other	140 (56)	188 (51)
CNI/MMF ± other	93 (38)	93 (25)
Cyclophosphamide based	6 (2)	48 (13)
Other	9 (4)	19 (5)
T-cell depletion	0 (0)	5 (1)
CNI+sirolimus	0 (0)	14 (4)
GVHD organ distribution at treatment initiation, n (%)		
Stage 0 in all target organs*	4 (2)	5 (1)
Isolated skin	105 (42)	171 (47)
Isolated GI (upper and lower)	78 (31)	113 (31)
Isolated liver	0 (0)	3 (1)
≥2 organs involved	61 (25)	75 (21)
GVHD grade at treatment initiation,[18] n (%)		
I†	75 (30)	112 (31)
II	106 (43)	185 (50)
III	53 (21)	57 (16)
IV	14 (6)	13 (4)
Onset Minnesota Risk,[19] n (%)		
Standard	190 (77)	308 (84)
High	58 (23)	59 (16)
Initial corticosteroid dose (methylprednisolone mg/kg)	1.5 (0.24-2.7)	1.0 (0.09-3.3)
Clinical response after 4 wk of treatment, n (%)		
CR	138 (56)	230 (63)
PR	32 (13)	37 (10)
NR	78 (31)	100 (27)
Long-term outcomes, n (%)		
1-y NRM	61 (25)	81 (22)
1-y Relapse rate	47 (19)	51 (14)
1-y OS	159 (64)	260 (71)

ATG, antithymocyte globulin; CNI, calcineurin inhibitor; MDS, myelodysplastic syndromes; MMF, mycophenolate mofetil; MPN, myeloproliferative neoplasms; MTX, methotrexate.

*Reasons for treatment of stage 0 GVHD in all target organs include biopsy-proven GVHD without clinical symptoms.

†Grade 0 and I were combined for all analyses.

defined as MAP < 0.141, Ann Arbor 2 as 0.141 ≤ MAP ≤ 0.290, and Ann Arbor 3 as MAP > 0.290. Following treatment, a single threshold (0.290) divides MAPs into 2 groups with significantly different NRMs.[6] This validated scoring system of prognostic biomarkers is now widely used and commercially available. Missing biomarker data for patients who died before 4 weeks of treatment were imputed by carrying forward the last measurement before death. If data were missing for reasons other than early death, no imputation was made.

Statistical analyses

The effects of the change in clinical symptoms (complete or partial response vs no response) and the MAP after week 4 of treatment (high vs low) on the hazard of 6-month NRM were evaluated in univariable and multivariable competing risk regression models that considered relapse and second allogeneic HCT as competing risks.[20] We developed univariable competing risk regression models of 6-month NRM using all clinical variables with sufficient NRM events in the training cohort, with Minnesota risk as the measure of clinical GVHD severity. We then tested significant variables in a multivariable model for their impact on the ability of the MAP to predict 6-month NRM in the validation cohort. The cumulative incidence of NRM and relapse were measured for 12 months after treatment and differences between groups were compared using Gray test.[21] Crude proportions of 6-month NRM were compared using χ^2 or Fisher's exact test as appropriate. Overall survival (OS) was estimated via the Kaplan-Meier method and differences between the aforementioned groups were calculated using the log-rank test. The area under receiver operating characteristic (ROC) curves (AUC) were compared using the DeLong method.[22] The competing risks regression model to predict 6-month NRM that combined clinical responses and MAP considered relapse and second allogeneic HCT as competing risks. Differences in MAP between groups were compared using the

Mann-Whitney *U* test. Patients with a single biomarker evaluation at baseline because of death before a second measurement (n = 3) were not included in the analysis of changes in MAPs. *P* values were corrected for family-wise multiple comparisons using the Holm-Bonferroni method.[23] All analyses were performed using R statistical package, version 3.5.1 (R Core Team 2018).

Results

Patient characteristics

Clinical data and samples were available for 615 patients with acute GVHD whose first-line treatment included systemic corticosteroids. Patients were divided by date of HCT into a training cohort (n = 248) and a validation cohort (n = 367). Pretransplant clinical characteristics (Table 1) that differed between the 2 cohorts reflect more recent transplant practices in the validation cohort, including an increased use of haploidentical donors and the use of posttransplant cyclophosphamide-based GVHD prophylaxis. OS at 1 year in the validation cohort reflects improving trends in survival for HCT patients.[24] Results of all analyses are displayed for the validation cohort. All participating MAGIC centers had similar cumulative incidences of 6-month NRM (data not shown).

Clinical responses and MAPs after 4 weeks of GVHD treatment

We first evaluated the incidence of NRM in patients with acute GVHD who experienced a complete resolution (CR) of clinical symptoms, a partial resolution of clinical symptoms (PR), or no resolution of clinical symptoms after 4 weeks of systemic treatment. Patients with a CR or PR had similar incidences of NRM (15% and 16% respectively), and these groups were therefore combined for all analyses (supplemental Figure 3). As expected, patients with no resolution of clinical symptoms experienced significantly greater 12-month NRM than patients with a CR or PR. Patients in the validation cohort were treated with a broad range of initial doses of corticosteroids, but the dose did not correlate with long-term outcomes (supplemental Table 4).

We developed both univariable and multivariable competing risk regression models to evaluate the effect of clinical responses and MAPs measured after 4 weeks of treatment on the hazard of 6-month NRM. The hazard ratio (HR) for both clinical responses (4.97; 95% confidence interval [CI], 2.95-8.36) and MAPs (9.84; 95% CI, 5.82-16.6) were significant in the univariable model (supplemental Figure 4A). When we created a multivariable model using both of these metrics, we observed the HR of MAPs to be greater than twice that of clinical responses (HR = 6.91 vs 2.60). We then created ROC curves for each metric and found the AUC for MAPs to be significantly greater than that of clinical responses (0.86 vs 0.70, $P < .0001$; supplemental Figure 4B). When the 12 patients who died within the first month were excluded in a landmark analysis, the AUC for MAPs remained 20 points higher than that for clinical responses (0.84 vs 0.65, $P < .0001$). A previous study identified a single threshold (0.290) that divides MAPs obtained after 1 week of treatment of acute GVHD into 2 groups with distinctly different 12-month NRM.[6] Application of that threshold to patient samples after 4 weeks of treatment again identified groups with significantly different 12-month NRM (supplemental Figure 5). The MAP did not correlate with relapse and thus the differences in 12-month NRM were reflected in OS (supplemental Figure 5). These differences were not especially sensitive to a threshold effect because multiple thresholds both above and below 0.290 classified patients into groups with >40% differences in 6-month NRM (supplemental Table 6).

To test the possibility that the prediction of 6-month NRM by the week 4 MAP might be influenced by clinical variables, we performed a univariable analysis in the training cohort of key pretransplant and GVHD clinical variables listed in Table 1. Only Minnesota risk at GVHD onset and age were significant predictors of 6-month NRM (supplemental Table 4). In multivariable analyses of the validation cohort, age was not an independent predictor of NRM. Minnesota (high) risk remained a significant but less powerful predictor of 6-month NRM than the high week 4 MAP (HR, 2.0 and 8.5, respectively) (supplemental Table 5).

To evaluate further the ability of clinical response to predict NRM, we next created cumulative incidence curves of NRM for each clinical grade of GVHD at the time of treatment. Relapse and second allogeneic HCT were considered competing risks. When we analyzed each clinical grade at the time of treatment, the clinical response did not predict NRM for the 30% of patients treated for grade 1 GVHD, but classification by high vs low MAP predicted 6-month NRM in all clinical grades (I, II, III/IV) (supplemental Figure 6). Among all patients with a clinical response, the minority (10%) had high MAPs and experienced threefold greater 12-month NRM than the low MAP group (40% vs 12%, $P < .0001$) (Figure 1A, left). In patients with no clinical response, the majority (57%) had low MAPs and experienced almost threefold lower 12-month NRM than those with high MAPs (25% vs 65%, $P < .0001$) (Figure 1A, right). With respect to 6-month NRM, the poor positive predictive value of 35% for no clinical response improved significantly to 51% for a high MAP because of increases in both sensitivity and specificity (supplemental Table 7). Because lower GI GVHD is the principal driver of NRM, we also analyzed outcomes for patients with and without lower GI symptoms. MAPs categorized patients independently of the presence of lower GI GVHD either during or at the end of the first month of therapy (supplemental Figure 7). These results are consistent with our previous study that showed the GI biomarker REG3α more accurately reflects overall damage to GI crypts than the severity of lower GI symptoms.[13]

Because patients with incomplete responses can have persistent symptoms of differing severity, we also examined NRM within each clinical GVHD grade after 4 weeks of treatment. As expected, 6-month NRM increased with each clinical GVHD grade at the end of 4 weeks of treatment, but clinical nonresponders within each clinical grade had the same risk for 6-month NRM as clinical responders (Figure 1B, left). MAP, however, classified patients within all 3 clinical GVHD grades into 2 significantly different groups (Figure 1B, right), demonstrating their utility throughout the spectrum of clinical GVHD severity after 4 weeks of treatment.

Weekly clinical responses and MAPs

Because both clinical responses and MAPs after 4 weeks of treatment remained significant in the multivariable analysis of 6-month NRM, we hypothesized that their combination might improve the prediction of long-term outcomes compared with either metric alone. We therefore created a new model to predict 6-month NRM using the clinical response, the concentration of REG3α, and the concentration of ST2, all measured after 4 weeks of therapy.

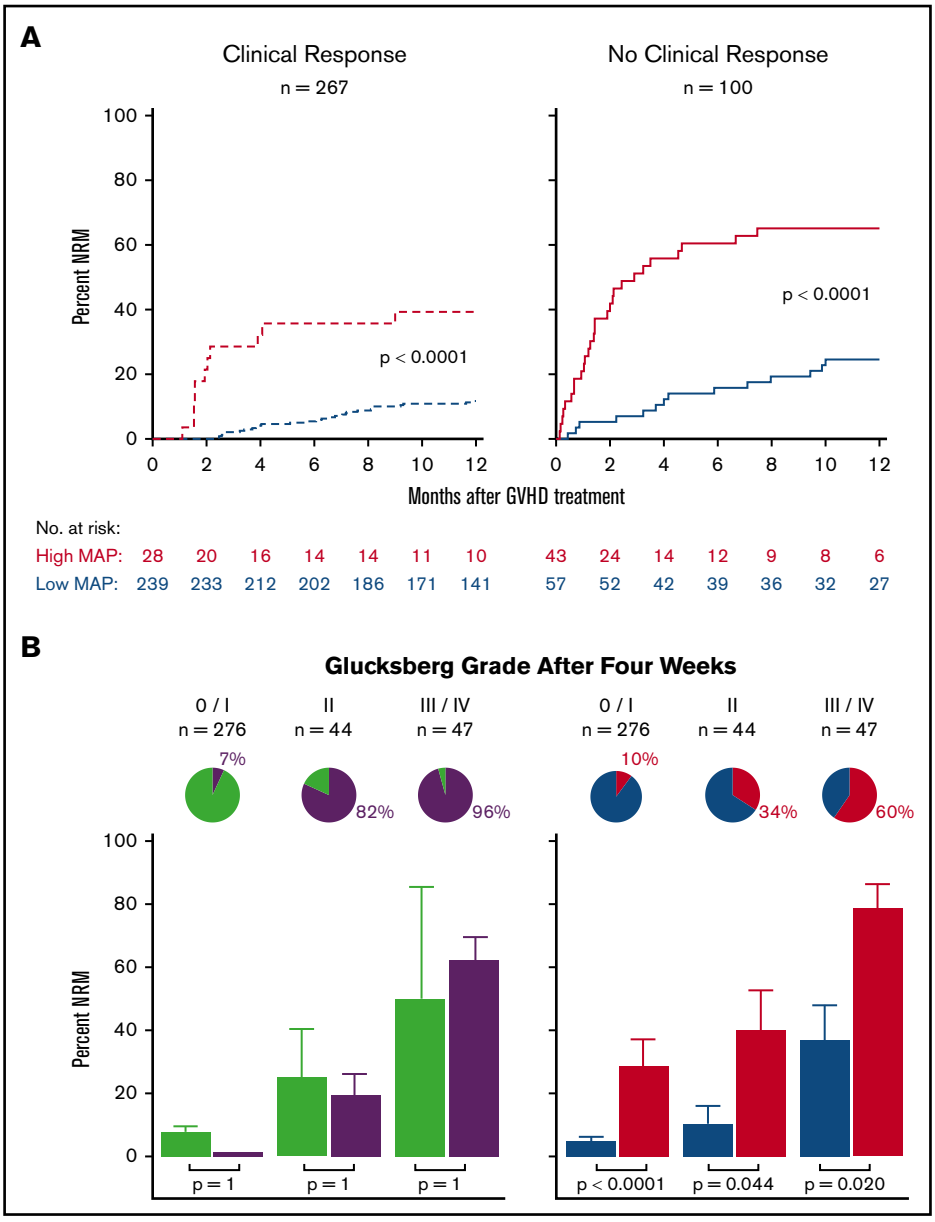

Figure 1. Prediction of NRM by MAP and clinical responses after 4 weeks of systemic therapy for GVHD. (A) Cumulative incidence of NRM for patients according to clinical response to GVHD therapy (left) and no response (right) analyzed by high (red dash) or low (blue dash) MAP. (B) Crude proportion of 6-month NRM (± standard error) for each clinical GVHD grade after 4 weeks of treatment according to response (light green box) or no response (purple box) of clinical symptoms (left), or low (blue box) or high (red box) MAP (right).

Contrary to our hypothesis, the AUC of the new algorithm did not differ from that of the MAP when applied to the validation cohort, although both were significantly more accurate than clinical responses alone (Figure 2).

Because GVHD is a dynamic process in which symptoms can change after therapy, we evaluated whether the MAP could serve as a monitoring biomarker of acute GVHD, or that that can be used serially over time to measure disease burden.[17] We computed ROC curves of 6-month NRM for clinical responses, MAPs, and the combined algorithm after 1 and 2 weeks of treatment. The AUC of the new algorithm did not differ from that of the MAP at any time point, although both algorithms containing biomarker concentrations were more accurate than clinical responses considered in isolation (Figure 2). Thus biomarker probabilities, either alone or in combination with clinical responses, are better monitoring biomarkers than clinical responses alone and provide a better surrogate end point for long-term NRM after acute GVHD treatment.

MAP as response biomarker

We next formally analyzed whether the MAP could serve as a response biomarker for the treatment of acute GVHD by comparing it before and after 4 weeks of treatment. The MAP at the initiation of treatment determines the Ann Arbor GVHD score, which classifies patients into 3 groups with distinctly different risks of long-term NRM.[15] The average MAP in each Ann Arbor score reflected the cumulative incidence of 6-month NRM (supplemental Table 8). In the validation cohort, all 3 Ann Arbor scores were present within each clinical GVHD grade at the onset of treatment (supplemental Figure 8), confirming the findings of earlier studies in completely independent groups of patients.[15,16] Six-month NRM clustered in patients with the greatest increases in MAP after

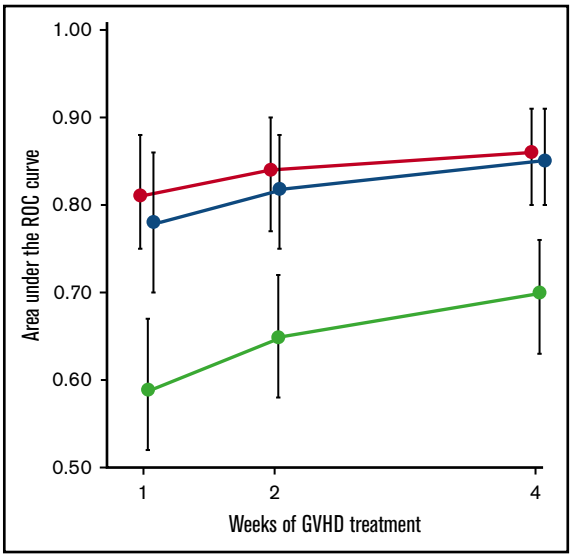

Figure 2. Prediction of 6-month NRM during first month of GVHD therapy. AUCs of ROC curves (±95% CI) for MAPs (red circle), clinical responses (green circle), and a new algorithm combining biomarkers and clinical responses (blue circle) after 1, 2, and 4 weeks of treatment of GVHD. Patients with both clinical responses and biomarker values were included at each timepoint (week 1: n = 321; week 2: n = 323; week 4: n = 367). The difference between MAP and clinical responses alone was $P < .0001$ at all time points. The AUC of MAPs at week 1 (0.81) is significantly greater than that of clinical response at week 4 (0.70) ($P = .0029$).

4 weeks, whereas survival clustered in patients with the greatest decreases (Figure 3A). These changes in MAP between patients with and without 6-month NRM were significant in all 3 Ann Arbor groups (Figure 3B). In patients with the lowest MAP before treatment (Ann Arbor 1), 6-month NRM predominated in patients with the largest increases in MAP, whereas in patients with the highest MAP before treatment (Ann Arbor 3), survival predominated in patients with the largest decreases in MAP. The MAP also served as a response biomarker in patients who had received posttransplant cyclophosphamide-based GVHD prophylaxis or when the analysis was restricted to patients receiving systemic corticosteroids alone as first-line therapy for GVHD (supplemental Figure 9).

We next hypothesized that movement of the MAP across a threshold of 0.290, as defined in an earlier study, would predict long-term survival within each Ann Arbor group.[6] As shown in Figure 4A, an increase above the threshold after 4 weeks of treatment predicted significantly worse 6-month NRM, and remaining below the threshold predicted significantly improved 6-month NRM. Among patients in the higher risk groups 27% of Ann Arbor 2 patients rose above the threshold, and 34% of Ann Arbor 3 patients decreased below the threshold. Overall, 14.7% of all patients crossed the threshold during the first month of therapy, which predicted dramatic differences in OS at 1 year within each Ann Arbor group (Figure 4B). A landmark analysis of patients surviving to day 28 produced nearly identical results (supplemental Figure 10). Thus we confirmed our central hypothesis and have validated the change in MAP after 4 weeks as a response biomarker of acute GVHD treatment.

We performed an identical set of analyses for changes in MAP after 2 weeks of treatment. There was a significant change in MAP between patients with and without 6-month NRM in both Ann Arbor 1 and 2 groups, and a trend toward significance in the Ann Arbor 3 group (supplemental Figure 11). Although there was a significant difference in 6-month NRM for all patients where the MAP crossed the 0.290 threshold, this change in NRM predicted significant differences in OS for Ann Arbor 1 and Ann Arbor 2 patients only (supplemental Figure 11). Thus although MAPs measured after 2 weeks provided useful information in the majority of patients, they were not as accurate as changes after 4 weeks of treatment in predicting long-term outcomes.

Discussion

In this large multicenter validation study, a computed probability, or MAP, calculated from the concentrations of 2 serum biomarkers more accurately predicted 6-month NRM than the change in clinical symptoms at every time point tested during the first month of systemic therapy for acute GVHD. Importantly, MAPs reclassified patients within each clinical response category, and MAPs after just 1 week of treatment more accurately predicted long-term outcomes

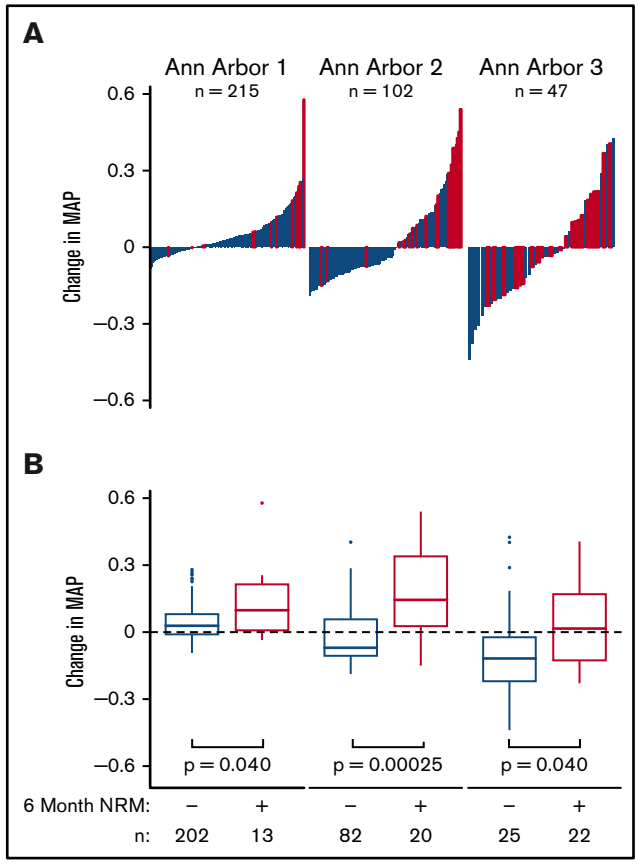

Figure 3. Changes in MAP after 4 weeks according to initial Ann Arbor score. (A) Reverse waterfall plots of changes in MAP and (B) box-and-whisker plots of changes in MAP in consecutive patients who provided samples before and after 4 weeks of treatment according to initial Ann Arbor score in patients with (red line) and without (blue line) 6-month NRM. (Left) Ann Arbor 1 patients (MAP < .141 at treatment initiation). (Center) Ann Arbor 2 patients (0.141 ≤ MAP ≤ .290 at treatment initiation). (Right) Ann Arbor 3 patients (MAP > .290).

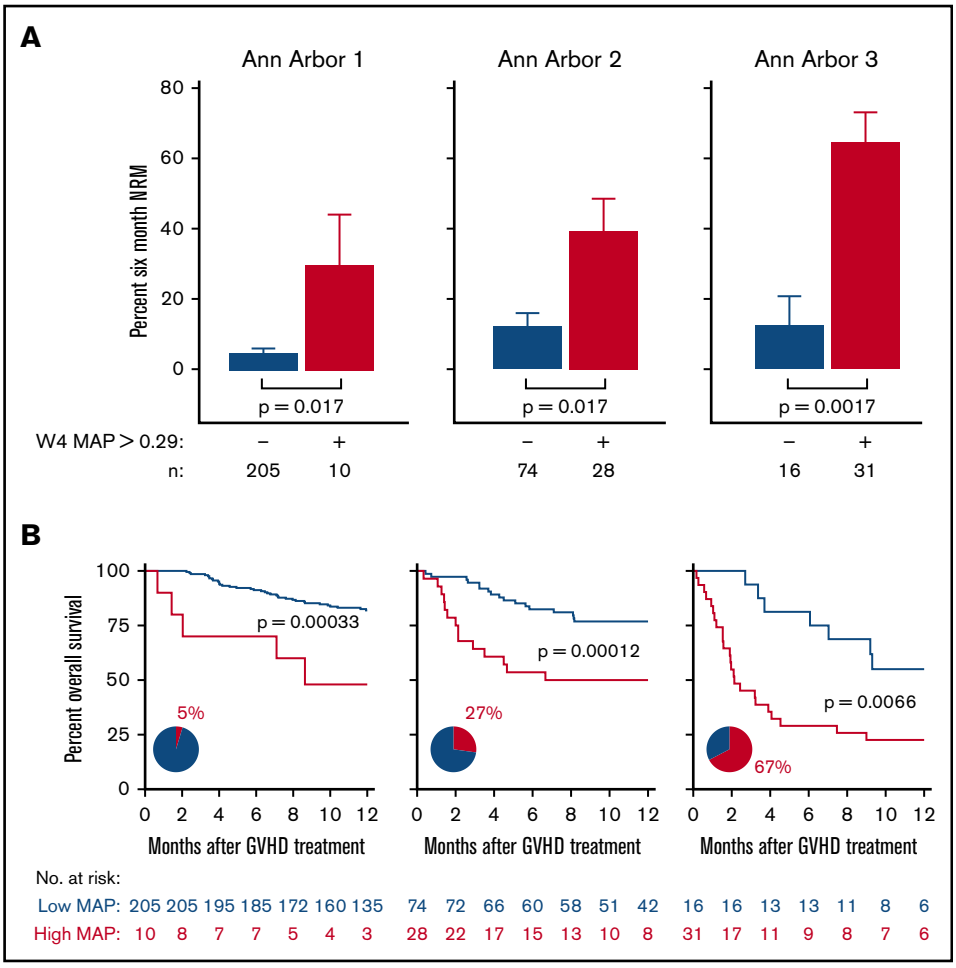

Figure 4. Long-term mortality by MAP threshold (0.290) after 4 weeks of treatment. (A) Crude proportions of 6-month nonrelapse mortality (± standard error) and (B) Kaplan-Meier estimates of overall survival according to Ann Arbor score for patients whose MAP after 4 weeks of treatment rose/remained above (red line) or fell/remained below (blue line) the threshold of 0.290. Ann Arbor scores were determined as in Figure 3.

than clinical responses measured 3 weeks later, the current gold standard for clinical trials of acute GVHD treatment (Figure 2). Furthermore, when patients were analyzed according to GVHD clinical grades after 4 weeks of treatment, MAPs further segregated all grades into 2 groups with significantly different 6-month NRM, whereas the clinical response was unable to predict differences in 6-month NRM in any grade. An additional novel finding in this study is that the changes in MAP after 4 weeks of treatment predicted long-term outcomes in all 3 Ann Arbor groups, and movement of the MAP across the threshold of 0.290 identified 15% of patients with significant differences in survival. Thus we have validated the MAP as a response biomarker for the treatment of acute GVHD and found that it is superior to the current gold standard, clinical response after 4 weeks. These results also suggest that a treatment goal for patients with acute GVHD would be to achieve a MAP of $\leq$0.290.

The ability of MAPs to forecast long-term outcomes is likely related to its accurate reflection of GVHD damage to crypts throughout the lower GI tract that are the primary sources of the biomarkers REG3α and ST2. The alarmin IL-33 binds its soluble receptor ST2 that is shed from damaged stromal, endothelial, and epithelial cells.[14] REG3α is released from damaged Paneth cells that are needed for crypt regeneration, helping to explain why this biomarker quantitates crypt damage better than the volume of diarrhea (supplemental Figure 7).[13,25] MAPs that improves over time presumably reflect healing of the lower GI tract and predict long-term survival regardless of the Ann Arbor score at start of treatment (Figure 4).

Even though the inclusion of the clinical response to treatment did not improve the predictive accuracy of the MAP, the clinical severity of GVHD remains an important consideration for individual patients. For example, in patients with high MAPs at the beginning of treatment (Ann Arbor 3), the 6-month NRM for patients with Glucksberg III/IV is twice that of Glucksberg II. Similarly, at the end of 4 weeks of treatment, the 12-month NRM in patients with high MAPs is twice as high if lower GI symptoms are present (supplemental Figure 7).

These results confirm and extend observations from previous studies that demonstrate a single measurement of the MAP predicts key long-term outcomes of acute GVHD.[6,15,16] Many clinicians now use the MAP in patients whose GVHD symptoms have not completely responded after 1 week of systemic corticosteroid therapy. The current study shows that the change in MAP more accurately predicts long-term outcomes than the change in clinical symptoms. This information is most likely to be useful in patients presenting with higher clinical staging at the time of treatment and whose first MAP determination coincides with the start of first-line therapy; the change in MAP after several weeks of therapy will provide a more

accurate picture of GVHD resolution or progression than the change (or lack thereof) in clinical symptoms, particularly in the critical target organ of the GI tract.

Our results also have important implications for the conduct of clinical trials of therapy for acute GVHD. The clinical response after 4 weeks of GVHD therapy currently serves as the primary end point of clinical trials because it predicts long-term outcomes better than the response after 2 weeks of therapy.[5] In this study, the clinical response to treatment did not provide additional prognostic information regarding NRM independently of the clinical GVHD grade at 4 weeks. This prognostic failure may be due to the fact that clinical response metrics weight the changes in clinical symptoms of all involved organs equally, despite the primary driver of NRM being damage to the GI tract. The ability of the MAP to predict NRM within each clinical grade after 4 weeks is likely from its more accurate estimation of the totality of damage to the lower GI tract than that provided by clinical symptom severity. In this respect, MAGIC biomarkers may be similar to other laboratory end points such as quantitative polymerase chain reaction and fluorescence-activated cell sorting analysis that more accurately quantitate disease burdens and that have replaced clinical surrogate end points in HIV infections and B-cell acute lymphoblastic leukemia, respectively.[26,27] Because both GVHD and its profoundly immunosuppressive treatments have potentially lethal complications, the ability to quantify response more accurately may help optimize therapies and finally improve outcomes for GVHD.

Our study has several important limitations. It does not determine whether the MAP predicts the response to a particular intervention, which is a critical question that must be answered to develop novel therapeutics for acute GVHD. Additionally, a large majority of patients who are treated for acute GVHD have low MAP at the start of treatment and low incidences of NRM. Prospective trials are needed to determine whether the MAP can be used to identify patients who will tolerate rapid tapers of systemic corticosteroids or who could benefit from through nonsteroidal strategies, thereby avoiding the significant morbidity and mortality associated with prolonged high-dose corticosteroid therapy. Clinical trials of such new approaches to GVHD treatment are now warranted using MAPs as either eligibility criteria, response end points, or both.

Acknowledgments

The authors thank the patients, their families, and the research staff for their participation.

This work was supported by National Institutes of Health, National Cancer Institute grants (P01CA03942 and P30CA196521) and a National Institutes of Health, National Center for Advancing Translational Sciences grant (TL1 TR001434).

Authorship

Contribution: H.K.S. wrote the report; J.E.L. and J.L.M.F. designed and supervised all aspects of the study; all authors designed the study, interpreted data, and contributed to writing the report; U.K., F.A., H.K.C., Z.D., A.E., S.A.G., S.K., E.O.H., W.J.H., C.L.K., S.M., P.M., R.O., M.A.P., M.Q., P.R., R.R., W.R., K.S.S., T.S., J.S., D.W., M.W., K.W., R.Y., and J.E.L. collected and reviewed the clinical data; H.K.S., M.A., K.B.-D., M.J.H., A.B.K., S.K., G.M., and H.M.-M. performed the laboratory analyses; and U.Ö. and J.-Y.L. performed the statistical analyses.

Conflict-of-interest disclosure: U.M., J.E.L., and J.L.M.F. are coinventors on a GVHD biomarker patent. The remaining authors declare no competing financial interests.

ORCID profiles: H.K.S., 0000-0001-9786-3010; Z.D., 0000-0002-7994-8974; M.J.H., 0000-0003-0095-1271; E.O.H., 0000-0002-1125-4060; A.B.K., 0000-0001-8277-9932; J.-Y.L., 0000-0002-0246-3514; S.M., 0000-0002-8325-9215; P.M., 0000-0001-6426-4046; M.A.P., 0000-0003-3030-8420; R.R., 0000-0003-2185-9546; K.S., 0000-0002-4051-4861; K.W., 0000-0001-9528-8681.

Correspondence: James L. M. Ferrara, Hess Center for Science and Medicine, Icahn School of Medicine at Mount Sinai, 1470 Madison Ave, 6th Floor, New York, NY 10029; e-mail: james.ferrara@mssm.edu.

References

1. Ferrara JLM, Levine JE, Reddy P, Holler E. Graft-versus-host disease. *Lancet*. 2009;373(9674):1550-1561.
2. Zeiser R, Blazar BR. Acute graft-versus-host disease - biologic process, prevention, and therapy. *N Engl J Med*. 2017;377(22):2167-2179.
3. Jagasia M, Arora M, Flowers MED, et al. Risk factors for acute GVHD and survival after hematopoietic cell transplantation. *Blood*. 2012;119(1):296-307.
4. MacMillan ML, Weisdorf DJ, Wagner JE, et al. Response of 443 patients to steroids as primary therapy for acute graft-versus-host disease: comparison of grading systems. *Biol Blood Marrow Transplant*. 2002;8(7):387-394.
5. MacMillan ML, DeFor TE, Weisdorf DJ. The best endpoint for acute GVHD treatment trials. *Blood*. 2010;115(26):5412-5417.
6. Major-Monfried H, Renteria AS, Pawarode A, et al. MAGIC biomarkers predict long-term outcomes for steroid-resistant acute GVHD. *Blood*. 2018;131(25):2846-2855.
7. MacMillan ML, Weisdorf DJ, Davies SM, et al. Early antithymocyte globulin therapy improves survival in patients with steroid-resistant acute graft-versus-host disease. *Biol Blood Marrow Transplant*. 2002;8(1):40-46.
8. Socié G, Vigouroux S, Yakoub-Agha I, et al. A phase 3 randomized trial comparing inolimomab vs usual care in steroid-resistant acute GVHD. *Blood*. 2017;129(5):643-649.
9. Martin PJ, Bachier CR, Klingemann H-G, et al. Endpoints for clinical trials testing treatment of acute graft-versus-host disease: a joint statement. *Biol Blood Marrow Transplant*. 2009;15(7):777-784.
10. Castilla-Llorente C, Martin PJ, McDonald GB, et al. Prognostic factors and outcomes of severe gastrointestinal GVHD after allogeneic hematopoietic cell transplantation. *Bone Marrow Transplant*. 2014;49(7):966-971.

11. Vander Lugt MT, Braun TM, Hanash S, et al. ST2 as a marker for risk of therapy-resistant graft-versus-host disease and death. *N Engl J Med*. 2013; 369(6):529-539.

12. Ferrara JLM, Harris AC, Greenson JK, et al. Regenerating islet-derived 3-alpha is a biomarker of gastrointestinal graft-versus-host disease. *Blood*. 2011; 118(25):6702-6708.

13. Zhao D, Kim Y-H, Jeong S, et al. Survival signal REG3α prevents crypt apoptosis to control acute gastrointestinal graft-versus-host disease. *J Clin Invest*. 2018;128(11):4970-4979.

14. Reichenbach DK, Schwarze V, Matta BM, et al. The IL-33/ST2 axis augments effector T cell responses during acute GVHD. *Blood*. 2015;125(20): 3183-3192.

15. Hartwell MJ, Özbek U, Holler E, et al. An early-biomarker algorithm predicts lethal graft-versus-host disease and survival. *JCI Insight*. 2017;2(3):e89798.

16. Levine JE, Braun TM, Harris AC, et al; Blood and Marrow Transplant Clinical Trials Network. A prognostic score for acute graft-versus-host disease based on biomarkers: a multicentre study. *Lancet Haematol*. 2015;2(1):e21-e29.

17. FDA-NIH Biomarker Working Group. BEST (Biomarkers, Endpoints, and Other Tools) Resource. Silver Spring, MD: US Food and Drug Administration; 2016. http://www.ncbi.nlm.nih.gov/books/NBK326791/. Accessed 21 June 2019.

18. Harris AC, Young R, Devine S, et al. International, multi-center standardization of acute graft-versus-host disease clinical data collection: a report from the MAGIC consortium. *Biol Blood Marrow Transplant*. 2016;22(1):4-10.

19. MacMillan ML, Robin M, Harris AC, et al. A refined risk score for acute graft-versus-host disease that predicts response to initial therapy, survival, and transplant-related mortality. *Biol Blood Marrow Transplant*. 2015;21(4):761-767.

20. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496-509.

21. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat*. 1988;16(3):1141-1154.

22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

23. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6(2):65-70.

24. Gooley TA, Chien JW, Pergam SA, et al. Reduced mortality after allogeneic hematopoietic-cell transplantation. *N Engl J Med*. 2010;363(22):2091-2101.

25. Levine JE, Huber E, Hammer STG, et al. Low Paneth cell numbers at onset of gastrointestinal graft-versus-host disease identify patients at high risk for nonrelapse mortality. *Blood*. 2013;122(8):1505-1509.

26. Murray JS, Elashoff MR, Iacono-Connors LC, Cvetkovich TA, Struble KA. The use of plasma HIV RNA as a study endpoint in efficacy trials of antiretroviral drugs. *AIDS*. 1999;13(7):797-804.

27. Gökbuget N, Dombret H, Bonifacio M, et al. Blinatumomab for minimal residual disease in adults with B-cell precursor acute lymphoblastic leukemia. *Blood*. 2018;131(14):1522-1531.