REGULAR ARTICLE

blood advances

The MAGIC composite response: a novel end point integrating clinical and biomarker parameters for acute GVHD

Yu Akahoshi,[1,2] Joseph Portelli,[1] Nikolaos Katsivelos,[1] Ioannis E. Louloudis,[1] Paibel Aguayo-Hiraldo,[3] Francis Ayuk,[4] Chantiya Chanswangphuwana,[5] Hannah K. Choe,[6] Matthias Eder,[7] Aaron M. Etra,[1] Elizabeth O. Hexner,[8] Carrie L. Kitko,[9] Sabrina Kraus,[10] Pietro Merli,[11] Timothy S. Olson,[12] Ivan Pasic,[13] Muna Qayed,[14] Ran Reshef,[15] Tal Schechter,[16] Julia Marx,[17] Evelyn Ullrich,[18] Ingrid Vasova,[19] Daniela Weber,[20] Matthias Wölfl,[21] Robert Zeiser,[22] Janna Baez,[1] Gilbert Eng,[1] Sigrun Gleich,[20] Steven Kowalyk,[1] George Morales,[1] Nikolaos Spyrou,[1] Rachel Young,[1] Zachariah DeFilipp,[23] William J. Hogan,[24] Ryotaro Nakamura,[25] John E. Levine,[1,*] and James L. M. Ferrara[1,*]

[1]Division of Hematology/Medical Oncology, The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY; [2]Division of Hematopoietic Stem Cell Transplantation, National Cancer Center Hospital, Tokyo, Japan; [3]Division of Bone Marrow Transplantation, Children's Hospital Los Angeles, University of Southern California, Los Angeles, CA; [4]Department of Stem Cell Transplantation, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [5]Division of Hematology and Center of Excellence in Translational Hematology, Faculty of Medicine, Chulalongkorn University and King Chulalongkorn Memorial Hospital, Bangkok, Thailand; [6]Division of Hematology, Blood and Marrow Transplantation Program, The Ohio State University Comprehensive Cancer Center, Columbus, OH; [7]Department of Hematology, Hemostasis, Oncology and Stem Cell Transplantation, Hannover Medical School, Hannover, Germany; [8]Department of Medicine and Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [9]Pediatric Hematology/Oncology Division, Vanderbilt University Medical Center, Nashville, TN; [10]Department of Internal Medicine II, University Hospital of Würzburg, Würzburg, Germany; [11]Department of Pediatric Hematology/Oncology and of Cell and Gene Therapy, Bambino Gesù Children's Hospital, Istituto di Ricovero e Cura a Carattere Scientifico, Rome, Italy; [12]Division of Oncology, Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA; [13]Division of Medical Oncology and Hematology, Princess Margaret Hospital, Toronto, Canada; [14]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA; [15]Division of Hematology/Oncology and Columbia Center for Translational Immunology, Columbia University Irving Medical Center, New York, NY; [16]Division of Hematology/Oncology/Blood and Marrow Transplant, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada; [17]Department of Medicine A, Hematology and Oncology, University Hospital Münster, Münster, Germany; [18]Department of Pediatrics, Experimental Immunology and Cell Therapy, Goethe University Frankfurt, Frankfurt am Main, Germany; [19]Department of Internal Medicine 5, Hematology and Oncology, Friedrich-Alexander-Universität Erlangen-Nürnberg and University Hospital Erlangen, Erlangen, Germany; [20]Department of Hematology and Oncology, Internal Medicine III, University of Regensburg, Regensburg, Germany; [21]Department of Pediatrics, Children's Hospital, University Hospital of Würzburg, Würzburg, Germany; [22]Department of Medicine I, University of Freiburg Medical Center, Faculty of Medicine, University of Freiburg, Freiburg, Germany; [23]Hematopoietic Cell Transplant and Cellular Therapy Program, Massachusetts General Hospital, Boston, MA; [24]Division of Hematology, Mayo Clinic, Rochester, MN; and [25]Department of Hematology/Hematopoietic Cell Transplantation, City of Hope, Duarte, CA

Key Points

- The MCR integrates clinical and biomarker data to predict long-term outcomes more accurately.
- When biomarkers are unavailable at onset, MCR remains superior to changes in clinical severity and can serve as a better trial end point.

Changes in the clinical symptoms of acute graft-versus-host disease (GVHD) are currently used to assess treatment responses. The Mount Sinai Acute GVHD International Consortium (MAGIC) consortium has recently revealed that the integration of serum biomarkers with clinical symptoms at the onset of treatment in a MAGIC composite score (MCS) more accurately predicts treatment response and 6-month nonrelapse mortality (NRM) than clinical symptoms alone. In this study, we evaluated whether the integration of serum biomarkers and clinical symptoms on day 28 (D28) would also better predict NRM than clinical response only (CRO). We analyzed data from 1135 patients receiving systemic treatment for acute GVHD and created a fourth MCS category for patients with complete resolution of symptoms and low-risk clinical biomarkers on D28. Using a classification and regression tree model with 6-month NRM as the end point, we identified status of MCS 0 or MCS 1 at D28 as responses, which we termed the MAGIC composite response (MCR). In the validation cohort (n = 309), MCR more accurately predicted 6-month NRM than CRO (area under the curve: 0.77 vs 0.69; P = .014) and demonstrated higher negative and positive

Submitted 9 May 2025; accepted 26 July 2025; prepublished online on *Blood Advances* First Edition 15 August 2025. https://doi.org/10.1182/bloodadvances.2025017116.

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

Data are available upon reasonable request from the corresponding authors, John E. Levine (john.levine@mssm.edu) and James L. M. Ferrara (james.ferrara@mssm.edu), or the author, Yu Akahoshi (akahoshiu@gmail.com).

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

predictive values. MCR correctly reclassified both clinical nonresponders and responders: 28 of 213 clinical responders (13%) became nonresponders with fivefold higher NRM (34.3% vs 6.8%, $P < .001$) and a larger group (29/96, 30%) of clinical nonresponders became responders with sixfold lower NRM (7.6% vs 50.7%, $P < .001$). These findings support the use of MCR as a superior surrogate end point for long-term GVHD control and survival in future clinical trials.

Introduction

Acute graft-versus-host disease (GVHD) remains a significant cause of morbidity and mortality after allogeneic hematopoietic cell transplantation (HCT).[1,2] Reduction in the severity of day 28 (D28) clinical symptoms after treatment initiation serves as the gold standard for response in clinical trials.[3-7] These clinical response criteria have notable limitations including modest prediction of nonrelapse mortality (NRM) because all target organs are treated equally despite clear evidence that gastrointestinal (GI) tract involvement disproportionately influences the risk of NRM.[8,9]

Several studies have highlighted the potential of serum biomarkers to enhance the precision of response assessments afforded by clinical manifestations alone.[10-19] The biomarkers ST2 and REG3α that measure damage to the GI tract crypts are combined in a single value termed the Mount Sinai Acute GVHD International Consortium (MAGIC) algorithm probability (MAP), which has emerged as a reliable indicator of disease activity and prognosis in acute GVHD.[17,18,20] Recently, we developed the MAGIC composite score (MCS), which integrates biomarker and clinical data to stratify the risk of GVHD at symptom onset and more accurately predicts response to therapy and NRM compared with clinical assessments alone (supplemental Table 1).[14] The MCS grading system used data from a single time point when GVHD was first treated. We undertook the current study to explore whether the combination of biomarkers and clinical symptoms in MCS would also be valuable in assessing treatment response. We used a similar approach to that used in developing MCS to define the MAGIC composite response (MCR). We again chose 6-month NRM as the key criterion and hypothesized that MCR criteria would demonstrate superior predictive accuracy compared with current clinical response only (CRO) approach.

Methods

Patient selection

Clinical data and serum samples were obtained from the MAGIC database and biorepository, which encompasses 26 HCT centers in North America, Europe, and Asia, using a prospective-specimen-collection, retrospective-blinded-evaluation (PRoBE) study design.[21,22]

We included both pediatric and adult patients who received their first HCT between 2014 and 2023, all of whom received systemic treatment for acute GVHD. We excluded patients who developed primary relapse of malignancy or who received donor lymphocyte infusion or second HCT before systemic GVHD treatment. Acute GVHD was diagnosed and staged according to the published criteria.[22] Minnesota risk, Manhattan risk, HCT-specific comorbidity index scores, intensity of conditioning regimens, and disease risk were classified as previously reported.[9,14,23,24] A complete response (CR) was defined as complete resolution of acute GVHD manifestations, and a partial response (PR) was defined as a decrease in at least 1 organ stage without worsening of other organs, provided that the improvement was less than a CR.[5] CRO included D28 CR or PR after systemic treatment, and overall response was defined by D28 CR or PR after systemic treatment without intervening secondary treatment. The training set excluded patients who died or relapsed before D28 of treatment (n = 6) for clarity in developing the MCR, but the validation set included them as nonresponders (deaths = 13, relapses = 7).

Serum samples

Serial serum samples were collected prospectively, cryopreserved, and shipped to a central laboratory. Serum concentrations of ST2 and REG3α were analyzed by enzyme-linked immunosorbent assays, as previously reported.[25,26] The MAP was calculated as a single value between 0.001 and 0.999 according to the following formula: $\log[-\log(1 - \text{MAP})] = -11.263 + 1.844(\log_{10}\text{ST2}) + 0.577(\log_{10}\text{REG3}\alpha)$.[12,27] We classified Ann Arbor (AA) scores using previously validated thresholds (AA1 < 0.141; $0.141 \leq$ AA2 < 0.291; AA3 ≥ 0.291).[11] D28 AA3 was considered high risk because previous works had revealed this threshold to create a risk separation of groups.[28,29]

Statistical analysis

The initiation of systemic treatment served as the starting point for all analyses. The primary end point was 6-month NRM, and outcomes were censored at 6 months. The cumulative incidence of 6-month NRM was estimated according to the Gray method, and relapse and second allogeneic HCT were considered as competing risks. We used the Kaplan-Meier method and the log-rank test to estimate and compare overall survival probabilities. We compared categorical variables using the Fisher exact test and continuous variables using the Mann-Whitney *U* test. We used the area under the receiver operating characteristic curve (AUROC) analysis and the DeLong test to compare the prognostic value of the different models. The ΔAUROC and its corresponding 95% confidence intervals (CIs) were calculated using a 1000 bootstrap method.[30] AUROC and negative predictive value (NPV) were used as the primary performance metric for comparisons. *P* values for NPV were calculated using the permutation test.[31] Time-dependent ROC analysis was performed as previously described.[32]

We analyzed the data as follows. First, we temporally divided patients into a training cohort that underwent transplant from 2014 to 2020 and a validation cohort that underwent transplant from 2021 to 2023. We deliberately chose a more recent validation set to reflect the current practices of GVHD prophylaxis, such as

Table 1. Patient characteristics (N = 1135)

Recipient age	Training (2014-2020) n = 826	Validation (2021-2023) n = 309	*P* values
Median age at HCT (range), y	54 (0-79)	57 (0-77)	.021
Recipient age, category (%)			
<18	111 (13.4)	44 (14.2)	.018
18-54	314 (38.0)	90 (29.1)	
≥55	401 (48.5)	175 (56.6)	
Recipient sex (%)			
Female	346 (41.9)	134 (43.4)	.703
Male	480 (58.1)	175 (56.6)	
Sex mismatch (%)			
Female to male	129 (15.6)	49 (15.9)	1
Other	696 (84.4)	260 (84.1)	
Race (%)			
White	691 (83.7)	204 (66.0)	<.001
Asian	38 (4.6)	13 (4.2)	
Black	44 (5.3)	11 (3.6)	
Others	4 (0.5)	4 (1.3)	
Unknown	49 (5.9)	77 (24.9)	
Primary disease (%)			
Acute leukemia	430 (52.1)	156 (50.5)	.829
MDS/MPN	231 (28.0)	95 (30.7)	
Malignant lymphoma	73 (8.8)	26 (8.4)	
Other	92 (11.1)	32 (10.4)	
HCT-CI (%)			
0-2	580 (70.2)	183 (59.2)	.001
≥3	246 (29.8)	126 (40.8)	
Conditioning (%)			
MAC (TBI < 8 Gy)	359 (43.5)	132 (42.7)	.794
MAC (TBI ≥ 8 Gy)	144 (17.4)	50 (16.2)	
RIC	323 (39.1)	127 (41.1)	
Donor type (%)			
HLA matched related	156 (18.9)	48 (15.5)	.004
HLA matched unrelated	447 (54.1)	188 (60.8)	
HLA mismatched related	5 (0.6)	1 (0.3)	
HLA mismatched unrelated donor	94 (11.4)	30 (9.7)	
Haploidentical	82 (9.9)	40 (12.9)	
Umbilical cord blood	42 (5.1)	2 (0.6)	
Donor source (%)			
Bone marrow	176 (21.3)	35 (11.3)	<.001
Peripheral blood	608 (73.6)	272 (88.0)	
Umbilical cord blood	42 (5.1)	2 (0.6)	
In vivo T-cell depletion (%)			
No	431 (52.2)	165 (53.4)	.765
Yes	395 (47.8)	144 (46.6)	
PTCy (%)			
No	691 (83.7)	248 (80.3)	.208
Yes	135 (16.3)	61 (19.7)	
Ex vivo T-cell depletion (%)			
No	803 (97.2)	292 (94.5)	.042
Yes	23 (2.8)	17 (5.5)	
GVHD grade at Tx			
I	229 (27.7)	91 (29.4)	.384
II	414 (50.1)	143 (46.3)	
III/IV	153 (18.5)	75 (24.3)	
MCS at Tx			
MCS 1	487 (59.0)	172 (55.7)	.059
MCS 2	286 (34.6)	104 (33.7)	
MCS 3	53 (6.4)	33 (10.7)	
Second-line treatment before D28	159 (19.3)	81 (26.2)	.011
Initial corticosteroid dose, median prednisolone (range), mg/kg	1.20 (0.5-4.0)	1.27 (0.5-3.74)	.332

HCT-CI, hematopoietic cell transplantation-specific comorbidity index; MAC, myeloablative conditioning; MDS/MPN, myelodysplastic syndromes/myeloproliferative neoplasms; PTCy, posttransplant cyclophosphamide; RIC, reduced intensity conditioning; TBI, total body irradiation; Tx, treatment.

increased use of posttransplant cyclophosphamide.[33] We, then, categorized patients on D28 of treatment according to MCS criteria.[14] We created a new category, MCS 0, for patients with a complete resolution of symptoms on D28 and low-risk biomarkers (AA1). Patients without symptoms but higher biomarker scores were assigned an MCS according to their 6-month NRM. Next, we used a classification and regression tree algorithm to create responder and nonresponder groups according to the incidence of 6-month NRM after treatment onset.[34] The criteria to separate groups included a maximum depth of 2 levels with a complexity parameter of 0.1.[34]

All *P* values were two-tailed, with a threshold of <.05 indicating statistical significance. All statistical analyses were performed with R version 4.3.2 (R Foundation) and EZR version 1.63.[35]

Patients were prospectively monitored for acute GVHD symptoms for 100 days according to institutional review board–approved protocols. Informed consent was obtained from all participants in accordance with the Declaration of Helsinki.

Results

Patient characteristics

We divided 1135 patients who fulfilled all the inclusion criteria into a training cohort (n = 826) that underwent transplant from 2014 to 2020 and a validation cohort (n = 309) that underwent transplant from 2021 to 2023 (Table 1). Patients who died or relapsed before D28 of treatment (n = 6) were excluded from the training set used to develop the MCR but were included as nonresponders (n = 20) in the validation set. The population of the training set differed significantly from those in the validation set in that they were older, had a higher HCT comorbidity index,

Table 2. Development of MCR

Training set			
At D28	**6-Month NRM (95% CI)**	**MCR***	**n (%)**
Grade 0 + AA1	3.7% (2.1-6.0)	Response	353 (42.7)
Grade 0 + AA2	13.3% (8.0-20.1)	Response	120 (14.5)
Grade 0 + AA3	28.8% (17.2-41.6)	Nonresponse	52 (6.3)
MCS 1	4.6% (1.9-9.2)	Response	130 (15.7)
MCS 2	31.7% (23.7-40.0)	Nonresponse	123 (14.9)
MCS 3	70.8% (55.4-81.8)	Nonresponse	48 (5.8)
Validation set			
At D28	**6-Month NRM (95% CI)**	**MCR***	**n (%)**
Grade 0 + AA1	3.6% (1.2-8.3)	Response	121 (39.2)
Grade 0 + AA2	18.5% (7.3-33.6)	Response	34 (11.0)
Grade 0 + AA3	35.4% (12.0-60.1)	Nonresponse	15 (4.9)
MCS 1	7.2% (2.3-16.0)	Response	59 (19.1)
MCS 2	41.8% (26.3-56.5)	Nonresponse	43 (14.9)
MCS 3	49.0% (22.6-71.1)	Nonresponse	17 (5.9)

*Categorized by classification and regression tree analyses based on 6-month NRM.

included more haploidentical and fewer cord blood donors, and used less ex vivo T-cell depletion, all of which reflect evolving trends in HCT. The median initial dose of prednisolone was approximately 1.2 mg/kg per day, and approximately 20% of patients received second-line treatment before D28 of treatment. Approximately 28% of patients with grade 1 GVHD received systemic treatment with at least 0.25 mg/kg prednisone in both cohorts. The median follow-up of survivors after treatment initiation was 23 and 17 months in the training and validation cohorts, respectively.

MAGIC composite scores at D28

We categorized patients at D28 of treatment in the training set by both clinical severity and biomarker scores according to previous MCS criteria.[14] Patients with complete resolution of their symptoms on D28 and who thus were not included in the original MCS groups were categorized as clinical grade 0 (n = 525, 63.6%) and divided into subgroups by D28 serum biomarker AA scores. We classified patients with clinical grade 0 and low-risk biomarkers, AA1, as MCS 0 to identify patients who have both resolved all clinical symptoms and possess low biomarker scores at the time of response assessment. MCS 0 patients experienced very low 6-month NRM (3.7%). Patients with clinical grade 0 and intermediate-risk biomarkers, AA2, had higher NRM (13.3%) and were classified as MCS 1; patients with clinical grade 0 and high-risk biomarkers, AA3, experienced yet higher NRM (28.8%) and were classified as MCS 2 (Table 2). The cumulative incidences of NRM of these 4 groups are found in Figure 1. Active GVHD caused a very small percentage of deaths in patients with MCS 0 and MCS 1 but most deaths in patients with MCS 2 and MCS 3 (supplemental Table 2).

MCR

When we applied the classification and regression tree algorithm to these 4 groups to generate MCR, it categorized them as responders or nonresponders according to 6-month NRM (Table 2). Both MCS 0 (highly favorable) and MCS 1 (favorable) groups were categorized as responders in the training and validation sets with NRM of ~5% to 15%, whereas MCS 2 and MCS 3 (unfavorable) experienced high NRM of >30% (Table 2). MCS 2

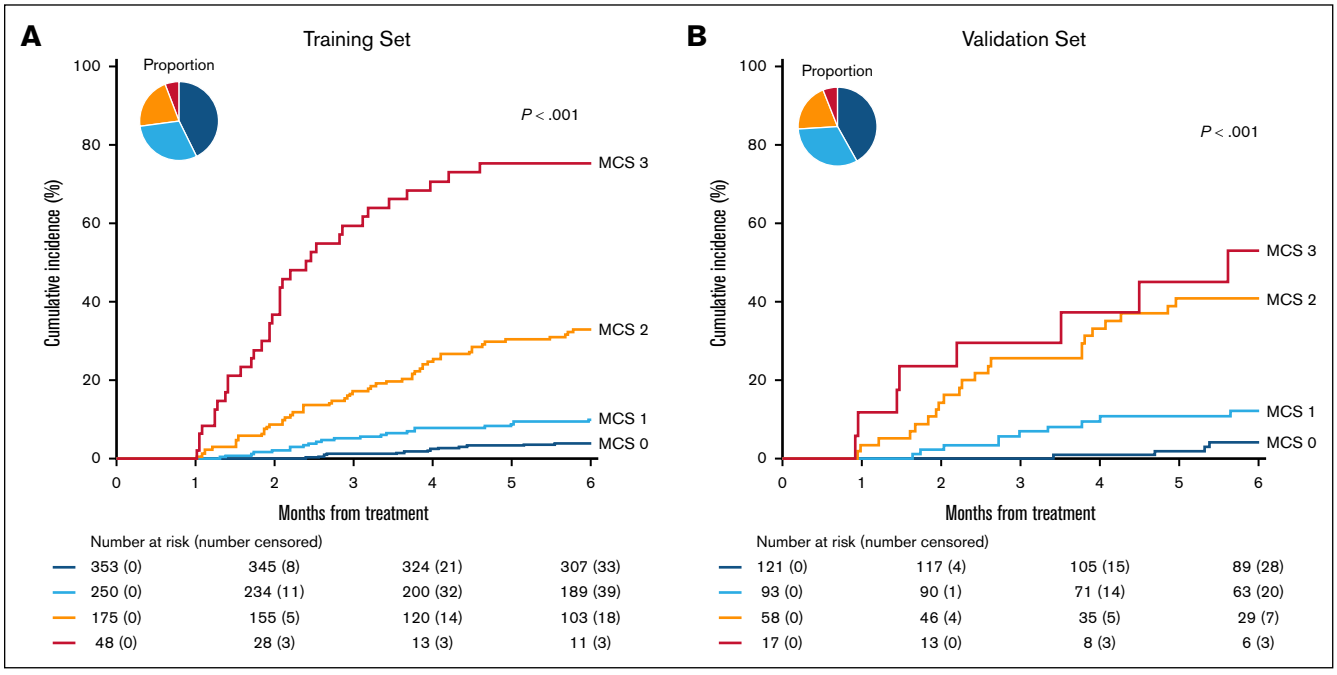

Figure 1. Cumulative incidence of NRM at 6 months by MCS at D28. (A) Training set: MCS 0 (dark blue): 3.7% (95% CI, 2.1%-6.0%), MCS 1 (light blue): 8.8% (95% CI, 5.7%-12.7%), MCS 2 (orange): 30.9% (95% CI, 24.2%-37.8%), MCS 3 (red): 70.8% (95% CI, 55.4%-81.8%). (B) Validation set: MCS 0: 3.6% (95% CI, 1.2%-8.3%), MCS 1: 11.3% (95% CI, 5.7%-18.9%), MCS 2: 40.1% (95% CI, 27.0%-52.9%), MCS 3: 49.0% (95% CI, 22.6%-71.1%).

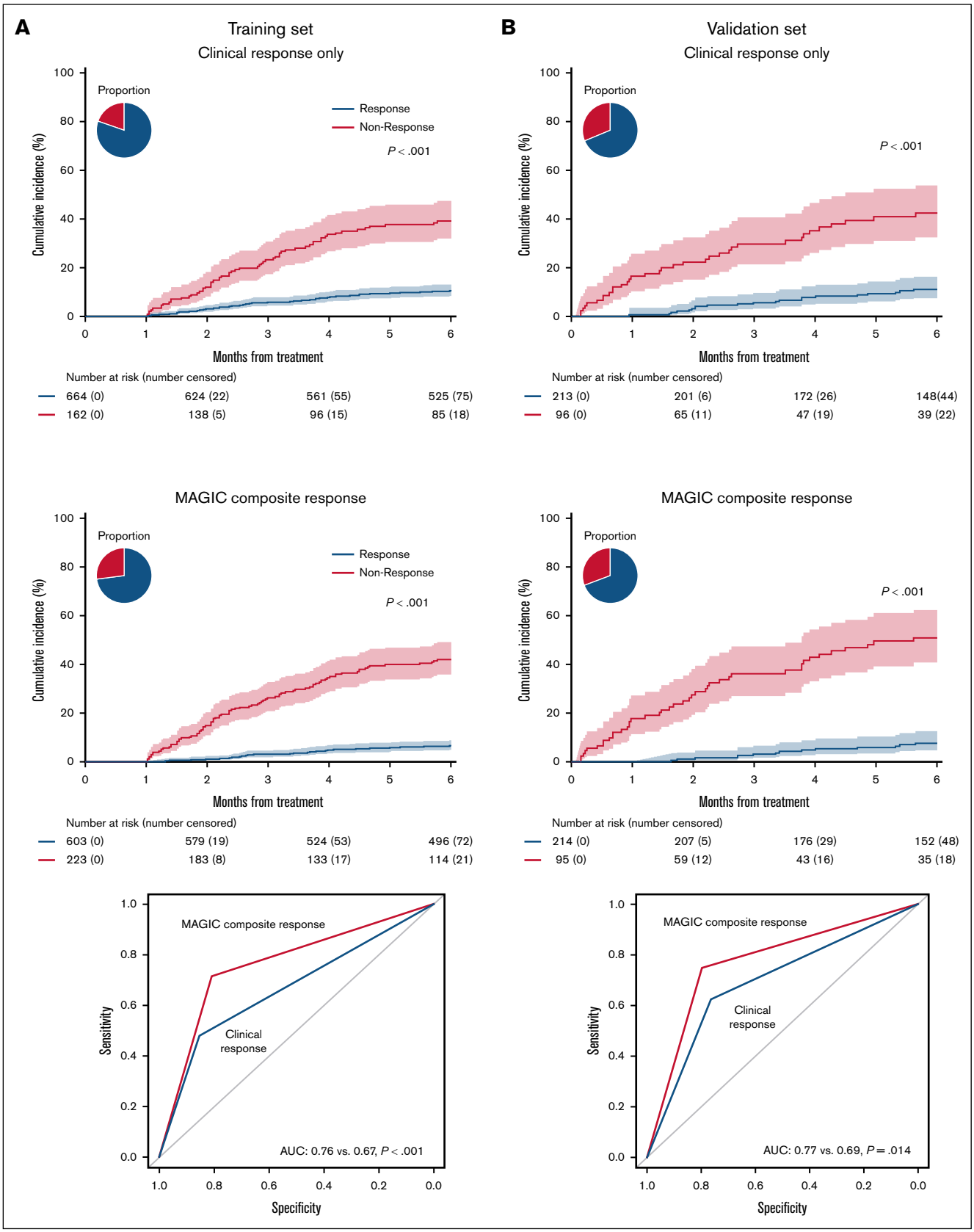

Figure 2. Cumulative incidence of NRM at 6 months based on clinical and MAGIC response criteria in the training and validation sets. (A) Training set. Left: clinical responders: 9.7% (95% CI, 7.6%-12.0%), clinical nonresponders: 36.4% (95% CI, 29.0%-43.8%). Middle: MAGIC composite responders: 5.8% (95% CI, 4.1%-7.9%),

Table 3. Predictive performances for 6-month NRM

Response criteria	Response rates	Sensitivity	Specificity	Balanced accuracy*	PPV	NPV	*P* values for NPV	AUC	*P* values for AUC	ΔAUC (95% CI)
Training set										
Clinical response	80.4%	48.0%	85.3%	66.7%	36.4%	90.3%	<.001	0.67	<.001	0.096 (0.053-0.138)
MCR	73.1%	71.5%	80.8%	76.2%	39.5%	94.2%		0.76		
Validation set										
Clinical response	68.9%	61.9%	76.4%	69.2%	37.7%	89.7%	.017	0.69	.014	0.078 (0.014-0.143)
MCR	69.3%	74.4%	79.5%	77.0%	45.8%	93.1%		0.77		

AUC, area under the curve; PPV, positive predictive value.
*Balanced accuracy is the average of sensitivity and specificity, calculated as their sum divided by 2.

and MCS 3 patients were combined because of their poor outcomes and the very small number of patients in the MCS 3 group.

The MCR algorithm identified 2 important subgroups whose 6-month NRM did not completely align with a change in clinical symptoms. In the first group, patients with complete resolution of their symptoms (grade 0) and high biomarkers (AA3) at D28 of treatment, classified as MCS 2, experienced a 6-month NRM of 28.8% and were classified as nonresponders. In the second group, patients with persistence of mild symptoms (see subsequent discussion) and who were categorized as MCS 1 after D28 of treatment experienced low 6-month NRM (5%-7%) and were therefore classified as responders (Table 2).

Comparison of MCR to clinical response only

We, then, compared the 6-month NRM of responders and nonresponders predicted by the MCR algorithm to that predicted by the CRO system. The proportion of patients categorized as responders in both systems was similar (68.9% vs 69.3%). Compared with CRO, MCR produced a greater separation in 6-month NRM between groups in both the training set (33.7% vs 26.7%) and the validation set (38.9% vs 27.5%) (Figure 2, left and middle panels). This improvement resulted in a significantly greater AUROC in both the training set (0.76 vs 0.67, $P < .001$) and the validation set (0.77 vs 0.69, $P = .014$) (Figure 2, right panels). Importantly, MCR also significantly increased the NPV, which in these systems corresponds to survival (Table 3). We did not include initiation of second-line treatment in the initial calculation of either CRO or MCR because subjective considerations unrelated to treatment response are common in observational studies where treatment parameters are not clearly prespecified. In the training set, 144 of 159 patients (91%) had at least grade 2 GVHD when they received second-line therapy before D28. When initiation of second-line therapy was considered a nonresponse, as is customary in clinical trials, MCR remained superior to CRO in predicting long-term outcomes (Table 4). There was no difference in relapse between response groups according to either system, and thus MCR produced a larger separation in overall survival between groups (supplemental Figures 1 and 2).

To understand why the MCR criteria predicted long-term NRM better than the CRO criteria, we evaluated the reclassification of CRO groups by MCR. In both data sets, the MCR criteria identified 15% of CRO responders with a fivefold greater incidence of NRM compared with nonresponders and who were therefore classified as MCR nonresponders (training: 30.7% vs 5.9%, $P < .001$; validation: 34.3% vs 6.8%, $P < .001$) (Figure 3A). In the validation set, 26 of these 28 reclassified patients (93%) had a high MAP at D28 of treatment (supplemental Table 3). Conversely, MCR identified nearly one-third of CRO nonresponders with sevenfold less NRM and who were therefore classified as MCR responders (training: 5.0% vs 46.7%, $P < .001$; validation: 7.6% vs 50.7%, $P < .001$). All 29 of these reclassified patients in the validation set had a low MAP at D28 of treatment (supplemental Table 4). Further analyses revealed that most MCS 1 patients (80%-85%) who were nonresponders by clinical criteria had unresolved mild skin and/or upper GI symptoms (supplemental Table 5). These observations help to explain how MCR improves sensitivity with minimal loss of specificity in predicting 6-month NRM (Figure 2) and highlight the utility of biomarkers in the improvement of long-term GVHD control. A time-dependent analysis of AUROC for NRM at time points between 2 and 12 months confirmed the consistent superiority of MCR criteria at all time points (supplemental Figure 3). The difference between groups of failure-free survival, defined as survival without subsequent lines of systemic therapy, recurrent malignancy, or death, was also greater based on MCR than CRO (supplemental Figure 4).

We used the entire data set to also compare MCR with CRO in several subsets that mirror potential clinical trial designs for clinically significant GVHD where a strong correlation between early response and long-term outcomes is crucial to the assessment of the intervention. Because biomarker values are not always known at the onset of treatment, we analyzed 3 common treatment scenarios in which biomarkers were not included as inclusion criteria at the initiation of treatment: grades 2 to 4 GVHD, grade 3 of 4 GVHD, and Minnesota high-risk GVHD. MCR correctly identified more nonresponders compared with CRO in all 3 scenarios resulting in significantly improved correlation between response and survival as evidenced by significantly higher NPVs (supplemental Table 6). Thus, inclusion of biomarkers at D28 significantly improved the ability of the response metric to predict long-term outcomes of GVHD even when biomarkers were not

Figure 2 (continued) MAGIC composite nonresponders: 39.5% (95% CI, 33.0%-45.8%). Right: area under the curve (AUC) by receiver operating characteristic for 6-month NRM. (B) Validation set. Left: clinical responders: 10.3% (95% CI, 6.6%-15.0%), clinical nonresponders: 37.7% (95% CI, 27.8%-47.6%). Middle: MAGIC composite responders: 6.9% (95% CI, 4.0%-11.0%), MAGIC composite nonresponders: 45.8% (95% CI, 35.3%-55.8%). Right: AUC by receiver operating characteristic for 6-month NRM. Pie charts depict percentage of responders/nonresponders.

Table 4. Predictive performances for 6-month NRM treating second line as a nonresponse

Response criteria	Response rates	Sensitivity	Specificity	Balanced accuracy*	PPV	NPV	*P* values for NPV	AUC	*P* values for AUC	ΔAUC (95% CI)
Training set†										
Overall response	73.2%	57.7%	78.6%	68.2%	32.1%	91.4%	.001	0.68	<.001	0.065 (0.028-0.105)
MCR	68.0%	74.0%	75.4%	74.7%	34.5%	94.3%		0.74		
Validation set‡										
Overall response	58.3%	71.0%	62.8%	66.9%	31.8%	90.6%	.095	0.69	.056	0.054 (−0.001 to 0.102)
MCR	59.2%	77.9%	69.1%	73.5%	36.1%	93.0%		0.74		

AUC, area under the curve; PPV, positive predictive value.
*Balanced accuracy is the average of sensitivity and specificity, calculated as their sum divided by 2.
†In the training set, 59 (7.1%) and 41 (5.0%) were reclassified by the second-line treatment when using clinical response and MCR, respectively.
‡In the validation set, 33 (10.7%) and 31 (10.0%) were reclassified by the second-line treatment when using clinical response and MCR, respectively.

available or known at the onset of treatment. As expected, analysis of patients with MCS 2/3 GVHD also revealed statistically significant improvement in AUROC and NPV (supplemental Table 6).

Previously, we have revealed that MAP biomarkers on D28 better predicted 6-month NRM than CRO.[28] Comparison of MCR status to the MAP at D28 revealed that MCR was superior in both key metrics of NPV and area under the curve (supplemental Table 7). This finding confirmed the superiority of a composite metric that includes both clinical and biomarker values in predicting long-term outcomes.

Discussion

The current standard for response in clinical trials of acute GVHD, the reduction of GVHD clinical symptoms on D28 of treatment, modestly correlates with long-term outcomes such as 6-month NRM. We evaluated whether MCR, which integrates serum biomarkers with clinical symptoms on D28, could redefine responses to improve predictive accuracy as has recently been found with MCS at the initiation of treatment.[14] The MCR system was superior to the change in clinical symptoms alone in predicting 6-month outcomes as determined by key metrics such as AUROC and NPV even when biomarkers were available only on D28 and not at the beginning of treatment. Importantly, the MCR system includes patients as responders with low biomarkers and mild, but stable symptoms, who have favorable outcomes, even though they would be classified as nonresponders by clinical response criteria alone. Thus, a composite status integrating clinical symptoms and biomarkers after D28 of treatment better predicts long-term outcomes than conventional criteria that rely only on changes in clinical symptom severity.

This study suggests that the end points of clinical trials of GVHD treatment can be improved. Severe acute GVHD can be lethal, and the severity of clinical symptoms at onset does not correlate well with treatment response or eventual mortality.[13,14] Recent advances that include the use of serum biomarkers can define groups at the initiation of treatment with very different risks of treatment response and NRM.[14] But, currently, we use the same definition of response as the end point for both high-risk and low-risk GVHD, despite the differing treatment goals for each group. In high-risk trials,[36] the primary objective is to increase the response rate, with the expectation that this will lead to reduced mortality. Thus, the end point for trials treating high-risk disease should measure a response that more accurately correlates with survival. However, the current definitions of response include changes in clinical symptoms that do not differentiate survival outcomes from nonresponding patients (eg, patients whose lower GI GVHD improves from stage IV to II have similarly high NRM as patients whose lower GI GVHD remains in stage IV).[37] The MCR with its integration of symptoms and serum biomarkers predicts survival of responders (as measured by NPV) better than CRO and could be useful in such trials. In contrast, treatment of patients with low-risk disease, who have low mortality and high survival rates, should minimize treatment toxicity but maintain a high response rate. The response end point for low-risk trials[38] could include reduction in toxicities such as severe infections[39,40] or repeat hospitalizations provided these responses are associated with favorable long-term outcomes (eg, survival). Favorable biomarkers in patients with mild, stable disease could support their inclusion as "responders" if they meet the other salient end point criteria.

This retrospective study has several strengths. First, the analysis was performed using a large data set that enabled the development of the system in a training cohort and its validation in a more recent cohort, which better reflects current practice. Second, the observational data derive from 2 dozen HCT centers, which represent real-world evidence and increase the likelihood of their broad applicability. Third, the data and samples were collected prospectively and evaluated using a rigorous PRoBE study design.[21,22] Fourth, serum biomarkers were evaluated in a single central laboratory with rigorous quality controls, ensuring accuracy and consistency.

This study also has limitations. First, although the data were prospectively collected using a PRoBE design, such real-world evidence admits to greater heterogeneity of the patient populations than when they are subjected to inclusion and exclusion criteria. Second, the MCR model was based on observational data that do not specify which changes in therapy qualify as a nonresponse to treatment. Third, the availability of newly approved second-line therapies could also influence treatment decisions in clinical trials. Thus, to be useful as a trial end point, MCR status should be formally validated with data from clinical trial patients who were subject to clear inclusion and exclusion criteria and prespecified treatment plans. Third, the existence of differences in baseline characteristics between training and validation sets might have

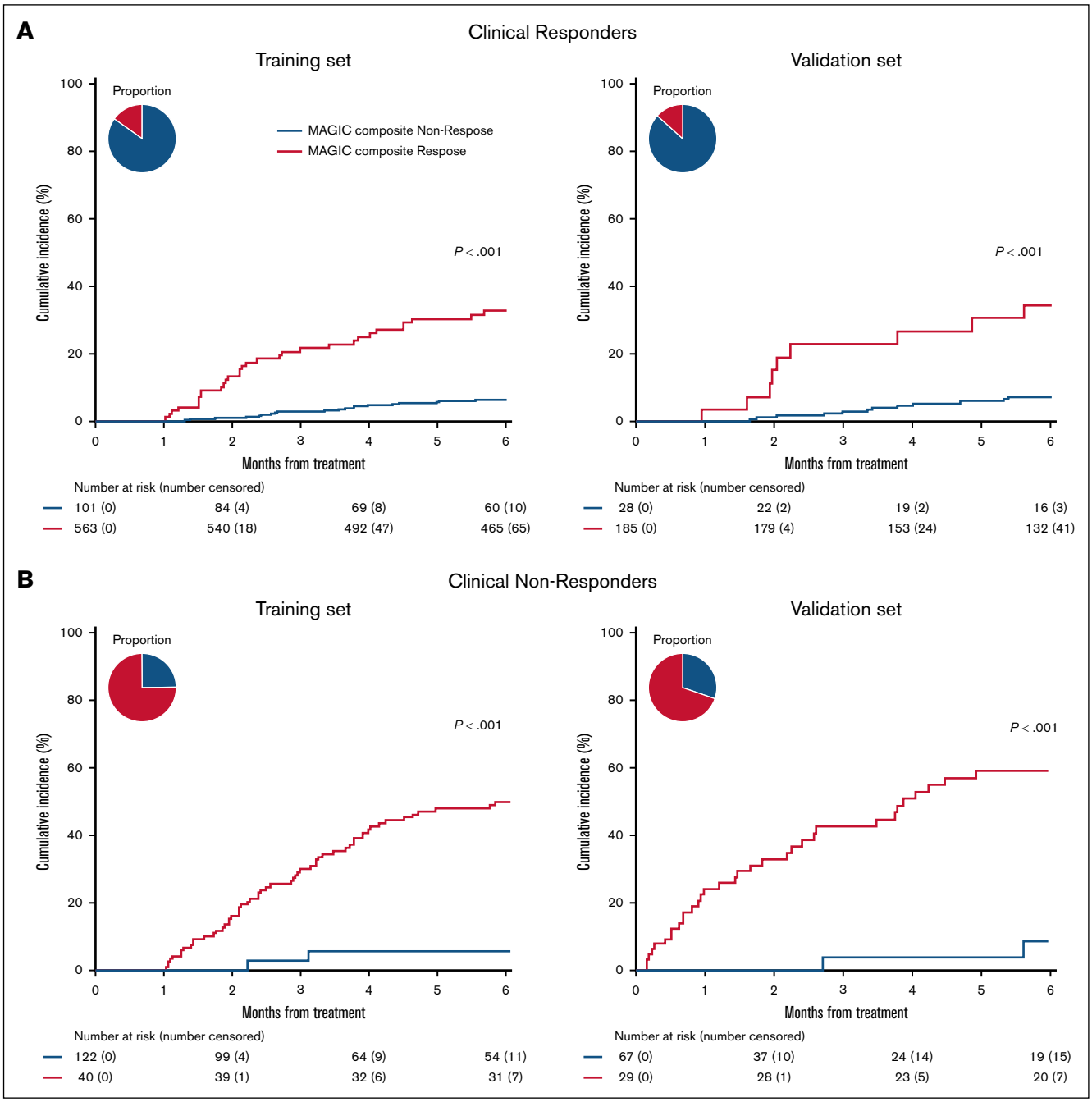

Figure 3. NRM of MCR reclassification of clinical responders and nonresponders. The 6-month cumulative incidence of NRM by clinical response reclassified by MCR criteria. (A) Clinical responders reclassified by MCR. Left: training set MAGIC composite responders: 5.9% (95% CI, 4.1%-8.0%), MAGIC composite nonresponders: 30.7% (CI, 22.0%-39.8%). Right: validation set MAGIC composite responders: 6.8% (95% CI, 3.7%-11.2%), MAGIC composite nonresponders: 34.3% (95% CI, 16.9%-52.6%). (B) Clinical nonresponders reclassified by MCR. Left: training set MAGIC composite responders: 5.0% (95% CI, 0.9%-15.0%), MAGIC composite nonresponders: 46.7% (95% CI, 37.6%-55.3%). Right: validation set MAGIC composite responders: 7.6% (95% CI, 1.3%-22.0%), MAGIC composite nonresponders: 50.7% (95% CI, 37.8%-62.2%). Pie charts depict percentage of responders/nonresponders.

biased the generation of the model, again underscoring the need to validate it using clinical trial data.

This study demonstrates that the combined clinical/biomarker status of a patient after D28 of treatment better predicts long-term outcomes than a change in clinical symptoms. This superiority depends in large part on the reclassification of patients as responders whose initially mild symptoms persisted for 4 weeks but whose biomarkers were low at D28. This change addresses a longstanding anomaly in GVHD clinical trials regarding non-responding patients who nevertheless have very favorable long-term outcomes. But it must be acknowledged that a more

accurate prediction of long-term outcomes at D28 is not especially useful in guiding therapeutic decisions which usually occur much earlier. Future studies should determine whether such a composite of clinical symptoms and biomarkers at earlier time points, such as after 1 or 2 weeks of treatment, can also robustly predict long-term outcomes[41]; if so, clinical trials of GVHD could reduce the length of time when patients are subjected to experimental therapies and allow for earlier initiation of next-line therapies.

Acknowledgments

The authors greatly appreciate the patients, their families, medical staff, and data managers in the MAGIC centers.

Y.A. is a recipient of the Japan Society for the Promotion of Science Postdoctoral Fellowship for Research Abroad program. This work was supported by National Institutes of Health, National Cancer Institute grants P01 CA039542 and P30 CA196521, the Pediatric Cancer Foundation, the Gillespie and Solomon-Gillespie GVHD Research Fund, and German Jose Carreras Leukemia Foundation grants DJCLS 01 GVHD 2016 and DJCLS 01 GVHD 2020.

Authorship

Contribution: Y.A. designed the study, collected the clinical data, conducted the statistical analysis, and wrote the manuscript; J.P. collected the clinical data and performed statistical analysis; N.K., I.E.L., and N.S. collected the clinical data, advised statistical methods, and reviewed and revised the manuscript; P.A.-H., F.A., C.C., H.K.C., M.E., A.M.E., E.O.H., C.L.K., S. Kraus, P.M., T.O., I.P., M.Q., R.R., T.S., J.M., E.U., I.V., D.W., M.W., R.Z., Z.D., W.J.H., and R.N. collected the clinical data and reviewed and revised the manuscript; J.B., G.E., S.G., and R.Y. collected and reviewed the clinical data; S. Kowalyk and G.M. performed the laboratory analysis; J.E.L. and J.L.M.F. designed the study, interpreted the data, advised the methods, reviewed and revised the manuscript, and organized this project; and all authors contributed to the writing of the report and approved the final version of the article.

Conflict of interest disclosure: Y.A. reports honoraria from Novartis, AstraZeneca, and Takeda. P.A.-H. reports consulting or advisory fees from Incyte. F.A. reports honoraria from Kite/Gilead, Novartis, Bristol Myers Squibb, Takeda, Medac, Miltenyi Biomedicine, AbbVie, Mallinckrodt/Therakos, and Janssen; and research funding from Mallinckrodt/Therakos and Neovii. H.K.C. reports consulting or advisory fees from Actinium, Incyte, Ironwood, Orca Bio, RegImmune, and Sanofi; research funding from GlaxoSmithKline and Incyte; and honoraria from MJH Holding, Physician Education Resource, and Plexus Communications. E.O.H. reports advisory, committee, or consulting roles for AbbVie, Disc Medicine, PharmaEssentia, Blueprint Medicines, Cabaletta Bio, and the American Board of Internal Medicine; and research funding from AbbVie, Blueprint Medicines, Disc Medicine, Kartos Therapeutics, and Kite Pharma. C.L.K. reports honoraria from Sanofi, Incyte, and Mesoblast; and consulting fees from CSL Behring, Alexion, and Physician Education Resource. P.M. reports consulting fee from Miltenyi, Amgen, and Pierre Fabre. T.O. reports consulting agreements with bluebird bio, Pfizer, Medexus, Alexion, and Elixirgen Therapeutics. R.R. reports consulting or advisory role with Allogene, Gilead Sciences, Bayer, Incyte, TScan, Orca Bio, Pierre Fabre Pharmaceuticals, CareDx, Quell Biotherapeutics, Sana Biotechnology, Sail Biomedicines, and Autolus; and research funding from Atara Biotherapeutics, Incyte, Sanofi, Immatics, AbbVie, Takeda, Gilead Sciences, CareDx, TScan, Cabaletta, Synthekine, Bristol Myers Squibb, Johnson & Johnson, Allogene, Genentech, Vittoria Therapeutics, AstraZeneca, and Imugene. Z.D. reports research support from Incyte, Corp, RegImmune, Corp, Taiho Oncology, Inc, and Kura Oncology, Inc; and consulting fees from Sanofi, Incyte, Corp, Inhibrx, RegImmune, Corp, MaaT Pharma, Forte Biosciences Inc, Medexus Pharmaceuticals, Inc, and Mesoblast Ltd. R.N. reports consultant feeds from Sanofi, MaaT Pharma, and Ono Pharmaceutical; and research support from Miyarisan. J.E.L. reports research support from Equillium, Incyte, MaaT Pharma, and Mesoblast; and consulting fees from Editas, Equillium, Kamada, Mesoblast, Sanofi, bluebird bio, Inhibrx, and X4 Pharmaceuticals. J.L.M.F. reports research support from Equillium, Incyte, MaaT Pharma, and Mesoblast; and consulting fees from Editas, Equillium, Kamada, Mesoblast, Alexion, Realta, Medpace, Viracor, AlloVir, and Physician Education Resource. The remaining authors declare no competing financial interests.

ORCID profiles: Y.A., 0000-0001-6825-9340; J.P., 0009-0008-6236-4471; N.K., 0009-0004-7955-1119; I.E.L., 0009-0008-2716-5792; P.A.-H., 0000-0002-0196-806X; E.O.H., 0000-0002-1125-4060; P.M., 0000-0001-6426-4046; T.S.O., 0000-0003-1288-1960; M.Q., 0000-0001-7689-343X; R.R., 0000-0003-2185-9546; E.U., 0000-0001-8530-1192; Z.D., 0000-0002-7994-8974; W.J.H., 0000-0002-5841-4105; R.N., 0000-0002-9082-0680; J.E.L., 0000-0002-5611-7828.

Correspondence: John E. Levine, The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, Box 1410, New York, NY 10029; email: john.levine@mssm.edu; and James L. M. Ferrara, The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029; email: james.ferrara@mssm.edu.

References

1. Ferrara JL, Levine JE, Reddy P, Holler E. Graft-versus-host disease. *Lancet*. 2009;373(9674):1550-1561.
2. Zeiser R. Advances in understanding the pathogenesis of graft-versus-host disease. *Br J Haematol*. 2019;187(5):563-572.
3. Levine JE, Logan B, Wu J, et al. Graft-versus-host disease treatment: predictors of survival. *Biol Blood Marrow Transpl*. 2010;16(12):1693-1699.
4. Martin PJ, Bachier CR, Klingemann HG, et al. Endpoints for clinical trials testing treatment of acute graft-versus-host disease: a joint statement. *Biol Blood Marrow Transpl*. 2009;15(7):777-784.
5. MacMillan ML, DeFor TE, Weisdorf DJ. The best endpoint for acute GVHD treatment trials. *Blood*. 2010;115(26):5412-5417.

6. Saliba RM, Couriel DR, Giralt S, et al. Prognostic value of response after upfront therapy for acute GVHD. *Bone Marrow Transpl*. 2012;47(1):125-131.
7. Inamoto Y, Martin PJ, Storer BE, Mielcarek M, Storb RF, Carpenter PA. Response endpoints and failure-free survival after initial treatment for acute graft-versus-host disease. *Haematologica*. 2014;99(2):385-391.
8. Harris AC, Ferrara JL, Braun TM, et al. Plasma biomarkers of lower gastrointestinal and liver acute GVHD. *Blood*. 2012;119(12):2960-2963.
9. 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 Transpl*. 2015;21(4):761-767.
10. Paczesny S, Krijanovski OI, Braun TM, et al. A biomarker panel for acute graft-versus-host disease. *Blood*. 2009;113(2):273-278.
11. Akahoshi Y, Spyrou N, Hogan WJ, et al. Incidence, clinical presentation, risk factors, outcomes, and biomarkers in de novo late acute GVHD. *Blood Adv*. 2023;7(16):4479-4491.
12. Hartwell MJ, Özbek U, Holler E, et al. An early-biomarker algorithm predicts lethal graft-versus-host disease and survival. *JCI Insight*. 2018;3(16):e124015.
13. Spyrou N, Akahoshi Y, Ayuk F, et al. The utility of biomarkers in acute GVHD prognostication. *Blood Adv*. 2023;7(17):5152-5155.
14. Akahoshi Y, Spyrou N, Weber D, et al. Novel MAGIC composite scores using both clinical symptoms and biomarkers best predict treatment outcomes of acute GVHD. *Blood*. 2024;144(9):1010-1021.
15. 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.
16. Etra A, Gergoudis S, Morales G, et al. Assessment of systemic and gastrointestinal tissue damage biomarkers for GVHD risk stratification. *Blood Adv*. 2022;6(12):3707-3715.
17. 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.
18. Ferrara JL, 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.
19. Katsivelos N, Spyrou N, Weber D, et al. Serial clinical and biomarker monitoring during graft-versus-host disease treatment identifies distinct risk strata including an ultra-low risk group. *Transpl Cell Ther*. 2025;31(1):10.e1-10.e9.
20. Akahoshi Y, Spyrou N, Hoepting M, et al. Flares of acute graft-versus-host disease: a Mount Sinai Acute GVHD International Consortium analysis. *Blood Adv*. 2024;8(8):2047-2057.
21. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst*. 2008;100(20):1432-1438.
22. Harris AC, Young R, Devine S, et al. International, multicenter standardization of acute graft-versus-host disease clinical data collection: a report from the Mount Sinai Acute GVHD International Consortium. *Biol Blood Marrow Transpl*. 2016;22(1):4-10.
23. Bacigalupo A, Ballen K, Rizzo D, et al. Defining the intensity of conditioning regimens: working definitions. *Biol Blood Marrow Transpl*. 2009;15(12):1628-1633.
24. Sorror ML, Storer B, Storb RF. Validation of the hematopoietic cell transplantation-specific comorbidity index (HCT-CI) in single and multiple institutions: limitations and inferences. *Biol Blood Marrow Transpl*. 2009;15(6):757-758.
25. Zhang J, Ramadan AM, Griesenauer B, et al. ST2 blockade reduces sST2-producing T cells while maintaining protective mST2-expressing T cells during graft-versus-host disease. *Sci Transl Med*. 2015;7(308):308ra160.
26. Zhao D, Kim YH, 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.
27. Rios CAO, Qayed M, Etra AM, et al. Differences in acute graft-versus-host disease (GVHD) severity and its outcomes between black and white patients. *Transpl Cell Ther*. 2024;30(11):1061-1061.e10.
28. Srinagesh HK, Özbek U, Kapoor U, et al. The MAGIC algorithm probability is a validated response biomarker of treatment of acute graft-versus-host disease. *Blood Adv*. 2019;3(23):4034-4042.
29. Qayed M, Kapoor U, Gillespie S, et al. A validated risk stratification that incorporates MAGIC biomarkers predicts long-term outcomes in pediatric patients with acute GVHD. *Transpl Cell Ther*. 2024;30(6):603.e1-603.e11.
30. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. 1st ed. Taylor & Francis; 1994.
31. Park C, Park SY, Kim HJ, Shin HJ. Statistical methods for comparing predictive values in medical diagnosis. *Korean J Radiol*. 2024;25(7):656-661.
32. Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*. 2013;32(30):5381-5397.
33. Bolaños-Meade J, Hamadani M, Wu J, et al. Post-transplantation cyclophosphamide-based graft-versus-host disease prophylaxis. *N Engl J Med*. 2023;388(25):2338-2348.
34. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. 1st ed. Taylor & Francis; 1984.
35. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transpl*. 2013;48(3):452-458.

36. Al Malki MM, London K, Baez J, et al. Phase 2 study of natalizumab plus standard corticosteroid treatment for high-risk acute graft-versus-host disease. *Blood Adv*. 2023;7(17):5189-5198.

37. Akahoshi Y, Inamoto Y, Spyrou N, et al. Refinement of day 28 treatment response criteria for acute GVHD: a collaboration study of the JSTCT and MAGIC. *Blood Adv*. Published online 1 July 2025. https://doi.org/10.1182/bloodadvances.2025016233

38. Etra A, Capellini A, Alousi A, et al. Effective treatment of low-risk acute GVHD with itacitinib monotherapy. *Blood*. 2023;141(5):481-489.

39. Akahoshi Y, Kimura SI, Tada Y, et al. Cytomegalovirus gastroenteritis in patients with acute graft-versus-host disease. *Blood Adv*. 2022;6(2):574-584.

40. Akahoshi Y, Kimura SI, Inamoto Y, et al. Effect of cytomegalovirus reactivation with or without acute graft-versus-host disease on the risk of nonrelapse mortality. *Clin Infect Dis*. 2021;73(3):e620-e628.

41. Spyrou N, Akahoshi Y, Kowalyk S, et al. A day 14 endpoint for acute GVHD clinical trials. *Transpl Cell Ther*. 2024;30(4):421-432.