REGULAR ARTICLE

blood advances

Amphiregulin, ST2, and REG3α biomarker risk algorithms as predictors of nonrelapse mortality in patients with acute GVHD

Aaron Etra,[1] Najla El Jurdi,[2] Nikolaos Katsivelos,[1] Deukwoo Kwon,[3] Stephanie Gergoudis,[1] George Morales,[1] Nikolaos Spyrou,[1] Steven Kowalyk,[1] Paibel Aguayo-Hiraldo,[4] Yu Akahoshi,[1] Francis Ayuk,[5] Janna Baez,[1] Brian C. Betts,[2] Chantiya Chanswangphuwana,[6] Yi-Bin Chen,[7] Hannah Choe,[8] Zachariah DeFilipp,[7] Sigrun Gleich,[9] Elizabeth Hexner,[10] William J. Hogan,[11] Ernst Holler,[9] Carrie L. Kitko,[12] Sabrina Kraus,[13] Monzr Al Malki,[14] Margaret MacMillan,[2] Attaphol Pawarode,[15] Francesco Quagliarella,[16] Muna Qayed,[17] Ran Reshef,[18] Tal Schechter,[19] Ingrid Vasova,[20] Daniel Weisdorf,[2] Matthias Wölfl,[21] Rachel Young,[1] Ryotaro Nakamura,[14] James L. M. Ferrara,[1,*] John E. Levine,[1,*] and Shernan Holtan[2,*]

[1]The Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY; [2]Hematology, Oncology and Transplant, University of Minnesota, Minneapolis, MN; [3]Department of Population Health Science and Policy, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY; [4]Division of Hematology, Oncology, and Blood and Marrow Transplantation, Children's Hospital Los Angeles, Los Angeles, CA; [5]Department of Stem Cell Transplantation, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [6]Blood and Marrow Transplantation Program, Chulalongkorn University, Bangkok, Thailand; [7]Hematopoietic Cell Transplant and Cellular Therapy Program, Massachusetts General Hospital, Boston, MA; [8]Division of Hematology, James Cancer Center, The Ohio State University, Columbus, OH; [9]Department of Hematology and Oncology, Internal Medicine III, University of Regensburg, Regensburg, Germany; [10]Blood and Marrow Transplantation Program, Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; [11]Division of Hematology, Mayo Clinic, Rochester, MN; [12]Pediatric Stem Cell Transplant Program, Vanderbilt University Medical Center, Nashville, TN; [13]Department of Internal Medicine II, University Hospital of Würzburg, Würzburg, Germany; [14]Hematology/Hematopoietic Cell Transplant, City of Hope National Medical Center, Duarte, CA; [15]Blood and Marrow Transplantation Program, University of Michigan, Ann Arbor, MI; [16]Ospedale Bambino Gesu', Rome, Italy; [17]Aflac Cancer and Blood Disorders Center, Emory University, Atlanta, GA; [18]Blood and Marrow Transplantation Program, Columbia University Medical Center, New York, NY; [19]Division of Hematology/Oncology/BMT, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada; [20]Med. Klinik III/Poliklinik, Universitatsklinik Erlangen, Erlangen, Germany; and [21]Pediatric Blood and Marrow Transplantation Program, Children's Hospital, University of Würzburg, Würzburg, Germany

Key Points

- ST2, REG3α, and/or AREG at the time of acute GVHD diagnosis are excellent predictors of risk for 12-month NRM.
- The best biomarker algorithm and threshold for risk stratification may depend on the target population.

Graft-versus-host disease (GVHD) is a major cause of nonrelapse mortality (NRM) after allogeneic hematopoietic cell transplantation. Algorithms containing either the gastrointestinal (GI) GVHD biomarker amphiregulin (AREG) or a combination of 2 GI GVHD biomarkers (suppressor of tumorigenicity-2 [ST2] + regenerating family member 3 alpha [REG3α]) when measured at GVHD diagnosis are validated predictors of NRM risk but have never been assessed in the same patients using identical statistical methods. We measured the serum concentrations of ST2, REG3α, and AREG by enzyme-linked immunosorbent assay at the time of GVHD diagnosis in 715 patients divided by the date of transplantation into training (2004-2015) and validation (2015-2017) cohorts. The training cohort (n = 341) was used to develop algorithms for predicting the probability of 12-month NRM that contained all possible combinations of 1 to 3 biomarkers and a threshold corresponding to the concordance probability was used to stratify patients for the risk of NRM. Algorithms were compared with each other based on several metrics, including the area under the receiver operating characteristics curve, proportion of patients correctly classified, sensitivity, and specificity using only the validation cohort (n = 374). All algorithms were strong discriminators of 12-month NRM, whether or not patients were systemically treated (n = 321). An algorithm containing only ST2 + REG3α had the highest area under the receiver operating characteristics curve (0.757), correctly classified the most patients (75%), and

Submitted 27 June 2023; accepted 29 March 2024; prepublished online on *Blood Advances* First Edition 19 April 2024. https://doi.org/10.1182/bloodadvances.2023011049.

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

Data are available on request from the corresponding author, John Levine (john.levine@mssm.edu).

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

more accurately risk-stratified those who developed Minnesota standard-risk GVHD and for patients who received posttransplant cyclophosphamide-based prophylaxis. An algorithm containing only AREG more accurately risk-stratified patients with Minnesota high-risk GVHD. Combining ST2, REG3α, and AREG into a single algorithm did not improve performance.

Introduction

Acute graft-versus-host disease (GVHD) remains a major cause of morbidity and mortality after allogeneic hematopoietic stem cell transplantation (HCT), despite modern prophylaxis regimens such as posttransplant cyclophosphamide (PT-CY) that have reduced the maximum severity of GVHD but the overall incidence and the need for systemic treatment remain high.[1-5] The maximum severity of acute GVHD is mainly driven by gastrointestinal (GI) tract damage and correlates well with nonrelapse mortality (NRM) and survival, but can only be determined in retrospect after treatment.[6] The overall symptom severity at diagnosis, when the full extent of GI damage is not yet known, correlates modestly with response to treatment and long-term outcomes[7,8] but the overall grade is used nonetheless to determine eligibility for treatment trials tailored for different risks, such as less toxic therapy for low-risk[9-11] and more intensive therapy for high-risk GVHD[12] (also NCT05263999, NCT04167514). Furthermore, clinical practice is heterogeneous, with some clinicians choosing topical therapy, whereas others prescribe systemic corticosteroids for patients with similar mild acute GVHD presentations. Laboratory measures of GVHD severity at the time of diagnosis that more accurately predict treatment outcomes than clinical symptoms, such as GI GVHD biomarkers, are needed to design more efficient clinical trials and may help guide treatment selection, including the decision to use topical rather than systemic treatment.

Three serum biomarkers, regenerating family member 3 alpha (REG3α), suppressor of tumorigenicity-2 (ST2), and amphiregulin (AREG), quantify GI damage in the context of acute GVHD.[13-18] REG3α is an antimicrobial peptide secreted by Paneth cells that provides a key survival signal for intestinal stem cells necessary for the regeneration of GI crypts.[16,17] ST2 is a ligand for interleukin-33 (IL-33), a protein secreted by damaged epithelial cells.[19] The interaction between IL33 and ST2 is thought to be anti-inflammatory under normal conditions but it can potentiate GI tissue damage during the inflammation of GVHD.[20] AREG, a weak epidermal growth factor ligand, promotes GI epithelial barrier repair and is secreted by effector cells, such as innate lymphoid cell-2s, gut-associated lymphoid tissue, and alloreactive T cells when stimulated by IL-33.[21-25] ST2 and AREG are both strongly associated with the repair and inflammatory cascade known as the ST2-IL-33 axis, although the exact role of these biomarkers in GVHD pathogenesis is still an area of active study.[26,27]

Several groups have validated acute GI GVHD biomarker-based algorithms to predict short- and long-term outcomes at the time of diagnosis. The Mount Sinai Acute GVHD International Consortium (MAGIC) algorithm probability uses the concentrations of 2 biomarkers, ST2 and REG3α, to predict the response to systemic therapy, risk of NRM, and survival.[28-30] This algorithm was recently validated as a prognostic tool that is superior to clinical prediction models, such as the Minnesota GVHD risk system.[6,8,31] A group at the University of Minnesota developed AREG as a prognostic biomarker that predicts both the risk of NRM and OS and is superior to the Minnesota risk system.[32,33] Both algorithms have been used to select patients with high- or low-risk GVHD for clinical trials testing primary treatment[34-36] (also clinical trials NCT05123040, NCT02525029, NCT04291261, and NCT05090384).

A recent publication showed that the combination of ST2 and REG3α was the most accurate of a panel of 5 biomarkers in predicting GVHD outcomes but AREG was not included in that analysis.[28] In this study, we expanded upon our prior work using the same large cohort of patients and identical statistical techniques to evaluate the combination of ST2, REG3α, and AREG that best stratifies patients with GVHD according to risk for 12-month NRM.

Methods

Study design and oversight

The MAGIC database and biorepository uses a PRoBE (prospective-specimen collection, retrospective-blinded-evaluation) design in which serum samples and clinical data are prospectively collected before clinical outcomes are known, biomarker concentrations are determined without knowledge of the patient's clinical status or outcome, and unbiased methods (eg, random assignment) are used to include subjects in analyses.[37] In this study, we included 715 patients from the MAGIC database and biorepository diagnosed with acute GVHD as defined by the MAGIC criteria,[38] with sufficient remaining serum from a prior 730 patient study that compared GI and systemic biomarkers for predicting GVHD outcomes (supplemental Table 1).[28] All the patients received topical and/or systemic therapy for acute GVHD upon diagnosis. Patients were divided into a training cohort (n = 341), who underwent allogeneic HCT between May 2004 and October 2015, and a validation cohort (n = 374), who underwent allogeneic HCT between November 2015 and April 2017, as previously reported (supplemental Table 2).[28] PT-CY-based GVHD prophylaxis has become increasingly prevalent; thus, we supplemented the PT-CY subset with an additional 77 patients from the MAGIC database and biorepository who underwent allogeneic HCT between 2020 and 2022 and developed GVHD (supplemental Table 3). To avoid selection bias, we included patients who sequentially underwent transplantation in the reverse order from the most recent patient with 12 months follow-up. The size of the PT-CY subset (n = 133) relative to the total validation cohort (n = 451) approximates the proportion of patients (29%) in the MAGIC database and biorepository who received PT-CY prophylaxis between 2020 and 2022. All patients, parents, and legal guardians provided informed consent on an institutional review board-approved protocol.

Biomarker determination and algorithm development

We measured ST2 and REG3α concentrations at the Icahn School of Medicine at Mount Sinai[28] and measured AREG concentrations at the University of Minnesota using enzyme-linked immunosorbent assay according to published protocols.[32] ST2 and AREG were expressed in picograms per milliliter and REG3α was expressed in nanograms per milliliter. All biomarker values were log-10 transformed for use in algorithms. Competing risk regression that considered relapse and second transplant as competing risks was used in the training cohort to create biomarker algorithms for all 7 possible combinations of 1, 2, or 3 biomarkers to predict the probability of 12-month NRM from the time of diagnosis of GVHD. Each algorithm calculated the predicted probability of 12-month NRM as a value from 0.001 to 0.999 using the complementary log-log link for each individual patient in the training cohort and then identified the threshold to separate low and high risk according to the concordance probability (the value that maximizes sensitivity and specificity).[39] Threshold performance was assessed using several metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), balanced accuracy, and the proportion of patients correctly classified as high or low risk. Balanced accuracy was defined as the average sensitivity and specificity.[40] Patients were deemed correctly classified as high-risk if they died from NRM within 12 months from the diagnosis of GVHD and as low-risk if they did not experience NRM. All assessments of performance and comparisons among algorithms used data from only the validation cohort.

NRM was defined as death within 12 months of GVHD onset from any cause other than relapse. Treatment response on day 28 of systemic therapy was defined as follows: a complete response required resolution of all GVHD symptoms; a partial response required improvement of at least 1 stage in at least 1 organ without worsening in any other organ; initiation of second-line systemic therapy, death before day 28, and all other responses were categorized as nonresponses.

Statistical methods

Patient characteristics between the training and validation cohorts were compared using the χ^2 or Wilcoxon 2-sample tests as appropriate. Correlations between individual biomarker algorithms were evaluated using the Pearson correlation coefficient. Areas under the receiver operating characteristic curves were compared using DeLong test,[41] and *P* value were adjusted for the false discovery rate.[42] The cumulative incidences of NRM and relapse were calculated using Fine and Gray method.[43] Differences in cumulative incidences were compared using Gray test[44] except when the cumulative incidence curves crossed, in which case the cumulative incidence rates at 12 months were compared using χ^2 tests.[45] Overall survival (OS) was estimated using the Kaplan-Meier method, and the differences between groups were compared using the log-rank test. *P*-values were corrected for multiple comparisons using the Benjamini-Hochberg method.[46] All tests were 2-sided and statistical significance was considered at $P < .05$. All analyses were performed using SAS version 9.4 and R statistical package version 4.0.3 (R Core Team 2020).

All patients, parents, and legal guardians provided informed consent on an institutional review board-approved protocol.

Results

Patient characteristics

Patient characteristics are shown in supplemental Table 2 (training and validation cohorts) and supplemental Table 3 (expanded PT-CY subset). The significant differences in age distribution, indication for transplantation, conditioning intensity, donor type, and GVHD prophylaxis between cohorts reflect changes in transplant practices between the earlier training (2004-2015) and later validation (2015-2017) cohorts. Despite these changes, there were no significant differences in GVHD characteristics, such as target organ involvement, severity at diagnosis, maximum severity, systemic treatment, or 12-month NRM. There were more patients with late acute GVHD in the validation cohort, whose median day of GVHD onset was 2 days later. When we applied the algorithms and thresholds for ST2 + REG3α and AREG that are currently in use in clinical trials[29,32,34,36] (also NCT05123040 and NCT02525029) to the full cohort (n = 715), both algorithms performed similarly well (supplemental Figure 1).

Algorithm creation and validation

We used the training cohort to create algorithms that predicted the risk of 12-month NRM for all 7 possible combinations of ST2, REG3α, and AREG (supplemental Table 4). Each individual biomarker was an independent discriminator of the 12-month NRM, either alone or in a pairwise combination. When all 3 biomarkers were combined, REG3α and AREG remained significant predictors of 12-month NRM risk but ST2 was no longer significant, a finding that likely reflects that the correlation among biomarkers was the highest for AREG and ST2 (supplemental Figure 2). We used the validation cohort to calculate the area under the curve (AUC) of the receiver operating characteristic curves for each algorithm (Table 1). All algorithms were strong discriminators of the 12-month NRM but algorithms that combined biomarkers had greater AUCs than the single biomarker algorithms; the largest AUC belonged to the combination of ST2 + REG3α. There were no statistically significant differences among the algorithms (supplemental Table 5) and all algorithms produced 2 groups with significantly different risk of NRM. Similar findings were observed when the analyses were limited to the subset of validation cohort patients who underwent systemic treatment for acute GVHD (n = 321) (Table 1; supplemental Table 6), in which the ST2 + REG3α algorithm again had the largest AUC (0.739). Therefore, we

Table 1. AUC for each algorithm applied to the validation cohort

	All (n = 374)		Systemically treated subset (n = 321)	
	AUC	*P* value*	AUC	*P* value*
ST2	0.710	<.001	0.694	<.001
REG3α	0.711	<.001	0.698	<.001
AREG	0.707	<.001	0.693	<.001
ST2 + AREG	0.734	<.001	0.718	<.001
ST2 + REG3α	0.757	<.001	0.739	<.001
REG3α + AREG	0.736	<.001	0.721	<.001
ST2 + REG3α + AREG	0.752	<.001	0.735	<.001

*For comparison of observed AUC to 0.5 as the null.

Table 2. **Performance characteristics (threshold corresponds to concordance probability)**

Algorithm	Threshold	% high risk	Sensitivity	Specificity	PPV	NPV	Balanced accuracy	Correctly classified
AREG	0.231	45%	0.69	0.61	0.29	0.89	0.65	62%
ST2 + REG3α	0.247	30%	0.63	0.77	0.39	0.90	0.70	75%
ST2 + REG3α + AREG	0.204	48%	0.77	0.59	0.30	0.92	0.68	60%

focused our comparisons on the algorithms AREG and ST2 + REG3α, given their use in clinical trials, and the combination of ST2 + REG3α + AREG.

Risk stratification

We identified thresholds for each algorithm in the training cohort corresponding to the concordance probability that maximized sensitivity and specificity and used those thresholds to risk stratify patients in the validation cohort (Table 2; supplemental Table 7). ST2 had the highest specificity, REG3α had the highest sensitivity, and the combination of the ST2 + REG3α algorithm had the highest balanced accuracy; these findings were unchanged when only patients who were systemically treated were analyzed (supplemental Table 7). All algorithms stratified patients into low- and high-risk groups with large and highly statistically significant differences in the 6-month and 12-month NRM (supplemental Table 8). The differences in 12-month NRM created by the ST2 + REG3α algorithm (29%) were considerably larger than that created by the AREG algorithm (18%) (Figure 1A-B) because of the higher specificity of the ST2 + REG3α algorithm, which correctly classified more patients as low risk than the AREG algorithm (70% vs 55%). Similar differences in 12-month NRM were observed when the algorithms were applied only to patients who were systemically treated (Figure 1C-D; supplemental Table 8). The inclusion of all 3 biomarkers in an algorithm resulted in risk groups that were close in size and PPV to the algorithm of AREG alone (Table 2). As expected from prior studies, the cumulative incidence of relapse was not significantly different with any algorithm; thus, all differences in the 12-month NRM translated into statistically and clinically significant differences in the 12-month OS (supplemental Table 8). As expected, patients at high risk for NRM were less likely to respond to systemic corticosteroid treatment than patients at low risk in the subset of the validation cohort patients who received systemic treatment (321/374, 86%), with the largest difference between groups again observed using the ST2 + REG3α algorithm (supplemental Table 9). When we chose a second threshold of 80% specificity as determined in the training cohort and applied to it the validation cohort, we found similar results in which the ST2 + REG3α algorithm correctly classified the greatest number of patients and produced the largest differences in NRM between groups for all patients as well as the subset of patients who were systemically treated (supplemental Tables 10 and 11).

Analyses of key subsets helped explain these modest differences in performance between the ST2 + REG3α and AREG algorithms, although it is important to note that there were no statistically significant differences among the AUCs (Table 3). Both algorithms successfully stratified patients who received systemic treatment for GVHD (n = 321; AREG 12% vs 30%, $P < .001$; ST2 + REG3α: 12% vs 39%, $P < .001$), and both algorithms stratified patients with lower GI GVHD at diagnosis effectively (n = 109, AREG: 8% vs 45%, $P < .001$; ST2 + REG3α: 12% vs 48%, $P < .001$). Minnesota risk classification stratifies patients for risk of NRM. AREG further risk stratified the Minnesota high-risk subset (Figure 2A-C; supplemental Table 12) and ST2 + REG3α more accurately classified the standard-risk subset (Figure 2D-F; supplemental Table 12). AREG correctly classified 7% more patients with Minnesota high-risk GVHD than ST2 + REG3α, but this group composed a small proportion of patients with GVHD (53/374, 14%); thus, the overall net effect was a correct classification of 1% more of the total population. In contrast, in patients with Minnesota standard-risk GVHD, which comprised most patients, ST2 + REG3α correctly classified 17% more patients than AREG (79% vs 62%); therefore, the overall net effect was a correct classification of 15% more patients in the total population. Further subset analysis showed that in patients with only skin rash at diagnosis (n = 199), ST2 + REG3α created distinct risk strata (8% vs 22%, $P < .001$) but AREG did not (10% vs 13%, $P = .218$). In addition, patients who received PT-CY-based GVHD prophylaxis (n = 133) were successfully stratified for risk of NRM using the ST2 + REG3α algorithm (12% vs 35%, $P < .001$) but not by the AREG algorithm (14% vs 21%, $P = .22$). These findings were consistent when the analyses were limited to patients who received systemic treatment for GVHD (Table 3). Given the increased use of PT-CY as GVHD prophylaxis and the large number of patients who presented only with rashes at the time of GVHD diagnosis, the ST2 + REG3α algorithm appears preferable for these important subgroups.

Discussion

Recent advances in GVHD prophylaxis have decreased the overall incidence of clinically severe (grade III/IV) GVHD but not the overall incidence of GVHD that requires treatment.[5,47,48] High-performance laboratory tests that predict GVHD outcomes are needed to tailor therapy based on the risk. The 2 validated GVHD risk stratification algorithms used in clinical trials used different biomarkers and both predicted GVHD outcomes well. In this study, we used a large international multicenter cohort to directly compare these 2 algorithms and evaluate whether novel combinations of biomarkers would improve their performance. Each of the 2 algorithms was an excellent discriminator of the 6-month and 12-month NRM in these patients when assessed using identical statistical methods. Although other combinations of these GI GVHD biomarkers performed well, adding AREG to ST2 + REG3α did not improve upon AREG or ST2 + REG3α, perhaps because AREG is a downstream component of the IL-33/ST2 axis.

The ST2 + REG3α algorithm had several modest advantages over the other algorithms: it more accurately classified patients, identified greater differences in 12-month NRM and OS between the

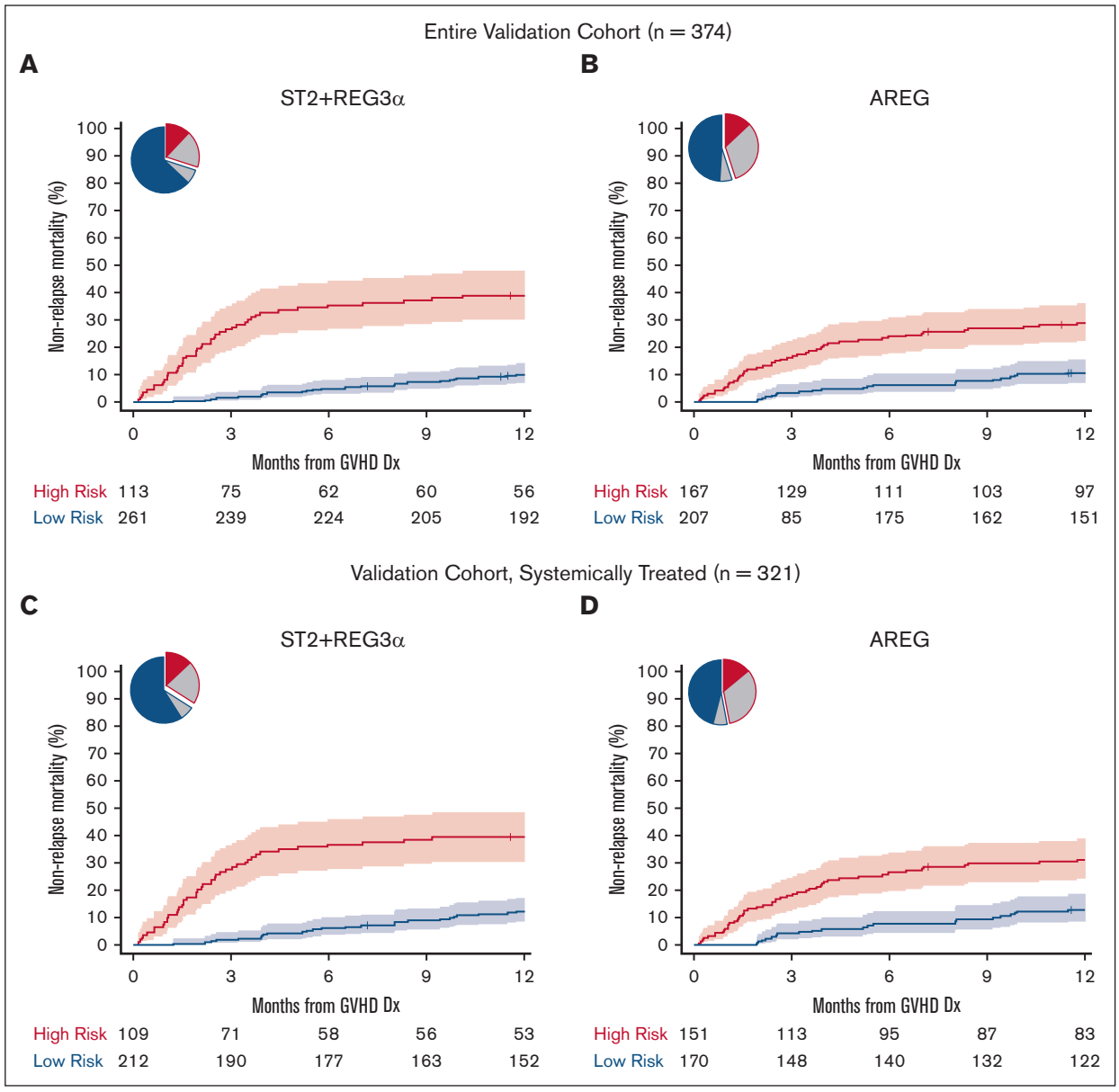

Figure 1. Twelve-month NRM by risk classification for AREG and ST2 + REG3α biomarker algorithms (validation cohort). Pie charts show the proportion of patients classified as high-risk (HR, red border) and low-risk (LR, blue border). The proportions correctly classified as HR or LR are shaded in red and blue, respectively. The proportion incorrectly classified is shaded in gray. The cumulative incidence curves show the 12-month NRM, with shaded regions representing 95% confidence intervals. (A and B): All validation cohort patients (C and D): systemically treated subset. (A) ST2 + REG3α: NRM 39% vs 10%, $P < .001$; (B) AREG: NRM 29% vs 11%, $P < .001$; (C) ST2 + REG3α: NRM 39% vs 12%, $P < .001$; (D) AREG: NRM 31% vs 13%, $P < .001$.

high- and low-risk groups and performed well in most patients (Minnesota standard risk GVHD), in which it was best at identifying patients who were at high risk of NRM despite the absence of high-risk clinical symptoms. This last finding may be because ST2 + REG3α identifies more patients as having a high risk for NRM before lower GI GVHD symptoms have manifested. Notably, the AREG algorithm was slightly better than ST2 + REG3α for patients with Minnesota high-risk GVHD. The ability to identify patients at low risk for NRM, despite the presence of high-risk clinical symptoms, is clinically important, as it may help avoid over-treatment. The addition of a third GI biomarker, AREG, did not improve the performance of ST2 + REG3α, which is likely due to the strong correlation between ST2 and AREG. The IL-33/ST2 axis plays a key role in the pathogenesis of GVHD, whereas IL-33 is both the ST2 ligand and an inducer of AREG secretion by type 2 innate lymphoid cells; however, the interactions surrounding these proteins in GVHD biology remain poorly described and an area of active investigation.[19,23,27,49]

Our study has several limitations. First, the training and validation cohorts were primarily obtained from a prior study.[28] The training cohort differed from the validation cohort, reflecting the evolution of transplant practices: patients were younger, had different indications for transplant, had a different donor mix, received

Table 3. Cumulative incidence of 12-month NRM for key subsets using the threshold corresponding to the concordance probability

		ST2+REG3α				AREG		
Subset	**AUC**	**CI 12 mo NRM**	***P* value**	**Correctly classified**	**AUC**	**CI 12 mo NRM**	***P* value**	**Correctly classified**
Validation cohort (n = 374)								
Minnesota high risk (n = 53)	0.689	24% vs 48%	.083	29 (55%)	0.780	8% vs 53%	.013	33 (62%)
Minnesota standard risk (n = 321)	0.720	9% vs 34%	<.001	250 (79%)	0.636	11% vs 21%	.003	200 (62%)
LGI involvement (n = 109)	0.790	12% vs 48%	<.001	73 (67%)	0.773	8% vs 45%	<.001	66 (61%)
Skin only involvement (n = 199)	0.629	8% vs 22%	<.001	155 (79%)	0.563	10% vs 13%	.218	129 (65%)
PT-CY prophylaxis (n = 133)	0.717	12% vs 35%	<.001	100 (75%)	0.595	14% vs 21%	.22	83 (62%)
Systemically treated (n = 321)	0.739	12% vs 39%	<.001	231 (72%)	0.693	12% vs 30%	<.001	195 (61%)
Systemically treated subset (n = 321)								
Minnesota high risk (n = 52)	0.679	26% vs 48%	.116	28 (54%)	0.780	8% vs 54%	.011	33 (63%)
Minnesota standard risk (n = 269)	0.700	12% vs 35%	<.001	203 (75%)	0.617	13% vs 23%	.022	162 (60%)
LGI involvement (n = 104)	0.777	14% vs 49%	<.001	68 (65%)	0.768	9% vs 46%	<.001	63 (61%)
Skin only involvement (n = 158)	0.610	11% vs 24%	.034	116 (73%)	0.555	13% vs 16%	.489	99 (63%)
PT-CY prophylaxis (n = 125)	0.711	13% vs 35%	.005	94 (75%)	0.592	16% vs 23%	.233	72 (58%)

LGI, lower gastrointestinal tract.

different GVHD prophylaxis, and were less likely to experience late-onset GVHD. The fact that the algorithms performed well in both cohorts is reassuring, but patients who underwent transplantation after 2017 were not included with the exception of the PT-CY subset; thus, these analyses do not fully reflect the most current transplant practices. It is noteworthy that the ST2 + REG3α algorithm successfully stratified patients who received PT-CY prophylaxis for risk of 12-month NRM, whereas the AREG algorithm did not. However, some other clinically relevant subsets were too small for analysis, such as patients whose GVHD treatment subsequently required treatment with ruxolitinib (n = 14). Second, previous ST2 + REG3α and AREG biomarker algorithms were developed from different data sets using different statistical methods and end points. In this study, we created new algorithms, including novel combinations, from the training set using identical statistical methods and 12-month NRM as the primary end point to compare algorithm performance with minimal bias. Thus, the new algorithms differ from the versions of the algorithms used in past clinical trials, although this new ST2 + REG3α algorithm yields the same results as the previously published algorithm (supplemental Figure 1). Furthermore, although categorical risk scores, such as high/low risk, are useful for separating patients into groups, the field would benefit from the development of calibrated risk scores that can be applied to individual patients. For example, it might be useful if a continuous value, such as the MAGIC Algorithm Probability or AREG concentration, accurately estimates an individual patient's risk for NRM. Third, although there were clinically meaningful differences in performance among the algorithms in specific subgroups, no algorithm was statistically superior to any other. Fourth, the small number of patients with specific characteristics, such as Minnesota high-risk GVHD, raises the possibility that some subset analyses were underpowered and further study of these groups is needed. Fifth, our data set was not optimized to evaluate certain late complications such as chronic GVHD. Future studies are needed to prospectively collect late clinical events to evaluate the ability of different algorithms to predict long-term outcomes other than NRM and survival. Finally, a larger data set than the one used here would be needed to detect significant differences between algorithms and identify scenarios in which 1 algorithm might be preferred over another.

In conclusion, biomarkers enhance clinical risk stratification strategies by identifying patients at increased risk of NRM within the Minnesota standard-risk population and those at decreased risk of NRM within the Minnesota high-risk population. Thus, biomarkers will play an increasingly important role in GVHD clinical trial design and ultimately in clinical practice. Because no single biomarker is universally ideal, the choice of algorithm and threshold should be guided by the research or clinical aim. If the goal is to deescalate GVHD treatment, high NPV and sensitivity facilitate the identification of patients most likely to respond to standard treatment and survive long-term. For example, a recent study showed that patients with Minnesota standard-risk GVHD and low-risk biomarker scores by ST2 + REG3α could be successfully treated with inhibition of JAK1 by itacitinib monotherapy, thereby avoiding exposure to the toxicity of systemic corticosteroids.[34] Conversely, high PPV and specificity are preferred for identifying a population at high risk of poor outcomes or for studying a potentially toxic intervention. For example, patients with Minnesota standard-risk GVHD who are at high risk for biomarkers may be appropriate for inclusion in clinical trials that intensify treatment given the risk of failure with standard treatment, even at the risk of more treatment-related toxicity. Similarly, patients with Minnesota high-risk GVHD who are at low risk for NRM by biomarkers might be excluded from clinical trials that intensify GVHD treatment. One could also consider adjusting thresholds to maximize NPV or PPV for different clinical scenarios and treatment goals. In summary, algorithms based on ST2 + REG3α and AREG, including previously published versions with their accompanying risk stratification thresholds, are suitable for identifying patients across a wide range of clinical presentations and are appropriate for use across a spectrum of clinical trial designs. Periodic re-examination of algorithms will be necessary as clinical practice evolves.

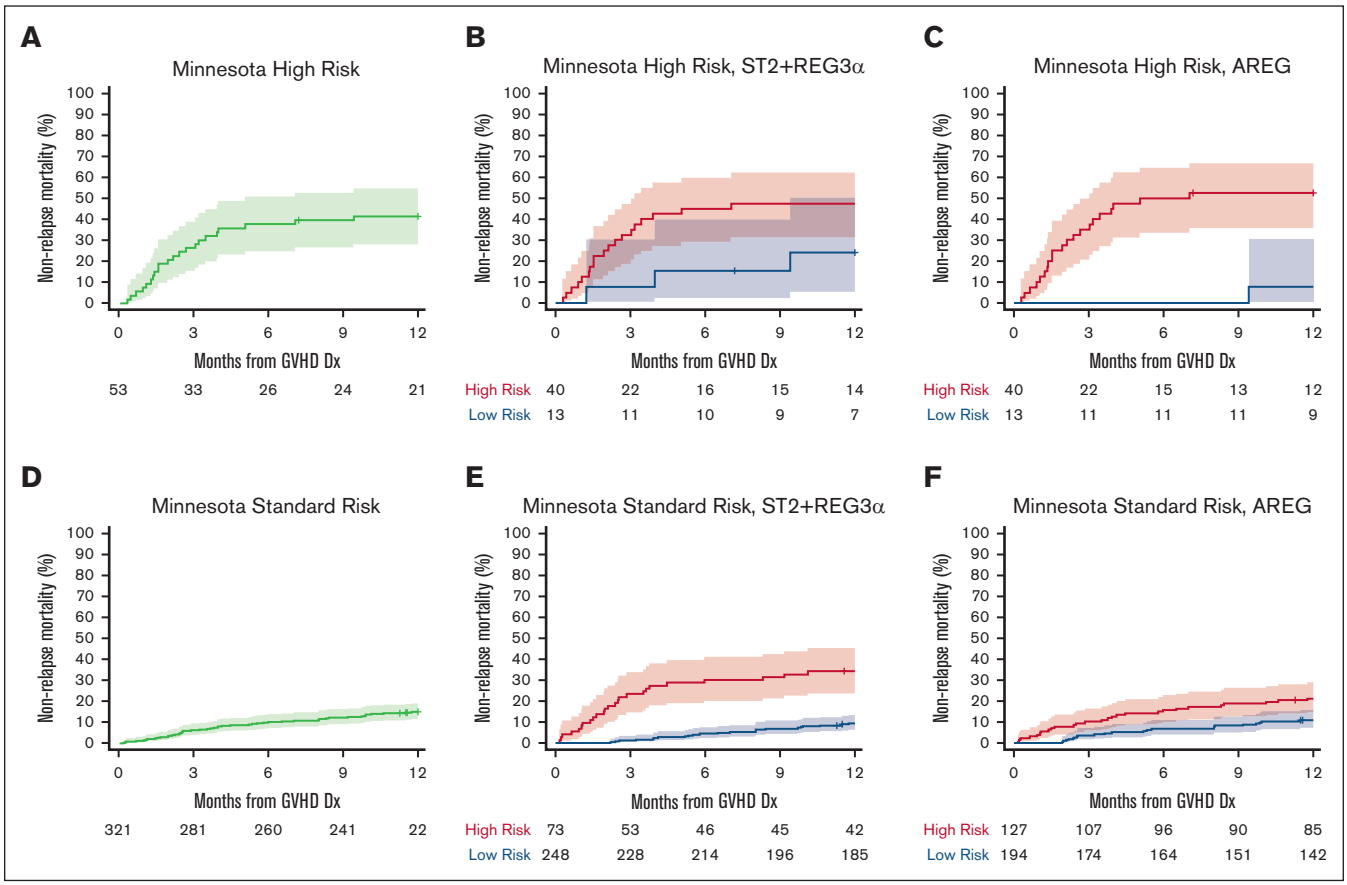

Figure 2. Twelve-month NRM by Minnesota risk and further stratification using ST2 + REG3α and AREG algorithms in the validation cohort. The cumulative incidence curves show 12-month NRM with shaded regions representing the 95% confidence intervals. (A) Minnesota high-risk acute GVHD: NRM, 42%. (B) Minnesota high-risk stratified by ST2 + REG3α: NRM 45% vs 24%, $P = .083$; (C) Minnesota high-risk stratified by AREG: 50% vs 8%, $P = .013$. (D) Minnesota standard-risk acute GVHD: NRM, 15%. (E) Minnesota standard risk stratified ST2 + REG3α: NRM 34% vs 9%, $P < .001$; (F) Minnesota standard risk-stratified by AREG: NRM 21% vs 11%, $P = .003$.

Acknowledgments

The authors thank the patients, their families, and the research staff for their participation. The authors also thank Gilbert Eng for the computer programming and database support.

This work was supported by the National Institutes of Health, National Cancer Institute grant PO1CA03942), University of Minnesota Foundation (GVHD Research Fund), Pediatric Cancer Foundation, and German Jose Carreras Leukemia Foundation grants DJCLS 01 GVHD 2016 and DJCLS 01 GVHD 2020.

Authorship

Contribution: A.E., D.K., J.L.M.F., J.E.L., and S.H. conceived and designed the study; A.E., P.A-H., F.A., J.B., C.C., Y-B.C., H.C., Z.D., S.G., E. Hexner, W.J.H., E. Holler, C.L.K., S.K., M.A.M., A.P., F.Q., M.Q., R.R., T.S., I.V., D.W., M.W., R.Y., R.N., and J.E.L. collected and reviewed the clinical data; S.G., G.M., S.K., and S.H. performed the laboratory analysis; A.E., N.K., N.S., and D.K. performed the statistical analysis; A.E., N.K., J.L.M.F., and J.E.L. wrote the report; and all authors reviewed and edited the manuscript.

Conflict-of-interest disclosure: B.C.B. is a coinventor of a CD83 CAR T cell licensed to CRISPR Therapeutics; received consulting fees from CTI BioPharma and Incyte; received research funding from Vitrac Therapeutics and CTI BioPharma; and is the current Director of Laboratory Science for American Society of Transplantation and Cellular Therapy. Y-B.C. received consulting fees from Incyte, Takeda, Vor Biopharma, Celularity, Equilium, and Pharmacosmos. H.C. received consulting fees from Incyte, Sanofi, Actinium, and REGiMMUNE, and research funding from Opna. C.L.K. received consulting fees from Horizon Therapeutics. M.A.M. received consulting fees from NexImmune, TScan, Hansa Biopharma, Stemline Therapeutics, CarDx, and Incyte; participated in a speakers' bureau for Sanofi; and received research funding from NexImmune and Gilead. M.Q. received honoraria from Novartis and Vertex. R.R. received consulting fees from Atara Biotherapeutics, Allogene, Gilead Sciences, Takeda, Incyte, Instil Bio, TScan, Synthekine, Orca, Quell Therapeutics, Capstan, and Jasper; served in an expert witness role with Bayer; and received research funding from Atara Biotherapeutics, Incyte, Sanofi, Immatics, AbbVie, TCR2, Takeda, Gilead Sciences, CareDx, TScan, Synthekine, Bristol Myers Squibb, Johnson & Johnson, Genentech, and Precision BioSciences. T.S. received consulting fees from Moderna. J.L.M.F. and J.E.L. are coinventors of a GVHD biomarker patent and receive royalties from its licensure. J.E.L. received consulting fees from bluebird bio, Editas,

Equillium, Incyte, Inhibrx, Kamada, Mesoblast, Sanofi, and X4 Pharmaceuticals, and research support from Genentech, Incyte, and Mesoblast. S.H. received consulting fees from Ossium Health; fees for clinical trial adjudication from CSL Behring; and research funding from Vitrac Therapeutics and Incyte. The remaining authors declare no competing financial interests.

ORCID profiles: A.E., 0009-0000-7360-3986; N.E.J., 0000-0002-9268-9655; N.K., 0009-0004-7955-1119; N.S., 0000-0001-5274-1749; P.A.-H., 0000-0002-0196-806X; Y.-B.C., 0000-0002-9554-1058; Z.D., 0000-0002-7994-8974; E.H., 0000-0002-1125-4060; W.J.H., 0000-0002-5841-4105; M.A.M., 0000-0001-8226-471X; F.Q., 0000-0002-7346-2738; M.Q., 0000-0001-7689-343X; D.W., 0000-0001-8078-8579; M.W., 0000-0002-9608-3482; R.N., 0000-0002-9082-0680; J.E.L., 0000-0002-5611-7828; S.H., 0000-0002-5054-9419.

Correspondence: John 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.

References

1. McDonald GB, Sandmaier BM, Mielcarek M, et al. Survival, nonrelapse mortality, and relapse-related mortality after allogeneic hematopoietic cell transplantation: comparing 2003-2007 versus 2013-2017 cohorts. *Ann Intern Med*. 2020;172(4):229-239.
2. Gooptu M, Antin JH. GVHD prophylaxis 2020. *Front Immunol*. 2021;12:605726.
3. Bolanos-Meade J, Reshef R, Fraser R, et al. Three prophylaxis regimens (tacrolimus, mycophenolate mofetil, and cyclophosphamide; tacrolimus, methotrexate, and bortezomib; or tacrolimus, methotrexate, and maraviroc) versus tacrolimus and methotrexate for prevention of graft-versus-host disease with haemopoietic cell transplantation with reduced-intensity conditioning: a randomised phase 2 trial with a non-randomised contemporaneous control group (BMT CTN 1203). *Lancet Haematol*. 2019;6(3):e132-e143.
4. Saliba RM, Alousi AM, Pidala J, et al. Characteristics of graft-versus-host disease (GvHD) after post-transplantation cyclophosphamide versus conventional GvHD prophylaxis. *Transplant Cell Ther*. 2022;28(10):681-693.
5. 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.
6. 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.
7. Cahn JY, Klein JP, Lee SJ, et al. Prospective evaluation of 2 acute graft-versus-host (GVHD) grading systems: a joint Societe Francaise de Greffe de Moelle et Therapie Cellulaire (SFGM-TC), Dana Farber Cancer Institute (DFCI), and International Bone Marrow Transplant Registry (IBMTR) prospective study. *Blood*. 2005;106(4):1495-1500.
8. MacMillan ML, DeFor TE, Holtan SG, Rashidi A, Blazar BR, Weisdorf DJ. Validation of Minnesota acute graft-versus-host disease risk score. *Haematologica*. 2020;105(2):519-524.
9. Mielcarek M, Furlong T, Storer BE, et al. Effectiveness and safety of lower dose prednisone for initial treatment of acute graft-versus-host disease: a randomized controlled trial. *Haematologica*. 2015;100(6):842-848.
10. Frairia C, Nicolosi M, Shapiro J, et al. Sole upfront therapy with beclomethasone and budesonide for upper gastrointestinal acute graft-versus-host disease. *Biol Blood Marrow Transplant*. 2020;26(7):1303-1311.
11. Gatza E, Braun T, Levine JE, et al. Etanercept plus topical corticosteroids as initial therapy for grade one acute graft-versus-host disease after allogeneic hematopoietic cell transplantation. *Biol Blood Marrow Transplant*. 2014;20(9):1426-1434.
12. Kekre N, Kim HT, Hofer J, et al. Phase II trial of natalizumab with corticosteroids as initial treatment of gastrointestinal acute graft-versus-host disease. *Bone Marrow Transplant*. 2021;56(5):1006-1012.
13. Ferrara JLM, Chaudhry MS. GVHD: biology matters. *Hematology Am Soc Hematol Educ Program*. 2018;2018(1):221-227.
14. 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.
15. 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.
16. 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.
17. Zhao D, Kim YH, Jeong S, et al. Survival signal REG3alpha prevents crypt apoptosis to control acute gastrointestinal graft-versus-host disease. *J Clin Invest*. 2018;128(11):4970-4979.
18. Amin K, Yaqoob U, Schultz B, et al. Amphiregulin in intestinal acute graft-versus-host disease: a possible diagnostic and prognostic aid. *Mod Pathol*. 2019;32(4):560-567.
19. Griesenauer B, Paczesny S. The ST2/IL-33 axis in immune cells during inflammatory diseases. *Front Immunol*. 2017;8:475.
20. Zeiser R, Blazar BR. Acute graft-versus-host disease - biologic process, prevention, and therapy. *N Engl J Med*. 2017;377(22):2167-2179.
21. Zaiss DMW, Gause WC, Osborne LC, Artis D. Emerging functions of amphiregulin in orchestrating immunity, inflammation, and tissue repair. *Immunity*. 2015;42(2):216-226.

22. Jansen SA, Nieuwenhuis EES, Hanash AM, Lindemans CA. Challenges and opportunities targeting mechanisms of epithelial injury and recovery in acute intestinal graft-versus-host disease. *Mucosal Immunol*. 2022;15(4):605-619.

23. Monticelli LA, Osborne LC, Noti M, Tran SV, Zaiss DM, Artis D. IL-33 promotes an innate immune pathway of intestinal tissue protection dependent on amphiregulin-EGFR interactions. *Proc Natl Acad Sci U S A*. 2015;112(34):10762-10767.

24. Walker JA, Barlow JL, McKenzie AN. Innate lymphoid cells–how did we miss them? *Nat Rev Immunol*. 2013;13(2):75-87.

25. Holtan SG, Hoeschen AL, Cao Q, et al. Facilitating resolution of life-threatening acute GVHD with human chorionic gonadotropin and epidermal growth factor. *Blood Adv*. 2020;4(7):1284-1295.

26. Ito T, Takashima S, Calafiore M, et al. Donor-derived amphiregulin drives CD4+ T cell expansion and promotes tissue pathology after experimental allogeneic BMT. *Blood*. 2022;140(Suppl 1):1152-1153.

27. 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.

28. 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.

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

30. Robin M, Porcher R, Michonneau D, et al. Prospective external validation of biomarkers to predict acute graft-versus-host disease severity. *Blood Adv*. 2022;6(16):4763-4772.

31. Spyrou N, Akahoshi Y, Ayuk FA, et al. The utility of biomarkers in acute GVHD prognostication. *Blood Adv*. 2023;7(17):5152-5155.

32. Holtan SG, DeFor TE, Panoskaltsis-Mortari A, et al. Amphiregulin modifies the Minnesota acute graft-versus-host disease risk score: results from BMT CTN 0302/0802. *Blood Adv*. 2018;2(15):1882-1888.

33. Holtan SG, Khera N, Levine JE, et al. Late acute graft-versus-host disease: a prospective analysis of clinical outcomes and circulating angiogenic factors. *Blood*. 2016;128(19):2350-2358.

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

35. Pidala J, Hamadani M, Dawson P, et al. Randomized multicenter trial of sirolimus vs prednisone as initial therapy for standard-risk acute GVHD: the BMT CTN 1501 trial. *Blood*. 2020;135(2):97-107.

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. 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.

38. 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 Transplant*. 2016;22(1):4-10.

39. Liu X. Classification accuracy and cut point selection. *Stat Med*. 2012;31(23):2676-2686.

40. Velez DR, White BC, Motsinger AA, et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet Epidemiol*. 2007;31(4):306-315.

41. 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.

42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B*. 1995;57(1):289-300.

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

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

45. Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. *Stat Med*. 2007;26(24):4505-4519.

46. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*. 2001;125(1-2):279-284.

47. Watkins B, Qayed M, McCracken C, et al. Phase II trial of costimulation blockade with abatacept for prevention of acute GVHD. *J Clin Oncol*. 2021;39(17):1865-1877.

48. Greinix HT, Eikema DJ, Koster L, et al. Improved outcome of patients with graft-versus-host disease after allogeneic hematopoietic cell transplantation for hematologic malignancies over time: an EBMT mega-file study. *Haematologica*. 2022;107(5):1054-1063.

49. Liew FY, Girard JP, Turnquist HR. Interleukin-33 in health and disease. *Nat Rev Immunol*. 2016;16(11):676-689.