HHS Public Access

Author manuscript

Lancet Haematol. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

Lancet Haematol. 2015 January 1; 2(1): e21–e29. doi:10.1016/S2352-3026(14)00035-0.

A PROGNOSTIC SCORE FOR ACUTE GRAFT-VERSUS-HOST DISEASE BASED ON BIOMARKERS: A MULTICENTER STUDY

John E. Levine, M.D.[1], **Thomas M. Braun, Ph.D.**[2], **Andrew C. Harris, M.D.**[1], **Ernst Holler, M.D.**[3], **Austin Taylor, B.A.**[1,4], **Holly Miller, D.O.**[1], **John Magenau, M.D.**[1], **Daniel J. Weisdorf, M.D.**[5], **Vincent T. Ho, M.D.**[6], **Javier Bolaños-Meade, M.D.**[7], **Amin M. Alousi, M.D.**[8], **L.M. Ferrara, M.D.**[1,4], and **for the Blood and Marrow Transplant Clinical Trials Network**

[1]Blood & Marrow Transplant Program, University of Michigan, Ann Arbor

[2]School of Public Health, University of Michigan, Ann Arbor

[3]Department of Hematology and Oncology, University of Regensburg, Regensburg, Germany

[4]The Tisch Cancer Institute, The Icahn School of Medicine at Mount Sinai Hospital, University of Minnesota, Minneapolis

[5]Blood & Marrow Transplant Program, University of Minnesota, Minneapolis

[6]Department of Medical Oncology, Dana Farber Cancer Institute, Boston

[7]Department of Oncology, Johns Hopkins School of Medicine, Baltimore

[8]Department of Stem Cell Transplantation, MD Anderson Cancer Center, Houston

SUMMARY

Background—Graft-versus-host disease (GVHD) is the major cause of non-relapse mortality (NRM) after allogeneic hematopoietic stem-cell transplantation (HCT). The severity of symptoms at the onset of GVHD does not accurately define risk, and thus most patients are treated alike with high dose systemic corticosteroids. We aimed to define clinically meaningful risk strata for patients with newly diagnosed acute GVHD using plasma biomarkers.

Methods—We prospectively collected plasma from 492 HCT patients with newly diagnosed acute GVHD and randomly divided them into training (n=328) and test (n=164) sets. We used the concentrations of three recently validated biomarkers (TNFR1, ST2, and REG3α) to create an algorithm that computed the probability of NRM six months after GVHD onset for individual

Corresponding Author: James L. M. Ferrara, MD, DSc, The Tisch Cancer Institute, The Icahn School of Medicine at Mount Sinai Hospital, 1470 Madison Avenue, 6th Fl., Rm. 110, New York, New York 10029 or james.ferrara@mssm.edu.

Contributors:

JEL, TB, and JLMF were responsible for the study design, data analysis and wrote the paper. AT and HM performed laboratory analyses. JEL, ACH, and EH collected data. All authors interpreted the data and contributed to writing and the paper. JEL and JLMF had full access to all the data in the study and jointly held final responsibility for the decision to submit for publication.

Declaration of interests

JEL, TB, and JLMF are co-inventors on a patent for the use of GVHD biomarkers. No other authors declare competing interests.

patients in the training set alone. We rank ordered the probabilities and identified thresholds that created three distinct NRM scores. We evaluated the algorithm in the testset, and again in an independent validation set of 300 additional HCT patients enrolled on multicenter clinical trials of primary therapy for acute GVHD.

Findings—In all three datasets, the cumulative incidence of twelve month NRM significantly increased as the GVHD score increased (8% [95% confidence interval (CI); 3%, 16%], 27% [95% CI; 20%%, 34%], and 46% [95% CI; 33%, 58%], for scores 1, 2 and 3 respectively in the multicenter validation set, $p<0 \cdot 0001$). Conversely, the response rates to primary GVHD treatment decreased as the GVHD score increased (86%, 67%, and 46%, for scores 1, 2 and 3 respectively in the multicenter validation set, $p<0 \cdot 0001$).

Interpretation—Biomarker-based scores can be used to guide risk-adapted therapy at the onset of acute GVHD.

INTRODUCTION

The ability of allogeneic hematopoietic stem cell transplantation (HCT) to cure hematologic malignancies is due, in part, to graft-versus-leukemia (GVL) effect mediated by alloreactive T cells in the donor graft. But GVL effects remain closely associated with graft versus host disease (GVHD) that is mediated by those same T cells as well as natural killer cells.[1] GVHD, which occurs in both acute and chronic forms, remains the major cause of death without relapse of primary disease, or non-relapse mortality (NRM).[2–4] The primary treatment of acute GVHD, high dose systemic glucocorticoids, has not changed in forty years.[5] Only one-third of patients achieve durable responses to initial corticosteroid therapy and survival among the remaining patients is poor.[6]

One important obstacle to the development of new therapies of acute GVHD is the inability to determine risk for an individual patient at the onset of symptoms. Mortality risk correlates with maximal clinical severity in current grading systems, which can only be assigned retrospectively after the response to treatment is known.[7–9]Thus, at disease onset most patients are treated alike with high dose corticosteroids resulting in significant numbers of patients who are both undertreated and overtreated. Overtreated patients who are likely to respond to low doses of glucocorticoids experience the additional infectious risks associated with profound immunosuppression as well as morbidities such as avascular necrosis of bone and diabetes mellitus.[10–13]Undertreated patients who develop steroid resistant acute GVHD experience a mortality rate in excess of 70–90%.[14–16]

In this study we have developed an algorithm using the concentration of plasma biomarkers to predict the probability of six month NRM at the onset of acute GVHD symptoms. This algorithm defines three scores with distinct mortality risks that may eventually prove useful as a guide to therapy for acute GVHD.

METHODS

Study population

The study population for training and test sets consisted of 792 patients with new onset acute GVHD grade I–IV. 492 patients from the University of Michigan and the University of Regensburg, Germany provided blood samples at the onset of acute GVHD on IRB-approved protocols at each center. Both centers used standardized guidance that was developed through a long-standing collaboration to minimize variability in the diagnosis and estimation of the severity of acute GVHD.[17]The initial dose of systemic corticosteroid therapy for GVHD treatment was between 1–2 mg/kg/day of methylprednisolone, as determined by the treating physician who used best medical judgment that considered a variety of factors such as GVHD severity and timing, donor source, infectious history, relapse risk, etc.300 patients from multiple centers who provided blood samples at the time of enrollment on Blood and Marrow Transplant Clinical Trial Network (BMT CTN) clinical trials of primary therapy for GVHD (see Supplemental Methods) formed an independent multicenter validation set. Patients from the University of Michigan who participated in BMT CTN clinical trials were included only in the training and test set.

Primary and secondary endpoints

The primary endpoint, NRM at six months from GVHD onset, was defined as any death without preceding relapse. Treatment response was a secondary endpoint that required improvement in overall clinical (modified Glucksberg) GVHD grade on day 28 after onset without additional systemic immunosuppressants. Complete response (CR) was defined as resolution of all target organ symptoms. Partial response was defined as an improvement of any organ stage by at least one stage without increase in any other target organ stage. All other treatment outcomes were classified as non-response. We categorized GVHD responses as durable if patients achieved CR by day 28 and remained in CR at six months post-onset. Patients who died before response assessments were considered non-responders. Six month GVHD staging data was available only for patients from the University of Michigan and the University of Regensburg (n=492).

Sample collection, preparation, and analysis

From 13 April 2000 to 7 May 2013, plasma samples were collected prospectively within 48 hours before or after the initiation of glucocorticoid therapy from patients who developed GVHD symptoms after HCT. Clinical GVHD grading was performed according to modified Glucksberg criteria[9] (see Supplementary Appendix). Enzyme-linked immunosorbent assays (ELISA) were performed as previously described.[18–21]

Statistical methods

We employed a competing risks regression model according to the methods of Fine and Gray[22] using log-transformed biomarker concentrations at the onset of GVHD from the training set alone to predict 6 month NRM in the training set. In the resulting algorithm, each biomarker was assigned a weight computed by the model that best fit the data of the training set alone. The sum of these weighted concentrations led to a predicted probability,

p, for each individual patient. Models with different numbers of biomarkers (from one to five) were fit to the training set alone. For a model to warrant examination, each weighted biomarker needed to be statistically significant and the model needed to be statistically superior (by the likelihood ratio test) to a model where all weights were zero.

The remaining models were then compared to each other using likelihood ratio tests and Akaike's Information Criterion (AIC)[23]. The most parsimonious model included TNFR1, REG3α, and ST2; models with either one or two biomarkers were statistically inferior and models with four or five biomarkers were not statistically superior. The final algorithm is shown below:

$$\log[-\log(1-p)] = -9.169 + 0.598(\log_2\text{TNFR1}) - 0.028(\log_2\text{REG3}\alpha) + 0.189(\log_2\text{ST2})$$

We then rank ordered the probability of NRM, p, in the training set and identified thresholds of p to define three scores such that 1 represented an excellent outcome (NRM ≤10%), 3 a poor outcome (NRM >40%), and thus NRM would increase by 15% on average with each increasing score. Multiple thresholds that met these criteria were evaluated in the test set; representative threshold pairs and their corresponding NRMs are shown in Table S1. Of note, we did not compare organ specific biomarkers to the algorithm in patients with single organ disease because of the relative paucity of such patients.

Overall differences in patient characteristics between the training, test and multicenter validation set were assessed with a chi-squared test of association for categorical values and a Wilcoxon Rank Sum test for continuous values. Estimation and inference for non-relapse mortality and relapse rates were based on the methods of Gray[24] and Fine and Gray.[22] Estimation and inference for overall survival were based on Cox regression, and estimation and inference for Day 28 CR and CR/PR rates were based on logistic regression. Empirical area under the receiver operating characteristic curves (AUC) for NRM by six months was computed nonparametrically. All analyses were performed in the statistical package R version 3.0.1 (R Development Core Team, Vienna, Austria).

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

RESULTS

Study population

The clinical characteristics of the two-center training and testsets and the multicenter BMT CTN validation set are shown in Table S2. The test and multicenter validation sets differed significantly in their overall distributions of age, stem cell source, indication for HCT, day of onset and severity of GVHD, and GVHD prophylaxis. Patients in the multicenter validation set were older (median of 52 vs 48 years), more likely to receive marrow as a stem cell source (21% vs 11%), and they developed GVHD later after HCT (median of 34 vs 27 days). They were also less likely to have high-risk disease (20% vs 43%), to have grade I

GVHD at onset (17% vs 37%), or to receive a CNI-containing GVHD prophylaxis (79% vs 96%).

Biomarker algorithm defines risk of NRM at GVHD onset

We hypothesized that plasma concentrations of one or more biomarkers at the diagnosis of acute GVHD could create an algorithm to predict 6 month NRM after GVHD diagnosis. To develop an algorithm that would be reliable in a multicenter setting, we collected blood samples from 492 HCT patients at two centers (the University of Michigan and the University of Regensburg) with similar diagnostic and therapeutic approaches to acute GVHD and randomly divided them into a training set (n=328) and a testset (n=164). For all 492 patients, we retrospectively analyzed plasma samples for concentrations of five biomarkers with prognostic value (IL2Rα, TNFR1, REG3α, Elafin, and ST2).[18–21,25] The two previously reported biomarkers with the weakest prognostic value (hepatocyte growth factor and IL8) were not included).[18] We used competing risks regression to develop an algorithm in the training set to compute a predicted probability (*p*) of NRM within six months of GVHD diagnosis. We then determined that an algorithm of the three biomarkers assigned the greatest weights (TFNR1, ST2, and REG3α) performed as well as the five-biomarker algorithm. Using the simpler algorithm, we determined the (*p*) for all patients in the training set, rank ordered them from lowest to highest, and identified thresholds that met predetermined desirable criteria for three GVHD scores (Ann Arbor 1 ≤10% and Ann Arbor 3 ≥40%) so that NRM would increase 15% on average with each increasing score. A range of thresholds met these criteria, and we chose one near the median of each range to define the Ann Arbor scores.

As seen in Figure 1A, this approach defined three distinct scores whose risk of NRM significantly increased with each increasing grade at both six months and twelve months after the onset of GVHD in the training set. We applied the algorithm to the testset (n=164) and found virtually identical risks of NRM (Figure 1B). We next applied the biomarker algorithm to an independent validation set of HCT patients enrolled on BMT CTN trials for primary GVHD therapy (n=300) and observed similarly significant differences in NRM (Figure 1C). Relapse, which was treated as a competing risk for NRM, did not significantly differ between the three Ann Arbor GVHD scores in any of the datasets (Figure 1D–F). The differences in NRM thus translated into significant differences in overall survival among these three scores after the onset of GVHD (Figure 1G–I).

Ann Arbor scores predict likelihood of treatment response

The response of GVHD to treatment 28 days later serves as a surrogate endpoint for long-term survival.[15,26] The proportion of all 792 patients who responded to therapy (generally systemic corticosteroids, Table S3 in the Supplementary Appendix) was highly statistically different for each of the Ann Arbor scores (1, 81%; 2, 68%; 3, 46%; p<0·0001 for all comparisons). We observed nearly identical rates in each dataset for CR and PR (Figure 2A–C) and for CR alone (Figure 2D–F).

A standard initial glucocorticoid dose for primary treatment of GVHD is 2 mg/kg/d of methylprednisolone[5], but clinicians may choose lower doses or, in the case of limited skin

GVHD (<50% body surface area), delay systemic treatment to avoid toxicity.[27,28] Intensity of initial steroid treatment of patients in the training and test sets (n=492) did not affect outcomes by Ann Arbor score. The responses by Ann Arbor score for patients with limited skin GVHD (Glucksberg grade I) who were treated (n=96) or not treated (n=98) with systemic steroids at diagnosis were similar (p=0·54). Likewise, the responses by Ann Arbor score for patients with Glucksberg ≥ grade II treated with 2 mg/kg/d of corticosteroids (n=137) or <2 mg/kg/d (n=64) were also similar (p=0.74). We assessed the durability of treatment response in all 492 Michigan and Regensburg patients. A durable response, defined as CR for at least six months without a recurrence of GVHD symptoms, was significantly less likely in patients with Ann Arbor 3 GVHD than patients with Ann Arbor 1 GVHD, regardless of organ involvement at GVHD onset (Table 1). Patients who presented with GVHD skin rash alone and were classified as Ann Arbor 1 were significantly more likely to achieve durable responses than those classified as Ann Arbor 3 (47% vs. 26%, p=0·0077). Likewise, patients with lower GI GVHD at onset were significantly more likely to achieve durable responses if their GVHD score was Ann Arbor 1 rather than Ann Arbor 3 (53% vs 13%, p<0·0001).

Ann Arbor scores predict development of GI GVHD

74 of 286 patients (26%) Michigan/Regensburg patients who presented with skin GVHD only subsequently developed lower GI GVHD; of this group, 14 of 34 (41%) presented with Glucksberg 1 and Ann Arbor 3. Thus the Ann Arbor score predicted the development of GI GVHD, but the extent of skin rash did not (Table 2). Patients with Ann Arbor 3 GVHD were 1.78 times more likely to later develop involvement of the GI tract(at a median of 12 days) than patients with Ann Arbor 1 GVHD (p=0·025).

Ann Arbor scores stratify for risk of NRM independently from clinical symptoms

We performed all subsequent analyses on the second, BMT CTN validation set because it represents a wide spectrum of supportive care and GVHD prophylaxis practices at a large number of centers, and in all patients the GVHD was considered significant enough to require treatment with systemic steroids as well as an experimental agent. As expected, the clinical grade of GVHD at onset did not always correlate with either response to treatment or with NRM (Figure S1). Despite the small sample sizes available for this subset analysis, the same biomarker algorithm defined three distinct risk strata for NRM within each Glucksberg grade (Figure 3A–C). Surprisingly, similar proportions of patients were assigned to each Ann Arbor score in each of the three Glucksberg grades. Patients with the higher Ann Arbor scores were also usually less likely to respond to treatment (Figure 3D–F). We did not find strong evidence for an interaction between severity of symptoms and Ann Arbor score on NRM (p=0·11), although statistical power was limited by the sample size.

The IBMTR acute GVHD severity index gives greater weight to GVHD of the skin relative to the Glucksberg grading system[7,8] (see Supplementary Appendix). When patients were categorized according to IBMTR grade the biomarker algorithm and thresholds defined GVHD scores with virtually identical risks of NRM and likelihood of treatment response (Figure S2).

Biomarker algorithm predicts risk of NRM better than Glucksberg grades

We used Akaike's Information Criterion (AIC)[23] to compare the biomarker algorithm to Glucksberg grades for their ability to model risk stratification of patients in the multicenter cohort. The AIC of Ann Arbor scoring was 12 units superior to Glucksberg grading. In order to better visualize this difference in AIC, we determined the hazard ratios (HR) for NRM in univariate models for both staging systems using moderate GVHD (Ann Arbor 2 or Glucksberg II) as the reference group (Table S4). We then fit a multivariate model with simultaneous adjustment for both Ann Arbor score and Glucksberg grades. As shown in Figure 4A, Ann Arbor 3 patients have significantly higher risk for NRM (p=0·0048) and Ann Arbor 1 patients have significantly less risk (p=0·0020) than patients with Ann Arbor 2. By contrast, the confidence intervals for the HRs of the Glucksberg grades encompass 1·0, demonstrating a lack of statistical significance between the grades. The area under the receiver operating characteristic curve for Ann Arbor scores (0·71) was also higher than that for Glucksberg grading (0·57), although this difference was not statistically significant (Figure 4B).

Biomarker algorithm defines risk independently of clinical risk factors

Several clinical risk factors, such as donor type, age, conditioning regimen intensity, and HLA-match, can predict for treatment response and survival in patients with GVHD.[15,29–31] Using Ann Arbor 2 as a reference, we found that Ann Arbor 1 predicted a lower risk of NRM (range 0·16–0·32) and Ann Arbor 3 a higher risk of NRM (range 1·4–2·9), regardless of the presence of these clinical risk factors (Table S5).

DISCUSSION

Maximal clinical severity of GVHD in symptom-based grading systems correlates with survival, but these systems are not often able to guide treatment at symptom onset. As a result, clinicians do not intensify immunosuppressive treatment of GVHD until primary therapy has failed. In this study, we have developed and validated an algorithm using biomarkers that defines three GVHD severity scores, each with a distinct risk of NRM. Ann Arbor GVHD scores defined risk across the full range of clinical presentations.

Importantly, a higher score predicted the development of GI GVHD in patients who presented without GI symptoms, which clinical grading did not. Most deaths of patients with GVHD that are not caused by relapse of primary disease are due to poor response to treatment of GVHD in the GI tract. It is therefore of significant interest that the three biomarkers included in this algorithm (TNFR1, ST2, and REG3α) all possess biological relevance to GI GVHD. TNFR1, a surrogate for TNFα, is produced by T cells and monocytes and amplifies GI injury.[32,33] TNFα regulates ST2 that, together with its ligand IL33 (a member of the IL1 family), influences inflammatory bowel disease activity.[34] REG3α, which we previously validated as a GI GVHD specific biomarker[20], is produced primarily by Paneth cells and protects GI epithelium from infectious damage.[35] The concentrations of these biomarkers at GVHD onset appear to reflect GI tract disease activity that does not correlate with the severity of GI symptoms at that time.

An important strength of this study is the biomarker algorithm's ability to define risk accurately despite differences in clinical severity at presentation and treatment intensity. In the dataset from the University of Michigan and University of Regensburg, approximately half of the patients who presented with rashes of less than 50% BSA (Glucksberg grade I), i.e. 20% of all patients, never required treatment with systemic steroids. In the multicenter BMT CTN dataset, all patients received treatment with systemic glucocorticoids and an experimental agent but the algorithm correctly identified patients at low risk of NRM.

Previous studies established correlations between either individual GVHD biomarkers or their combinations and clinical outcomes, but they lacked consistency among different clinical centers (panel).[18–21,25] The biomarker algorithm developed in this study advances the prior work but important limitations remain. First, although the algorithm predicts outcomes better than clinical symptoms, it still has relatively poor predictive power and is most useful for patients who score at either end. Second, the algorithm's ability to guide prospective therapy is yet to be demonstrated. Nevertheless, the algorithm should prove useful in the design of clinical trials. For example, a low score might be used as an exclusion criterion for patients with severe clinical symptoms (e.g., voluminous diarrhea) from a trial of an investigational agent: Such patients who are likely to respond to standard therapy, benefit by avoiding exposure to the risks of an experimental agent, and the trial also benefits by enrichment for patients who are less likely to respond to standard therapy. Conversely, a low score could be an inclusion criteria to limit exposure to lengthy glucocorticoid regimens. A high score (~23% of all GVHD) could be used as in inclusion criterion for a trial of intensive primary therapy. This approach would be particularly beneficial for patients with mild symptoms but who are less likely to respond to standard therapy and who might otherwise need to wait until primary treatment has failed before the initiation of an experimental modality. If a clinical trial is unavailable, a high score may lend confidence to the diagnosis of GVHD when a biopsy is equivocal, and a low score in a patient with a limited rash may support the use of topical treatment or watchful waiting.

Studies are currently underway to improve the predictive value of the algorithm. An attractive feature of the statistical methods used here is its ability to incorporate additional risk factors as they become known. For example, although donor type is not currently incorporated into GVHD grading systems, some studies show worse survival for patients with GVHD following an HCT from an unrelated volunteer donor.[15,26] Patients with HLA-mismatched donors were significantly more likely to have Ann Arbor 3 GVHD in all three datasets (Table S6–8). It is possible the incorporation of such a clinical characteristic, or even the nature of the GVHD symptoms at their onset, may improve the algorithm's predictive power. For example, in Michigan/Regensburg patients, the organ specific Glucksberg grade also correlated with durable response (Table S9). The algorithm has also not yet been adequately evaluated in patients who have both GVHD and other conditions, such as sinusoidal obstructive syndrome or bacterial sepsis, or who have uncommon GVHD presentations, such as isolated severe liver GVHD. It is also possible that the use of an algorithm at serial time points may prove useful, particularly in patients whose response to treatment is slow or partial. But much larger datasets will be required to test adequately such possibilities and combinations, probably on the order of several thousand patients. Yet we

anticipate future versions of this algorithm will prove increasingly useful and accelerate the development of precision medicine for HCT patients.

Panel: Research in context

Systematic review—We searched PubMed without language or date restrictions for articles with the following terms: “acute graft-versus-host disease” and “biomarkers” to September, 2014. We identified several relevant articles that showed prognostic significance for GVHD biomarkers at onset in single center studies.[18–21,25,36,37] However, there were no reports that validated GVHD biomarkers in multicenter studies.

Interpretation—To our knowledge, this study is the first to use biomarkers to classify patients at GVHD onset according to risk of treatment failure and non-relapse mortality outside of single centers. The biomarker algorithm was validated in patients from a broad spectrum of centers with a large variety of personnel and biases and was superior to clinical grading for determining risk. This study suggests that GVHD biomarker algorithm scores might be useful for designing risk-stratified trials of primary GVHD therapy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by grants #P01CA039542, #R21CA173459, and #P30CA046592 from the National Cancer Institute, #U10HL069294 from the National Heart, Lung, and Blood Institute, the National Cancer Institute and the Division of Allergy, Immunology, and Transplantation, National Institute of Allergy and Infectious Diseases, the Doris Duke Charitable Fund, the American Cancer Society, the Judith Devries Fund, and contributions by Eisai Inc., Hospira Inc., Roche Laboratories Inc., and Immunex Corporation, a wholly owned subsidiary of Amgen Inc. This manuscript was prepared using BMT CTN 0302 and BMT CTN 0802 Research Materials obtained from the NHLBI Biologic Specimen and Data Repository and the Information Coordinating Center, and the repository operated by the NMDP. It does not necessarily reflect the opinions or views of the BMT CTN 0302 or 0802 protocol teams or the NIH.

REFERENCES

1. Ferrara JL, Levine JE, Reddy P, Holler E. Graft-versus-host disease. Lancet. 2009; 373:1550–1561. [PubMed: 19282026]
2. Gooley TA, Chien JW, Pergam SA, et al. Reduced mortality after allogeneic hematopoietic-cell transplantation. N Engl J Med. 2010; 363:2091–2101. [PubMed: 21105791]
3. Anasetti C, Logan BR, Lee SJ, et al. Peripheral-blood stem cells versus bone marrow from unrelated donors. N Engl J Med. 2012; 367:1487–1496. [PubMed: 23075175]
4. Socie G, Ritz J, Martin PJ. Current challenges in chronic graft-versus-host disease. Biol Blood Marrow Transplant. 2010; 16:S146–S151. [PubMed: 19836455]
5. Martin PJ, Rizzo JD, Wingard JR, et al. First- and second-line systemic treatment of acute graft-versus-host disease: Recommendations of the American Society of Blood and Marrow Transplantation. Biol Blood Marrow Transplant. 2012; 18:1150–1163. [PubMed: 22510384]
6. Deeg HJ. How I treat refractory acute GVHD. Blood. 2007; 109:4119–4126. [PubMed: 17234737]
7. Rowlings P, Przepiorka D, Klein J, et al. IBMTR severity Index for grading acute graft-versus-host disease: restrospective comparison with Glucksberg grade. Brit J Haematol. 1997; 97:855–864. [PubMed: 9217189]
8. 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; 106:1495–1500. [PubMed: 15878974]

9. Przepiorka D, Weisdorf D, Martin P, et al. 1994 Consensus Conference on Acute GVHD Grading. Bone Marrow Transpl. 1995; 15:825–828.
10. Nucci M, Andrade F, Vigorito A, et al. Infectious complications in patients randomized to receive allogeneic bone marrow or peripheral blood transplantation. Transpl Infect Dis. 2003; 5:167–173. [PubMed: 14987200]
11. Holler E. Risk assessment in haematopoietic stem cell transplantation: GvHD prevention and treatment. Best Pract Res Clin Haematol. 2007; 20:281–294. [PubMed: 17448962]
12. Campbell S, Sun CL, Kurian S, et al. Predictors of a vascular necrosis of bone in long-term survivors of hematopoietic cell transplantation. Cancer. 2009; 115:4127–4135. [PubMed: 19536905]
13. Pidala J, Kim J, Kharfan-Dabaja MA, et al. Dysglycemia following glucocorticoid therapy for acute graft-versus-host disease adversely affects transplantation outcomes. Biol Blood Marrow Transplant. 2011; 17:239–248. [PubMed: 20637884]
14. Arai S, Margolis J, Zahurak M, Anders V, Vogelsang GB. Poor outcome in steroid-refractory graft-versus-host disease with antithymocyte globulin treatment. Biol Blood Marrow Transpl. 2002; 8:155–160.
15. Levine JE, Logan B, Wu J, et al. Graft-versus-host disease treatment: predictors of survival. Biol Blood Marrow Transplant. 2010; 16:1693–1699. [PubMed: 20541024]
16. Westin JR, Saliba RM, De Lima M, et al. Steroid-refractory acute GVHD: Predictors and outcomes. Adv Hematol. 2011
17. Levine JE, Hogan WJ, Harris AC, et al. Improved accuracy of acute graft-versus-host disease staging among multiple centers. Best Pract Res Clin Haematol. 2014
18. Paczesny S, Krijanovski OI, Braun TM, et al. A biomarker panel for acute graft-versus-host disease. Blood. 2009; 113:273–278. [PubMed: 18832652]
19. Paczesny S, Braun TM, Levine JE, et al. Elafin is a biomarker of graft-versus-host disease of the skin. Science Translational Medicine. 2010; 2:13ra2.
20. Ferrara JL, Harris AC, Greenson JK, et al. Regenerating islet-derived 3-alpha is a biomarker of gastrointestinal graft-versus-host disease. Blood. 2011; 118:6702–6708. [PubMed: 21979939]
21. 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; 369:529–539. [PubMed: 23924003]
22. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999; 94:496–509.
23. Akaike H. New look at statistical-model identification. IEEE T Automat Contr. 1974; Ac19:716–723.
24. Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. Annals of Statistics. 1988; 16:1141–1154.
25. Levine JE, Logan BR, Wu J, et al. Acute graft-versus-host disease biomarkers measured during therapy can predict treatment outcomes: a Blood and Marrow Transplant Clinical Trials Network study. Blood. 2012; 119:3854–3860. [PubMed: 22383800]
26. MacMillan ML, DeFor TE, Weisdorf DJ. The best endpoint for acute GVHD treatment trials. Blood. 2010; 115:5412–5417. [PubMed: 20388871]
27. Mielcarek M, Storer BE, Boeckh M, et al. Initial therapy of acute graft-versus-host disease with low-dose prednisone does not compromise patient outcomes. Blood. 2009; 113:2888–2894. [PubMed: 19001082]
28. Johnson ML, Farmer ER. Graft-versus-host reactions in dermatology. J Am Acad Dermatol. 1998; 38:369–392. [PubMed: 9520019]
29. Jagasia M, Arora M, Flowers ME, et al. Risk factors for acute GVHD and survival after hematopoietic cell transplantation. Blood. 2012; 119:296–307. [PubMed: 22010102]
30. Hurley CK, Woolfrey A, Wang T, et al. The impact of HLA unidirectional mismatches on the outcome of myeloablative hematopoietic stem cell transplantation with unrelated donors. Blood. 2013; 121:4800–4806. [PubMed: 23637130]

31. Storb R, Gyurkocza B, Storer BE, et al. Graft-versus-host disease and graft-versus-tumor effects after allogeneic hematopoietic cell transplantation. J Clin Oncol. 2013; 31:1530–1538. [PubMed: 23478054]

32. Schmaltz C, Alpdogan O, Muriglan SJ, et al. Donor T cell-derived TNF is required for graft-versus-host disease and graft-versus-tumor activity after bone marrow transplantation. Blood. 2003; 101:2440–2445. [PubMed: 12424195]

33. Hill GR, Ferrara JL. The primacy of the gastrointestinal tract as a target organ of acute graft-versus-host disease: Rationale for the use of cytokine shields in allogeneic bone marrow transplantation. Blood. 2000; 95:2754–2759. [PubMed: 10779417]

34. Pastorelli L, Garg RR, Hoang SB, et al. Epithelial-derived IL-33 and its receptor ST2 are dysregulated in ulcerative colitis and in experimental Th1/Th2 driven enteritis. Proc Natl Acad Sci U S A. 2010; 107:8017–8022. [PubMed: 20385815]

35. Ogawa H, Fukushima K, Naito H, et al. Increased expression of HIP/PAP and regenerating gene III in human inflammatory bowel disease and a murine bacterial reconstitution model. Inflamm Bowel Dis. 2003; 9:162–170. [PubMed: 12792221]

36. Ayuk F, Bussmann L, Zabelina T, et al. Serum albumin level predicts survival of patients with gastrointestinal acute graft-versus-host disease after allogeneic stem cell transplantation. Ann Hematol. 2014; 93:855–861. [PubMed: 24248672]

37. Magenau JM, Qin X, Tawara I, et al. Frequency of CD4(+)CD25(hi)FOXP3(+) regulatory t cells has diagnostic and prognostic value as a biomarker for acute graft-versus-host-disease. Biol Blood Marrow Transplant. 2010; 16:907–914. [PubMed: 20302964]

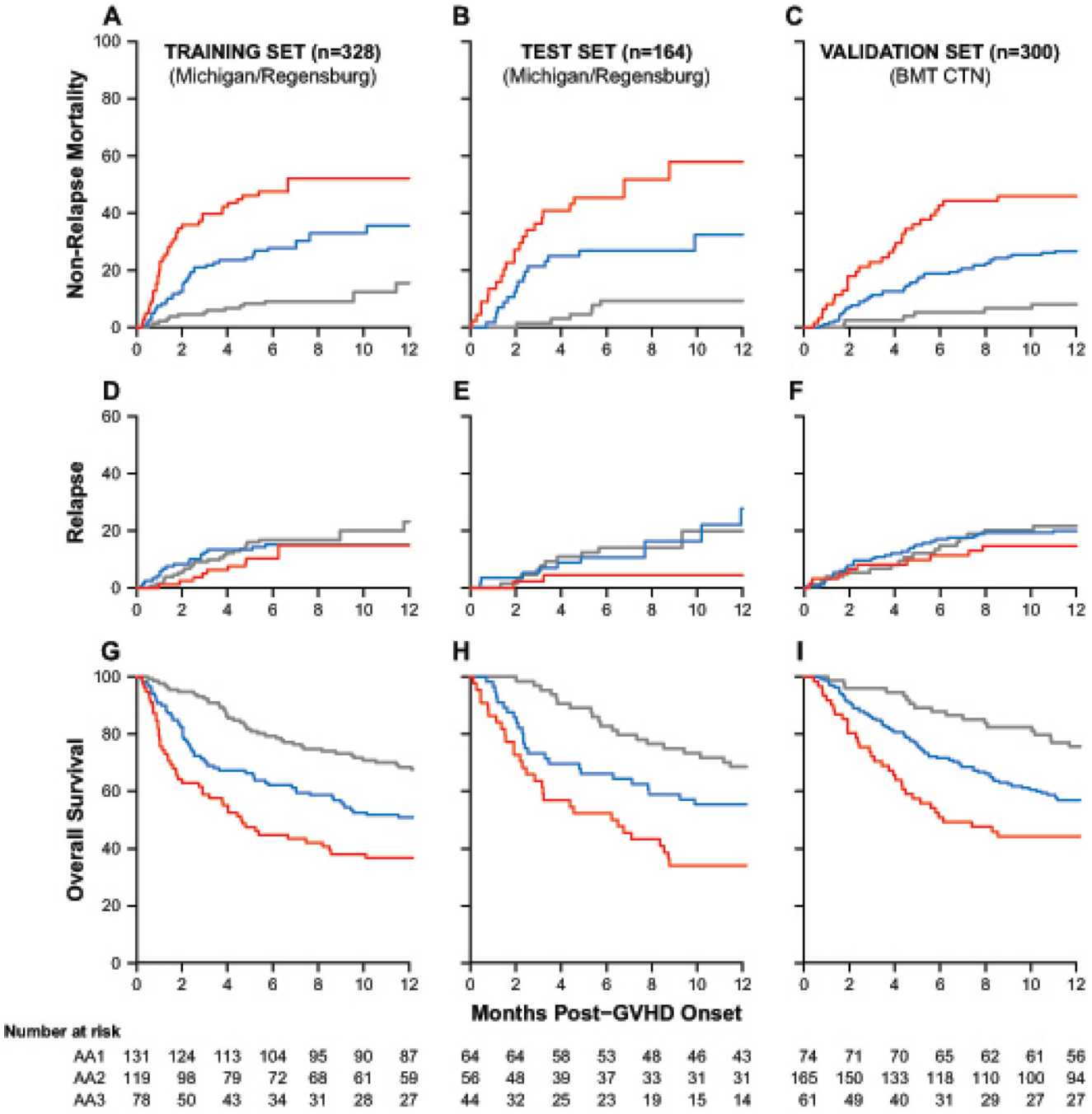

FIGURE 1. Outcomes by Ann Arbor score at GVHD onset

Cumulative incidence of non-relapse mortality is shown for the 328 patient training set (A, Ann Arbor 1 (grey) vs 2 (blue), p<0·0001, 2 vs 3 (red), p=0·0081), the 164 patient testset (B, Ann Arbor 1 vs 2, p=0·0069, 2 vs 3, p=0·024), and the 300 patient multicenter validation set (C, Ann Arbor 1 vs 2, p=0·0023, 2 vs 3, p=0·0021). Cumulative incidence of relapse was not significantly different in the training set (D), testset (E), or multicenter validation set (F). One-year survival is shown for the training set (G, Ann Arbor 1 vs 2, p=0·028, 2 vs 3,

p=0·015), the testset (H, Ann Arbor 1 vs 2, p=0·067, 2 vs 3, p=0·026), and the multicenter validation set (I, Ann Arbor 1 vs 2, p=0·0062, 2 vs 3, p=0·024).

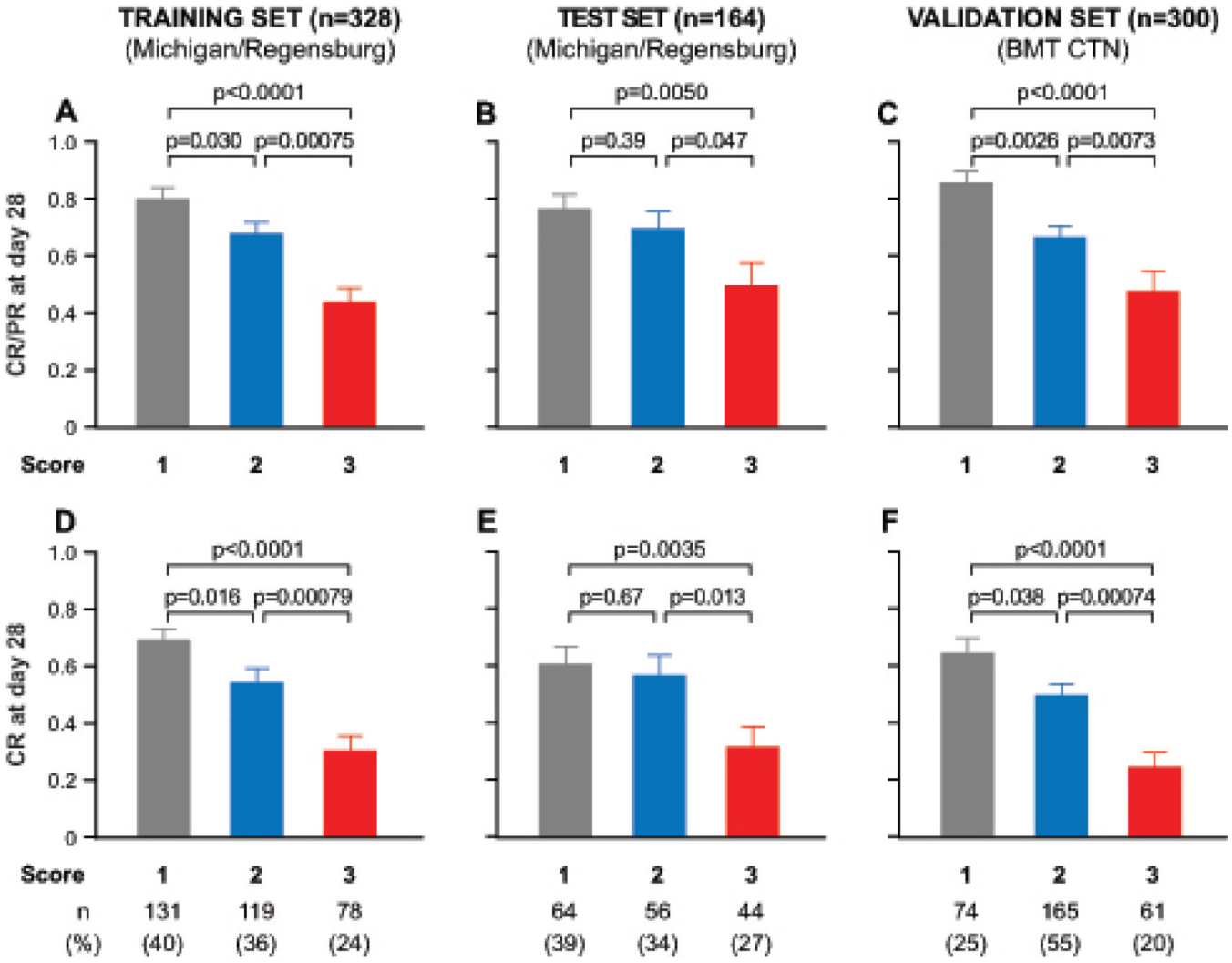

FIGURE 2. Response rate to primary GVHD therapy by Ann Arbor score

Shown are the proportion of patients with complete or partial response for (A) the training set, (B) the testset, and (C) the multicenter validation set. Complete response rates are shown in panels D–F. The numbers in parentheses are the proportion of patients assigned to each Ann Arbor score within each set.

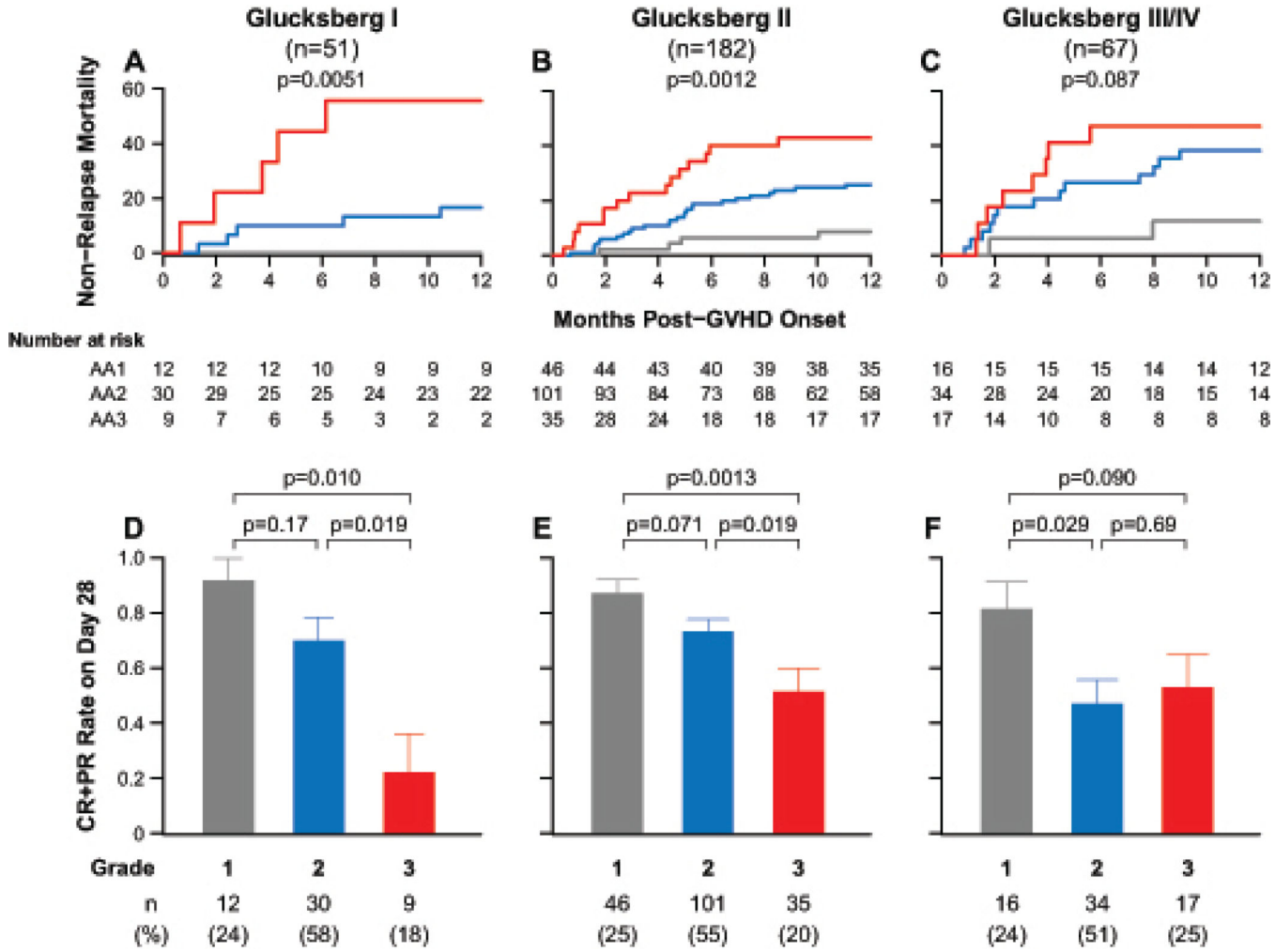

FIGURE 3. Outcomes for the multicenter validation set for each Glucksberg grade by Ann Arbor score

(1, grey; 2, blue; 3, red). Cumulative incidence of non-relapse mortality is shown for (A) Glucksberg grade I, p=0·0051; (B) grade II, p=0·0012; and (C) grade III/IV, p=0·087. P-values relate to any pairwise comparison of curves. The corresponding proportion of patients with complete or partial response are shown in Panels D–F. The numbers in parentheses are the proportion of patients assigned to each Ann Arbor score within each subset.

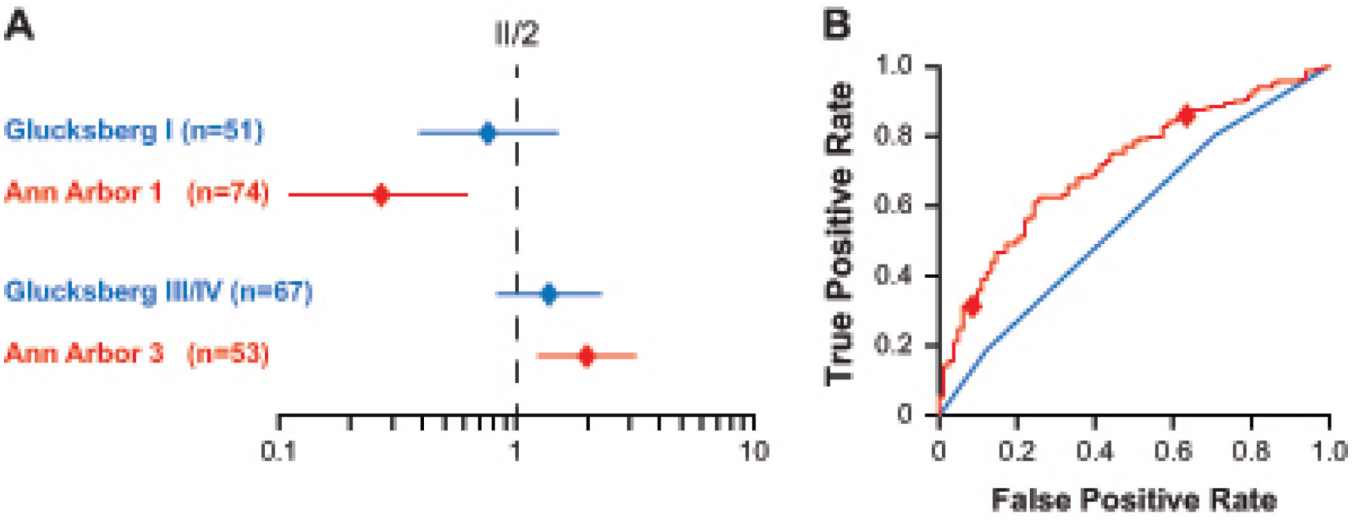

FIGURE 4. Comparison of Ann Arbor scores and Glucksberg grades to predict NRM risk (multicenter validation set)

(A) Multivariate comparison of Ann Arbor scoring (red) to Glucksberg grading (blue) for non-relapse mortality. Hazard ratios (diamonds) and their 95% confidence intervals [CI] (lines) for mild (grade I or Ann Arbor 1) and severe (grade III/IV or Ann Arbor 3) are shown relative to the reference group, moderate (II or 2) GVHD of each staging system. (B) ROC curves for the biomarker algorithm (red) or Glucksberg grades (blue) for prediction of non-relapse mortality. Diamonds indicate the thresholds that define Ann Arbor 1 and 3 GVHD. The AUC for the biomarker algorithm is 0.71 (95% CI, 0.64 to 0.75) and for Glucksberg grades is 0.57 (95% CI, 0.55 to 0.68).

Table 1

Ann Arbor Score Predicts Durability of Treatment Response

All presentations (n=492)					
Ann Arbor Score	**Sample Size**	**Durable Response**	**Percentage**	***p*-value**	
				vs AA1	**vs AA2**
1	195	97	50%	-	-
2	175	74	42%	0·15	
3	122	25	21%	<0·0001	<0·0001
Skin GVHD only at onset (stage 1–3) (n=286)					
1	129	60	47%	-	-
2	99	43	43%	0·64	-
3	58	15	26%	0·0077	0·028
Lower GI GVHD present at onset (± other organ involvement) (n=151)					
1	43	23	53%	-	-
2	55	18	33%	0·38	-
3	53	7	13%	<0·0001	0·016

Table 2

Proportion of Patients with Isolated Skin GVHD (stage 1–3) who Developed Lower GI GVHD Symptoms

Glucksberg Stage	N	Percent Develop Lower GI GVHD	Relative risk (vs Glucksberg 1)	p
1	81	28%	-	-
2	109	24%	0·84	0·48
3	96	26%	0·92	0·73
Ann Arbor Score	N	Percent Develop Lower GI GVHD	Relative risk (vs Ann Arbor 1)	p
1	129	19%	-	-
2	99	29%	1·51	0·081
3	58	34%	1·78	0·025