A prognostic score for acute graft-versus-host disease based on biomarkers: a multicentre study

John E Levine, Thomas M Braun, Andrew C Harris, Ernst Holler, Austin Taylor, Holly Miller, John Magenau, Daniel J Weisdorf, Vincent T Ho, Javier Bolaños-Meade, Amin M Alousi, James L M Ferrara, for the Blood and Marrow Transplant Clinical Trials Network

Summary

Background Graft-versus-host disease (GVHD) is the major cause of non-relapse mortality after allogeneic haemopoietic stem-cell transplantation (SCT). 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 Between April 13, 2000, and May 7, 2013, we prospectively collected plasma from 492 SCT patients with newly diagnosed acute GVHD and randomly assigned (2:1) using a random number generator, conditional on the final two digits of the patient’s date of birth, to a training set (n=328) and a test set (n=164). We used the concentrations of three recently validated biomarkers (TNFR1, ST2, and Reg3g) to create an algorithm that computed the probability of non-relapse mortality 6 months after GVHD onset for individual patients in the training set alone. We ranked ordered the probabilities and identified thresholds that created three distinct non-relapse mortality scores. We evaluated the algorithm in the test set, and again in an independent validation set of an additional 300 patients who underwent stem cell transplant and were enrolled on multicentre clinical trials of primary therapy for acute GVHD.

Findings In all three datasets (training, test, and validation), the cumulative incidence of 6-month non-relapse mortality significantly increased as the Ann Arbor GVHD score increased. In the multicentre validation set, scores were 8% (95% CI 3–16) for score 1, 27% (20–34) for score 2, and 46% (33–58) for score 3 (p<0·0001). Conversely, the mortality significantly increased as the GVHD score increased 86% for score 1, 67% for score 2, and 46% for score 3 in the multicentre validation set, p<0·0001).

Interpretation Biomarker-based scores can be used to guide risk-adapted therapy at the onset of acute GVHD. High risk patients with a score of 3 are candidates for intensive primary therapy, while low risk patients with a score of 1 are candidates for rapid taper of systemic steroid therapy.

Funding The National Cancer Institute, the National Heart, Lung, and Blood Institute, the Doris Duke Charitable Fund, the American Cancer Society, and the Judith Devries Fund.

Introduction The ability of allogeneic haemopoietic stem cell transplantation (SCT) to cure haematological malignancies is due, in part, to the graft-versus-leukaemia (GVL) effect mediated by alloreactive T cells in the donor graft. But GVL effects remain closely associated with graft-versus-host disease (GVHD), which is mediated by those same T cells and 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.2–4 The primary treatment of acute GVHD, high dose systemic glucocorticoids, has not changed in 40 years.5 Only a third of patients achieve durable responses to initial corticosteroid therapy and survival among the remaining patients is poor.6

An 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. Risk of mortality correlates with maximum clinical severity in current grading systems, which can only be assigned retrospectively after the response to treatment is known.7,8

Thus, at disease onset, most patients are treated alike with high dose corticosteroids resulting in substantial numbers of patients who are both undertreated and overtreated. Overtreated patients who are likely to respond to low doses of glucocorticoids have additional infectious risks associated with profound immunosuppression and morbidities such as avascular necrosis of bone and diabetes mellitus.9–11 An excess of 70–90% of undertreated patients who develop steroid resistant acute GVHD die.12–14

We aimed to define clinically meaningful risk strata for patients with newly diagnosed acute GVHD using plasma biomarkers.

Methods Clinical specimens and study design The study population for the training and test sets consisted of patients with new onset acute GVHD grade 1–IV from the University of Michigan, Ann Arbor, MI, USA (n=360) and the University of Regensburg, Germany (n=132), who provided blood samples at the onset of acute GVHD on institutional review board-approved protocols at

[www.thelancet.com/haematology Vol 2 January 2015]
each centre. Both centres used standardised guidance that was developed through a long-standing collaboration to minimise variability in the diagnosis and estimation of the severity of acute GVHD. Patients were randomly assigned (2:1) to the training or test set using a random number generator, conditional on the final two datasets having the same median day of onset. The initial dose of systemic corticosteroid therapy for GVHD treatment was between 1–2 mg/kg per day of methylprednisolone, as determined by the treating physician who used best medical judgment that considered a range of factors, including but not limited to, severity and timing of GVHD, donor source, infectious history, and relapse risk. 300 patients from multiple centres who provided blood samples at the time of enrolment on Blood and Marrow Transplant Clinical Trial Network (BMT CTN) clinical trials of primary therapy for GVHD (appendix p 2) formed an independent multicentre validation set. Patients from the University of Michigan who participated in BMT CTN clinical trials were included only in the training and test set. Between April 13, 2000, and May 7, 2013, plasma samples were collected prospectively within 48 h before or after the initiation of glucocorticoid therapy from patients who developed GVHD symptoms after SCT. Clinical grading of GVHD was done according to modified Glucksberg criteria9 (appendix p 3). The overall clinical grade (I–IV) combines the stages of three individual target organs (skin, liver, and gastrointestinal tract. ELISAs were done as previously described.18–21

Procedures
The primary endpoint, non-relapse mortality at 6 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 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 categorised GVHD responses as durable if patients achieved complete response by day 28 and remained in complete response at 6 months post-onset. Patients who died before response assessments were considered non-responders. Data for 6 month GVHD staging were available only for patients in the training and test sets from the University of Michigan and the University of Regensburg (n=492).

Statistical analysis
We used a competing risks regression model according to the methods of Fine and Gray22 using log-transformed biomarker concentrations at the onset of GVHD from the training set alone to predict 6 month non-relapse mortality in the training set. In the resulting algorithm, every 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 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. We then compared the remaining models to each other using likelihood ratio tests and Akaike’s Information Criterion (AIC) to determine which models were most parsimonious.23

We then rank ordered the probability of non-relapse mortality, p, in the training set and identified thresholds of p to define three scores such that 1 represented an excellent outcome (non-relapse mortality ≤10%) and 3 a poor outcome (non-relapse mortality >40%), and thus non-relapse mortality 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 non-relapse mortalities are shown in appendix p 7. 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 multicentre validation set were assessed with χ² 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 Gray24 and Fine and Gray, respectively.22 Estimation and inference for overall survival were based on Cox regression, and estimation and inference for day 28 complete response and complete or partial response rates were based on logistic regression. Empirical area under the receiver operating characteristic curves for non-relapse mortality by 6 months was computed nonparametrically. All analyses were done with R statistical package version 3.0.1 (R Development Core Team, Vienna, Austria).

Role of the funding source
The sponsors of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report. JEL and JLMF had full access to all data in the study and held joint final responsibility for the decision to submit for publication.

Results
The clinical characteristics of the two-centre training and test sets and the multicentre BMT CTN validation set are shown in appendix p 8. The multicentre validation set differed significantly from the test and training sets in their overall distributions of age, stem-cell source,
Articles

www.thelancet.com/haematology Vol 2 January 2015

To develop an algorithm that would be reliable in a multicentre setting, we collected blood samples from 492 SCT patients at two centres (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 test set (n=164). For all 492 patients, we retrospectively analysed plasma samples for concentrations of five biomarkers previously shown to have 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 interleukin 8) were not included.18 We used competing risks regression to develop an algorithm in the training set to compute a predicted probability of non-relapse mortality within 6 months of GVHD diagnosis. We then determined that an algorithm of the three biomarkers assigned the greatest weights (TNFR1, ST2, and REG3α) did as well as the five-biomarker algorithm. The final algorithm was:

\[
\log(1-p) = -9.169 + 0.598 \log(2^{\text{TNFR1}}) - 0.028 \log(2^{\text{REG3a}}) + 0.189 \log(\text{ST2}).
\]

Figure 1: Outcomes by Ann Arbor score at onset of graft-versus-host disease

Cumulative incidence of non-relapse mortality is shown for the 328 patient training set (A, Ann Arbor 1 [green] vs Ann Arbor 2 [blue], p<0.0001, Ann Arbor 2 vs Ann Arbor 3 [red], p=0.0081), the 164 patient test set (B, Ann Arbor 1 vs Ann Arbor 2, p=0.0069, Ann Arbor 2 vs Ann Arbor 3, p=0.024), and the 300 patient multicenter validation set (C, Ann Arbor 1 vs Ann Arbor 2, p=0.0023, Ann Arbor 2 vs Ann Arbor 3, p=0.0021). Cumulative incidence of relapse was not significantly different in the training set (Ann Arbor 1 vs Ann Arbor 2, p=0.35; Ann Arbor 2 vs Ann Arbor 3, p=0.59; D), test set (Ann Arbor 1 vs Ann Arbor 2, p=0.89, Ann Arbor 2 vs Ann Arbor 3, p=0.057; E), or multicentre validation set (Ann Arbor 1 vs Ann Arbor 2, p=0.014, Ann Arbor 2 vs Ann Arbor 3, p=0.37; F). 1-year survival is shown for the training set (G, Ann Arbor 1 vs Ann Arbor 2, p=0.028, Ann Arbor 2 vs Ann Arbor 3, p=0.015), the test set (H, Ann Arbor 1 vs Ann Arbor 2, p=0.067, Ann Arbor 2 vs Ann Arbor 3, p=0.026), and the multicentre validation set (I, Ann Arbor 1 vs Ann Arbor 2, p=0.0062, Ann Arbor 2 vs Ann Arbor 3, p=0.024).

The indication for SCT, day of onset and severity of GVHD, and GVHD prophylaxis (appendix p 8). Patients in the multicentre validation set were older than those in the other sets (median age 52 years [IQR 41–60] vs 48 years [34–56], more likely to receive marrow as a stem cell source (21% vs 11%), and they developed GVHD later after SCT (median 34 days vs 27 days). They were also less likely to have high-risk disease (17% vs 37%), or to receive a calcineurin inhibitor-containing GVHD prophylaxis (79% vs 96%).

To develop an algorithm that would be reliable in a multicentre setting, we collected blood samples from 492 SCT patients at two centres (the University of Michigan and the University of Regensburg) with similar diagnostic and therapeutic approaches to acute GVHD and...
Using the three biomarker 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 (upper threshold for Ann Arbor 1, non-relapse mortality ≤10%; and lower threshold for Ann Arbor 3 non-relapse mortality ≥40%) so that non-relapse mortality 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.

This approach defined three distinct scores whose risk of non-relapse mortality significantly increased with each increasing grade at both 6 months and 12 months after the onset of GVHD in the training set (figure 1A). We applied the algorithm to the test set \((n=164)\) and noted very similar risks of non-relapse mortality (figure 1B). We next applied the biomarker algorithm to an independent validation set of SCT patients enrolled on BMT CTN trials for primary GVHD therapy \((n=300)\) and noted similarly significant differences in non-relapse mortality (figure 1C). Relapse, which was treated as a competing risk for non-relapse mortality, did not differ significantly between the three Ann Arbor GVHD scores in any of the datasets (figure 1D–F). The differences in non-relapse mortality thus resulted in significant differences in overall survival at 1 year among these three scores after the onset of GVHD (figure 1G–I).

In all three datasets (training, test and validation), the cumulative incidence of non-relapse mortality significantly increased as the Ann Arbor score increased \((p<0·0001;\) table 1).

### Table 1: Cumulative incidence of NRM at 12 months

<table>
<thead>
<tr>
<th>Sample size ((n))</th>
<th>Cumulative incidence of NRM at 12 months, % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training set ((n=328))</strong></td>
<td></td>
</tr>
<tr>
<td>Ann Arbor score 1 131</td>
<td>16% (8–29)</td>
</tr>
<tr>
<td>Ann Arbor score 2 119</td>
<td>36% (26–49)</td>
</tr>
<tr>
<td>Ann Arbor score 3 78</td>
<td>52% (40–68)</td>
</tr>
<tr>
<td><strong>Test set ((n=164))</strong></td>
<td></td>
</tr>
<tr>
<td>Ann Arbor score 1 64</td>
<td>9% (4–20)</td>
</tr>
<tr>
<td>Ann Arbor score 2 56</td>
<td>32% (20–52)</td>
</tr>
<tr>
<td>Ann Arbor score 3 44</td>
<td>58% (41–82)</td>
</tr>
<tr>
<td><strong>Validation set ((n=300))</strong></td>
<td></td>
</tr>
<tr>
<td>Ann Arbor score 1 74</td>
<td>8% (3–16)</td>
</tr>
<tr>
<td>Ann Arbor score 2 165</td>
<td>27% (20–34)</td>
</tr>
<tr>
<td>Ann Arbor score 3 61</td>
<td>46% (33–58)</td>
</tr>
</tbody>
</table>

NRM=non-relapse mortality.

Figure 2: Response to primary graft-versus-host disease therapy after 28 days by Ann Arbor score
The proportion of patients with complete or partial response \((A–C)\) and the proportion of patients with complete response \((D–F)\). The numbers beneath the Ann Arbor score are the \(n\) (%) of patients with that score in each set. Error bars show SD.
The response of GVHD to treatment 28 days later serves as a surrogate endpoint for long-term survival.\textsuperscript{17,26} The proportion of all 792 patients who responded to therapy (generally systemic corticosteroids; appendix p 9) was significantly different for each of the Ann Arbor scores into which our algorithm categorised patients (81\% for score 1; 68\% for score 2; 46\% for score 3; p=0·0001 for all comparisons). We noted near similar proportions in each dataset for complete response and partial response (figure 2A–C) and for complete response alone (figure 2D–F).

A standard initial glucocorticoid dose for primary treatment of GVHD is 2 mg/kg per day of methylprednisolone,\textsuperscript{5} but clinicians might choose lower doses or, in the case of limited skin GVHD (<50\% body surface area), delay systemic treatment to avoid toxicity.\textsuperscript{7,28} Intensity of initial steroid treatment of patients in the training and test sets (n=492) did not affect outcomes by Ann Arbor score (appendix p 16). 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; appendix p 16). Likewise, the responses by Ann Arbor score for patients with Glucksberg grade II or greater treated with 2 mg/kg per day of corticosteroids (n=137) or less than 2 mg/kg per day (n=64) were also similar (p=0·74; appendix p 16). We assessed the durability of treatment response in all 492 patients from the Universities of Michigan and Regensburg. A durable response, defined as complete response for at least 6 months without a recurrence of GVHD symptoms, was significantly less likely in patients with Ann Arbor 3 GVHD than in patients with Ann Arbor 1 GVHD, regardless of organ involvement at GVHD onset (table 2). 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 who were classified as Ann Arbor 3 (table 3). Likewise, patients with lower gastrointestinal GVHD at onset were significantly more likely to achieve durable responses if their GVHD score was Ann Arbor 1 rather than was Ann Arbor 3. The organ specific Glucksberg grade also correlated with durable response in these patients (appendix p 15).

74 (26\%) of 286 patients from the Universities of Michigan and Regensburg (training and test sets) who presented with skin GVHD only subsequently developed lower gastrointestinal GVHD, of this group, 14 (41\%) of 34 presented with Glucksberg 1 and Ann Arbor 3 compared with 18 (23\%) of 67 with Ann Arbor 2 and 17 (19\%) of 89 with Ann Arbor 1 GVHD (p=0·042). Thus, the Ann Arbor score predicted the development of gastrointestinal GVHD, but the extent of skin rash did not (table 3). Patients with Ann Arbor 3 GVHD were 1·78 times more likely to later develop involvement of the gastrointestinal tract (at a median of 12 days [IQR 4–34]) than patients with Ann Arbor 1 GVHD (p=0·025).

We did all subsequent analyses on the validation set because it represented a wide range of supportive care and GVHD prophylaxis practices at a large number of centres, and in all patients the GVHD was deemed significant enough to require treatment with systemic steroids and an experimental drug. As expected, the clinical grade of GVHD at onset did not always correlate with either response to treatment or with non-relapse mortality (appendix p 5). Despite the small sample sizes available for this subset analysis, the same biomarker algorithm defined three distinct risk strata for non-relapse mortality 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 non-relapse mortality (p=0·11).
although statistical power was restricted by the sample size (data not shown).

The Center for International Blood and Marrow Transplant Research acute GVHD severity index gives greater weight to GVHD of the skin relative to the Glucksberg grading system\(^7\)\(^8\) (appendix p 3). When patients were categorised according to CIBMTR grade, the biomarker algorithm and thresholds defined GVHD scores with near similar risks of non-relapse mortality and likelihood of treatment response (appendix p 4).

The Akaike’s Information Criterion of Ann Arbor scoring (429) was lower than the Glucksberg grading (441), suggesting that Ann Arbor scores predict non-relapse mortality better than do Glucksberg grades.

---

**Table 1:** Outcomes for the multicentre validation set for each Glucksberg grade by Ann Arbor score

<table>
<thead>
<tr>
<th>Glucksberg</th>
<th>Ann Arbor 1</th>
<th>Ann Arbor 2</th>
<th>Ann Arbor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade I</td>
<td>12 (24%)</td>
<td>30 (58%)</td>
<td>9 (18%)</td>
</tr>
<tr>
<td>Grade II</td>
<td>16 (24%)</td>
<td>34 (51%)</td>
<td>15 (25%)</td>
</tr>
<tr>
<td>Grade III/IV</td>
<td>15 (24%)</td>
<td>15 (24%)</td>
<td>15 (25%)</td>
</tr>
</tbody>
</table>

Error bars show SD. GVHD = graft-versus-host disease.

**Figure 3:** Outcomes for the multicentre validation set for each Glucksberg grade by Ann Arbor score

Cumulative incidence of non-relapse mortality for (A) Glucksberg grade I, (B) grade II, and (C) grade III/IV. P values relate to any pairwise comparison of curves. (D–F) The corresponding proportion of patients with complete or partial response. The numbers beneath the Glucksberg grade are the n (%) of patients with that score in each set.

---

**Figure 4:** Comparison of Ann Arbor scores and Glucksberg grades to predict risk of non-relapse mortality (multicentre validation set)

(A) Multivariate comparison of Ann Arbor scoring (red) to Glucksberg grading (blue) for non-relapse mortality. Hazard ratios (diamonds) and their 95% 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) Receiver operating characteristic 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 Ann Arbor 3 GVHD. The area under curve for the biomarker algorithm is 0.71 (95% CI 0.64–0.75) and for Glucksberg grades is 0.57 (0.55–0.68). GVHD = graft-versus-host disease.
better visualise this difference in AIC, we determined the hazard ratios (HR) for non-relapse mortality in univariate models for both staging systems using moderate GVHD (Ann Arbor 2 or Glucksberg II) as the reference group (appendix p 10). We then fit a multivariate model with simultaneous adjustment for both Ann Arbor score and Glucksberg grades. Patients with Ann Arbor score 3 had significantly higher risk for non-relapse mortality than patients with Ann Arbor 2 (figure 4A, p=0.0048) and patients with Ann Arbor 1 have significantly less risk (p=0.0020) than patients with Ann Arbor 2. By contrast, the 95% CIs for the HRs of the Glucksberg grades encompass 1.0, showing an absence 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).

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.25–31 With Ann Arbor 2 as a reference, we noted that Ann Arbor 1 predicted a lower risk of non-relapse mortality (range 0.16–0.32) and Ann Arbor 3 a higher risk of non-relapse mortality (1.4–2.9), regardless of the presence of these clinical risk factors (appendix p 11). Indeed, patients with HLA-mismatched donors were significantly more likely to have Ann Arbor 3 GVHD in all three datasets (appendix pp 12–14). Regardless of the presence of these clinical risk factors, we noted that Ann Arbor 1 predicted a lower risk of non-relapse mortality (range 0.16–0.32) and Ann Arbor 3 a higher risk of non-relapse mortality (1.4–2.9), with Ann Arbor 2 as a reference.

Discussion

Maximum 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 our study, we have developed and validated an algorithm using biomarkers that define three GVHD severity scores, each with a distinct risk of non-relapse mortality. We observed that our Ann Arbor GVHD scores defined risk across the full range of clinical presentations.

A higher Ann Arbor score predicted the development of gastrointestinal GVHD in patients who presented without gastrointestinal 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 gastrointestinal tract. It is therefore of great interest that the three biomarkers included in this algorithm (TNFR1, ST2, and REG3α) possess biological relevance to gastrointestinal GVHD. TNFR1, a surrogate for TNFa, is produced by T cells and monocytes and amplifies gastrointestinal injury.19,20 TNFa regulates ST2 that, together with its ligand interleukin 33 (a member of the interleukin 1 family), affects inflammatory bowel disease activity.19 REG3α, which we previously validated as a gastrointestinal GVHD specific biomarker,19 is produced primarily by Paneth cells and protects the gastrointestinal epithelium from infectious damage.19 The concentrations of these biomarkers at GVHD onset seem to reflect gastrointestinal tract disease activity that does not correlate with the severity of gastrointestinal symptoms at that time.

An important strength of our 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, about half of patients who presented with rashes of less than 50% body surface area (Glucksberg grade 1)—ie, 20% of all patients, never required treatment with systemic steroids.

In the multicentre BMTCTN dataset, all patients received treatment with systemic glucocorticoids and an experimental drug, but the algorithm correctly identified patients at low risk of non-relapse mortality. Previous studies established correlations between either individual GVHD biomarkers or their combinations and clinical outcomes, but they were not consistent among different clinical centres (panel).25–27 The biomarker algorithm developed in our study advances the previous work, but important limitations remain. First, although the algorithm predicts outcomes better than clinical symptoms, it still has fairly 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 shown.

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 (eg, voluminous diarrhoea) from a trial of an investigational drug. Such patients who are likely to respond to standard therapy benefit by avoiding exposure to the risks of an experimental drug, 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 restrict exposure to lengthy glucocorticoid regimens. A high score (about 23% of all GVHD) could be used as an 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 could lend confidence to the diagnosis of GVHD when a biopsy is equivocal, and a low score in a patient with a limited rash might 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 grading systems of GVHD, some studies show worse survival for patients with GVHD after an SCT from an unrelated volunteer donor.12,13 Patients with HLA-mismatched donors were significantly more likely to have Ann Arbor 3 GVHD in all three datasets. Possibly, the incorporation of such a clinical characteristic, or even the nature of the GVHD symptoms at their onset, might improve the algorithm’s predictive power, and in the training and test sets, organ-specific Glucksberg grade was also correlated with durable response. The algorithm has also not yet been adequately assessed 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. The use of an algorithm at serial timepoints could 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 SCT patients.

Contributors
JEL, TMB, and JLMF were responsible for the study design, data analysis and wrote the report. AT’ and HM did the laboratory analyses. JEL, ACH, and EH collected the data. All authors interpreted the data and contributed to writing the report.

Declaration of interests
JEL, TB, and JLMF are coinventors on a patent for the use of GVHD biomarkers. All other authors declare no competing interests.

Acknowledgments
This study was supported by grant numbers P01CA039542, R21CA137459, and P30CA046592 from the National Cancer Institute, number U10HL092994 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 by contributions from Eisai Inc, Hospira Inc, Roche Laboratories Inc, and ImmuneX Corporation, a wholly owned subsidiary of Amgen Inc. This report was prepared with BMTCTN 0302 and BMTCTN 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 BMTCTN 0302 or 0802 protocol teams or the NIH.

References


