An acute graft-versus-host disease activity index to predict survival after hematopoietic cell transplantation with myeloablative conditioning regimens

Wendy M. Leisenring, Paul J. Martin, Effie W. Petersdorf, Anne E. Regan, Nada Aboulhosn, Jean M. Stern, Saundra N. Aker, Raymond C. Salazar, and George B. McDonald

Algorithms for grading acute graft-versus-host disease (GVHD) are inaccurate in assessing mortality risk. We developed a method to predict mortality by using data from 386 patients with acute GVHD. From the onset of GVHD to day 100, GVHD manifestations were scored for the skin, liver, and upper and lower gastrointestinal tract, and data were recorded for immunosuppressive treatment, performance, and fever. Logistic regression models predicting nonrelapse mortality (NRM) at day 200 were developed with data from 193 randomly selected patients and then validated in the remaining 193 patients. Clinical parameters were grouped to optimize predictive accuracy measured as the area under a receiver-operator characteristic (ROC) curve. The optimal model included the total serum bilirubin concentration, oral intake, need for treatment with prednisone, and performance score. When the overall burden of GVHD was measured by using average Acute GVHD Activity Index (aGVHDAI) scores for each patient in training and validation data sets, areas under ROC curves were 0.87 and 0.85, respectively. Contour lines were generated to reflect the predicted NRM at day 200 as a function of current aGVHDAI scores. These results demonstrate that clinical manifestations of GVHD severity can be used to accurately predict the risk of NRM in real time. (Blood. 2006;108:749-755)

Introduction

Acute graft-versus-host disease (GVHD) is a common complication of allogeneic hematopoietic cell transplantation (HCT), affecting about 60% of recipients of allogeneic donor cells following myeloablative conditioning regimens.[1,2] Grading of GVHD can serve a variety of purposes, including retrospective assessment of peak severity, real-time assessment of severity at prespecified time points, determination of the need for treatment, assessment of treatment response, prognostication for survival, and evaluation of new methods to prevent GVHD in prospective studies. The most widely used grading system for grading acute GVHD represents variations of criteria originally proposed by Glucksberg et al[3] in 1974 on the basis of clinical intuition.[4] Variations of the Glucksberg system have been published to improve its utility for specific purposes.[5,6]

Although these grading systems have descriptive validity and a general relationship to outcome, several problems hamper the application of current grading systems for the purpose of predicting outcomes among patients with acute GVHD: (1) Relation of disease severity in skin, gut, and liver to outcome was never evidence-based, but instead reflected the judgment of experienced clinicians.[3] (2) Assignment of a peak GVHD score is done retrospectively; clinicians cannot use the current grading system for peak score in real-time. (3) The systems do not account for the time to response after treatment. Thus, patients whose symptoms resolve completely after a short course of immunosuppressive therapy may be scored identically to patients who require months of high-dose immunosuppressive drug therapy to control symptoms. (4) Significant interobserver errors exist in the current grading systems, largely because of subjective biases.[7-10] (5) Assignment of grade IV GVHD is often used descriptively to indicate that GVHD caused a patient's death, irrespective of the severity of symptoms. In this situation, the grading reflects the outcome and cannot be used to predict the outcome. Indeed, neither of the grading algorithms described by Przepiorka et al[4] or Rowlings et al[5] performs well as a prognostic tool, because neither explains much of the variation in either early or late survival.[10]

We have developed a new system for assessing the severity of acute GVHD that scores GVHD activity at 10-day intervals to day 100 following transplantation. This Acute GVHD Activity Index (aGVHDAI) is scaled from 0 to 100, with the intent of providing clinicians with a means of predicting outcome in real-time on the basis of current GVHD activity, and providing investigators with an evidence-based measure of the burden of acute GVHD over time. Our aGVHDAI was designed to predict day 200 nonrelapse mortality (NRM). To develop our activity index, 386 patients with chronic myeloid leukemia (CML) who received allogeneic hematopoietic cell transplants from unrelated donors were randomly divided into a training data set and a validation data set. Signs and symptoms of acute GVHD were recorded to day 100; a model predicting day-200 NRM was optimized using receiver-operator characteristic (ROC) curves, and the final model was applied to the validation set of patients to assess the reproducibility of the

From the Clinical Research Division, Fred Hutchinson Cancer Research Center and the University of Washington School of Medicine, Seattle.

Submitted January 20, 2006; accepted March 1, 2006. Prepublished online as *Blood* First Edition Paper, March 14, 2006; DOI 10.1182/blood-2006-01-0254.

Supported by grants CA 18029 and CA 15704 from the National Institutes of Health, National Cancer Institute.

An Inside *Blood* analysis of this article appears at the front of this issue.

Correspondence: George B. McDonald, Gastroenterology/Hepatology Section (D2-190), Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 U.S.C. section 1734.

aGVHDAI in predicting mortality. We then prepared a graphic that allows clinicians to predict day-200 NRM on the basis of the calculated aGVHDAI at points in time up to day 100 after HCT.

Patients, materials, and methods

Patient selection

One of us (E.W.P.) randomly selected 411 patients with CML who received a first allogeneic hematopoietic cell transplant from unrelated donors with 0, 1, or 2 or more allele disparities at HLA-A, -B, -C, -DRB1, -DQB1 as determined by sequence-based methods.[11] All patients received conditioning therapy with cyclophosphamide (CY, 120 mg/kg) and total body irradiation (TBI, median dose 12 Gy), followed by bone marrow infusion, between 1987 and 1999 at the Fred Hutchinson Cancer Research Center (FHCRC).[12,13] All patients received GVHD prophylaxis with a calcineurin inhibitor (cyclosporine in almost all cases) plus methotrexate.[14] In addition, some patients received anti–T-cell monoclonal antibodies as prophylaxis (Table 1). Prophylaxis against infections during this time period varied, as described previously.[1] These patients had been retrospectively graded for acute GVHD by one observer (P.J.M.), but these grades were not revealed to those compiling the aGVHDAI until completion of the validation process. Among the 411 patients in the cohort, 386 developed grade I to IV acute GVHD and were included in this analysis. Review of patient charts and research records was carried out under the aegis of a protocol approved by the Institutional Review Board.

Assessment of the severity of GVHD by conventional grading

A date of onset of acute GVHD and a descriptive grade for its severity had been previously determined by one observer (P.J.M.), using methods as described by Przepiorka et al, with minor modifications.[1,4] Biopsies of skin, gut, and liver were done in 89%, 27%, and 2% of patients, respectively; overall, 93% of subjects had biopsies. The diagnosis of acute GVHD of the skin was established by the development of a characteristic erythematous or morbilliform rash first appearing generally between 10 and 70 days after transplantation. Stage 1 skin GVHD indicated rash involving less than 25% of body surface; stage 2, 25% to 50% of body surface; stage 3, more than 50% of body surface; and stage 4, bullous lesions. The diagnosis of acute GVHD of the liver was established by the development of hyperbilirubinemia, generally in the presence of a rash that is diagnostic of acute GVHD. Stage 1 liver GVHD indicated a peak total serum bilirubin concentration of 34.2 to 49.59 μM (2.0-2.9 mg/dL); stage 2, 51.3 to 100.9 μM (3.0-5.9 mg/dL); stage 3, 102.6 to 254.8 μM (6.0-14.9 mg/dL); and stage 4, 256.5 μM or higher (≥ 15.0 mg/dL). In cases where another cause of hyperbilirubinemia antedated the onset of rash, the liver score was decreased by one stage. The diagnosis of acute GVHD of the gut was established by development of upper gastrointestinal symptoms such as anorexia, nausea and vomiting, or diarrhea. Stage 1 gut GVHD indicated upper gastrointestinal symptoms or diarrhea with peak daily stool volume less than 1000 mL/d; stage 2, 1000 to 1499; stage 3, 1500 to 1999; and stage 4, 2000 mL/day or more or the presence of severe abdominal cramping, bleeding, or ileus caused by GVHD. In cases where peak gastrointestinal symptom severity was exacerbated by a cause other than GVHD, the gut score was decreased by one stage.

Overall grade I GVHD denoted stage 1 to 2 skin involvement with no liver or gut involvement, indicating that the prophylactic immunosuppressive regimen administered after the transplantation was not sufficient to prevent all manifestations of acute GVHD, but in most cases, the disease resolved spontaneously without treatment. Overall grade II GVHD denoted stage 3 skin involvement or stage 1 liver or gut involvement, indicating that the prophylactic immunosuppressive regimen administered after the transplantation was not sufficient to prevent manifestations of acute GVHD, but glucocorticoid treatment after the onset of GVHD was generally sufficient to control the disease. Overall grade III GVHD denoted stage 4 skin involvement or stage 2 to 4 liver or gut involvement without GVHD as a major contributing cause of death, indicating that the prophylactic immunosuppressive regimen was not sufficient to prevent manifestations of acute GVHD and that additional treatment after the onset of GVHD did not readily control the disease. Overall grade IV GVHD denoted stage 4 skin involvement or stage 2 to 4 liver or gut involvement with GVHD as a major contributing cause of death, indicating that the disease was resistant to both the prophylactic immunosuppressive regimen and any additional treatment after the onset of the disease.

Table 1. Patient characteristics

Characteristic	Training data set	Validation data set
Median age at transplantation, y (range)	36.7 (0.7-55.1)	35.8 (6.4-54.6)
No. younger than 18 y (%)	13 (7)	12 (6)
CML phase at time of transplantation, no. (%)		
Chronic	130 (67)	142 (74)
Accelerated	49 (25)	38 (20)
Blast crisis	12 (6)	13 (7)
Juvenile	2 (1)	0 (0)
HLA allele match status, no. (%)		
10 of 10 matched	91 (47)	92 (48)
9 of 10 matched	46 (24)	54 (28)
8 or fewer of 10 matched	56 (29)	47 (24)
No. male patients (%)	112 (58)	110 (57)
GVHD prophylaxis, no. (%)		
Calcineurin inhibitor + MTX*	173 (90)	168 (87)
Calcineurin inhibitor + MTX + other†	20 (10)	25 (13)

For training data set and validation data set, n = 193 each.

MTX indicates methotrexate.

*Tacrolimus was the calcineurin inhibitor for one patient from each of the training and validation data sets, respectively; the remaining patients received cyclosporine.

†Other refers to a CD25-specific immunotoxin for 4 and 7 patients and to a humanized CD25-specific antibody for 16 and 18 patients from the training and validation data sets, respectively.

Parameters of acute GVHD used for creation of an activity index

We assessed the severity of 7 GVHD parameters in each 10-day time period from day 0 to day 100, starting with the period in which the onset of GVHD occurred. Most of the raw data had been prospectively collected at the time of transplantation by our Clinical Nutrition, Nursing, and Pharmacy sections, for example, daily oral caloric intake, diarrheal volumes, body temperature, and medications administered. Physicians caring for patients wrote daily progress notes and dictated interim summaries according to a uniform format, enabling us to assign patient performance scores, as described.

We assigned a unique letter score for each degree of abnormality within each parameter, except for liver scoring, which used total serum bilirubin values. The purpose of using letters rather than numbers was to avoid the bias of numeric values and to facilitate future research regarding the relative accuracy of various weightings of each letter score.

Skin GVHD was scored by the nature and extent of abnormality during each 10-day period, as the worst score during this time. This scoring system excluded rashes of known viral or other origin such as varicella zoster virus, herpes simplex virus, and drug-related rashes. The letters "a, b, and c" described a macular erythematous rash involving less than one third, one third to two thirds, and more than two thirds of the body surface, respectively. The letters "d, e, and f" and "g, h, and k" similarly described maculopapular rashes and desquamation, respectively.

Liver GVHD was scored by the maximum total serum bilirubin concentration in micromolars per liter (μM) (or milligrams per deciliter [mg/dL]) during each 10-day interval after the onset of acute GVHD, rounded to the nearest whole number. If the total serum bilirubin concentration was less than 17.1 μM (< 1 mg/dL), then a score of 0 was recorded. Total serum bilirubin concentrations more than 256.5 μM (> 15 mg/dL) were recorded as 15, because previous studies have shown that bilirubin

concentrations more than 256.6 μM (> 15 mg/dL) do not predict a worse prognosis than values of 15.[15] No adjustments were made for the presence of concurrent liver dysfunction caused by diseases other than GVHD.

Upper gut GVHD was scored by reviewing data on daily caloric intake, as recorded by the FHCRC Clinical Nutrition Service during the posttransplantation period, and by clinical symptoms described in daily progress notes. These data compare the lowest oral caloric intake per day with the therapeutic caloric goal for that day. A score of 0 was given if oral calories were 90% or more of patient requirements and the patient had a normal appetite during the 10-day period. A letter score of "m" described oral calories 70% to 90% of patient requirements and mild anorexia or nausea; "n," oral calories 40% to 70% of patient requirements and moderate anorexia, nausea, and vomiting; and "p," oral calories less than 40% of patient requirements with poorly controlled symptoms.[16,17]

Lower gut GVHD was scored by reviewing daily stool volumes as recorded in chart notes by the Nursing and Clinical Nutrition services and by descriptions of diarrhea in daily progress notes. The score for each 10-day period reflected the highest daily stool output during this interval. A score of 0 was given if the largest stool volume in the interval was less than 200 mL/day and there was no record of diarrhea. A letter score of "s" described stool volumes 200 to 300 mL/day and a history of loose stools; "t", stool volumes 300 to 500 mL/day and frequent diarrhea; and "u," stool volumes more than 500 mL/day and unrelenting diarrhea.

Immunosuppressive therapy during each 10-day period was scored as the most intensive treatment prescribed to control GVHD. A score of 0 denoted no immunosuppressive drugs. A letter score of "y" meant that only prophylactic drugs such as cyclosporine or tacrolimus were prescribed during that period. Letters "w" and "x" meant that prednisone was prescribed at doses of less than 1 mg/kg/day and 1 to 2 mg/kg/day, respectively. Letter "y" meant that the patient had received secondary therapy to control GVHD during the 10-day period in question, for example, prednisone more than 2 mg/kg/day, antithymocyte globulin, sirolimus, or anti–T-cell monoclonal antibodies.

Fever scoring referred to the highest fever documented during each 10-day interval, with 0 assigned to patients whose temperature was normal (< 37.2°C) throughout this period, the letters "AA" for any recorded temperature from 37.2 to 38.5°C; and "BB" for any recorded temperature over 38.5°C.

Performance status during each 10-day period was scored using a modification of the Eastern Cooperative Oncology Group system.[18] Interim and discharge summaries from the FHCRC electronic database provided dates of admission and discharge between inpatient and outpatient status. Performance scores were based on narratives by patient care providers, where 0 described outpatients who felt well and fully active; "CC," outpatients who were unwell and fully ambulatory, but limited in strenuous activity; "DD," inpatients who were unwell, ambulatory, capable of self-care, and spending more than 50% of waking hours out of bed; "EE," inpatients who were unwell, limited in their self-care, and spending more than 50% of time in bed; and "FF," inpatients who were completely disabled, providing no self-care, and confined to bed.

Statistical modeling methods

The cohort of 386 patients was randomly divided into a training data set of 193 patients (64 of whom died by day 200) and a validation data set of 193 patients (62 of whom died by day 200). Using the training data set, logistic regression models were fitted to the parameters of interest to predict NRM by day 200. The coefficients for each factor in the model were then used to determine the weight for that factor in constructing the aGVHDAI. Rather than simply including the most significant factors, we were interested in optimizing the area under an ROC curve[19] for a scoring algorithm that best discriminated between patients who died without prior recurrent malignancy by day 200 and those who survived to day 200. ROC curves depict the trade-off in true-positive versus false-positive rates as the cut point for defining "positive" and "negative" is shifted along the full spectrum of aGVHDAI values. An area under the curve of 1.0 would indicate a perfect test, whereas 0.5 would represent a noninformative test. The area under the curve was computed using the trapezoidal rule (Stata software version 8.2, StataCorp, College Station, TX). Throughout the modeling exercise, GVHD parameters were grouped in different ways to optimize the scoring system's predictive abilities, while using the fewest parameters possible.

Two basic models were developed, the first for evaluation of the burden of acute GVHD across time. For this model, we used average aGVHDAI scores over all 10-day intervals from the time of onset of GVHD to death, departure from our center, or day 100, whichever occurred first. These data were used for model fitting and calculation of coefficients that could be used as score weights. A second model was developed for real-time prediction of NRM by day 200. For this model, the current value of each factor from each interval was included in a generalized estimating equation formulation of the logistic regression model, to account for the correlated longitudinal data from each subject.[20] All coefficients were positive (ie, we observed increased risk of NRM for higher categories of all covariates). Scaling was carried out by dividing the coefficients by the sum of the coefficients (using only the highest category for factors with multiple levels) and multiplying this proportion by 100. Thus, the score could range from 0 to 100, with 100 being worst. Once the weights were obtained from each model, we were able to construct a single weighting system by averaging the weights for the 2 models, without marked change in the overall accuracy of the aGVHDAI. We found that, within a similar realm of weights, the difference in areas under the ROC curves rests on how the scores from each interval are used, either as an average over all intervals to summarize experience for each subject (research use), or as an individual value from each interval, to predict NRM by day 200 (clinical use). All of the modeling was carried out using the training data set. Once the final weighting system was determined, the aGVHDAI was independently evaluated in the validation data set.

For each patient in the combined data sets (N = 386), the average GVHDAI was compared to the peak GVHD grade that had been recorded independently. We also examined the accuracy of GVHD grading according to Przepiorka et al[4] and Rowlings et al[5] for predicting mortality at day 200. In this analysis, overall grades were derived entirely from organ scores, without considering mortality. We then evaluated the true- and false-positive rates when these overall grades were dichotomized at each possible level for predicting NRM in the validation data set (n = 193). A nonparametric method was used to evaluate the areas under ROC curves defined by these scoring systems as compared to the area under the ROC curve for the average aGVHDAI in the validation data set.[21]

To illustrate the predictive value of the current value of aGVHDAI from different time intervals, we fitted a smooth cubic spline curve in logistic regression models for the probability of NRM by day 200 as a function of the calculated scores within different time intervals.

Results

Demographics of the study cohort

Subjects were selected on the basis of underlying disease (all had CML), donor relation (unrelated), and conditioning regimen (CY/TBI). The distribution of transplants matched for 10 of 10, 9 of 10, and 8 or fewer of 10 HLA-A, -B, -C, -DRB1, -DQB1 alleles was similar between the training data set and validation data set. Furthermore, the distribution of HLA mismatches in these 2 sets of transplants was similar to the overall HLA mismatch status of our unrelated donor transplantation population.[22] Other characteristics of patients in the training and validation data sets appeared similar (Table 1).

Development of an aGVHDAI in the training data set of 193 patients

Throughout the modeling exercise, the aim was to group the parameters of acute GVHD in different ways to optimize the scoring system's predictive qualities, while using the fewest parameters possible. For the optimal logistic regression model,

each patient's experience was summarized as the average of all 10-day interval aGVHDAI scores. Only parameters reflecting liver dysfunction, oral caloric intake, need for prednisone or secondary immunosuppressive therapy, and performance score entered the optimal model (Table 2).

A similar model to determine weights for use when predicting NRM based on the score in each time interval was fitted. Because the 2 models gave similar weights (data not shown), we elected to use a common weighting system based on the average from these 2 models (Table 2). There was little loss in the area under the ROC curve for either use of the score with this simplification (data not shown). There was no significant improvement in the area under the ROC curve when we added age as a continuous variable and HLA disparity (10 of 10 versus fewer than 10 of 10) to the optimal model for predicting NRM ($P = .51$).

Mild jaundice (total serum bilirubin concentration 34.2-68.4 μM [2-4 mg/dL]) conferred an increased risk of NRM, and deeper jaundice (≥ 85.5 μM [≥ 5 mg/dL]) a greater risk, but extreme bilirubin concentrations did not enter the model. Neither the severity of skin GVHD nor the amount of diarrhea entered the model. Uncontrolled symptoms of anorexia, nausea, and vomiting, with oral caloric intake less than 40% of caloric requirements, but not less severe symptoms, carried an increased risk of mortality. Any dose of prednisone, or more intensive immunosuppressive therapy, was associated with an increased risk of NRM, as was poor performance status (Table 2). Figure 1A-B shows the ROC curves with the weights shown in Table 2 for the training data set; the area under the curve is 0.87 when prediction of day-200 NRM is based on average aGVHDAI scores for each patient (Figure 1A) and 0.76 when based on scores for each interval (Figure 1B).

Validation of the aGVHDAI in an independent data set of 193 patients

ROC curves that resulted from the application of the weights shown in Table 2 to calculate the aGVHDAI in the randomly selected validation data set are shown in Figure 1. Panels C and D of Figure 1 show the ROC curves for the validation data set; the area under the curve is 0.85 when prediction of day-200 NRM is based on average aGVHDAI scores for each patient (Figure 1C) and 0.74 when based on scores for each interval (Figure 1D).

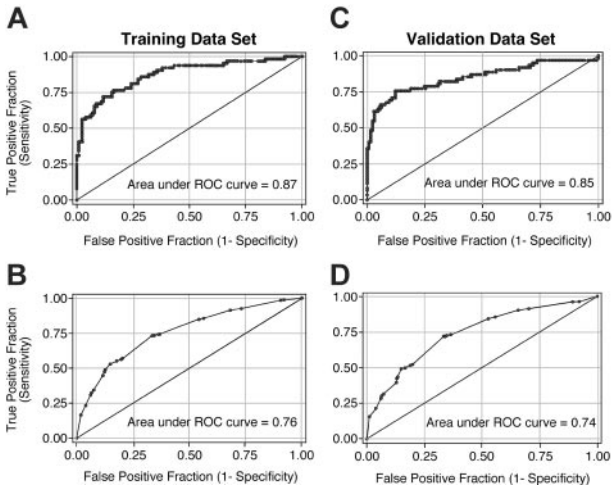

Figure 1. ROC curves illustrating the accuracy of the aGVHDAI across the full range of threshold values. All graphics used the common weighting system shown in Table 2. Panels A and C illustrate ROC curves when average aGVHDAI scores for each patient were used to predict day-200 NRM in the training data set and validation data set, respectively. Panels B and D illustrate ROC curves when aGVHDAI scores for each time interval were used to predict day-200 NRM in the training data set and validation data set, respectively.

Comparison of the average aGVHDAI values with conventional GVHD grading

Conventional GVHD grades were compared to the average aGVHDAI values for all patients in the original cohort of 386 patients. Results showed overlap of average aGVHDAI values among all peak grades of GVHD. Among the 386 subjects in the training and validation data sets classified as having developed acute GVHD, including 29 (8%) grade IV, 123 (32%) grade III, 233 (60%) grade II, and 1 (0.3%) grade I acute GVHD, average GVHDAI values ranged from 2.2 to 100; median GVHDAI values for grades II, III, and IV GVHD were 36.9, 54.1, and 81.6, respectively (Figure 2). One subject with grade I acute GVHD had an average aGVHDAI score of 9.3. Average aGVHDAI scores showed considerable overlap among patients with peak GVHD grades II, III, and IV. For example, the maximum average aGVHDAI score in the grade II acute GVHD group was 91.5.

Table 2. Logistic regression model for the aGVHDAI scaled to yield score values that range from 0 to 100

Factor, scoring level	Average aGVHDAI model			Current interval model			Final scaled weight factors
	Coefficient	*P*	95% CI	Coefficient	*P*	95% CI	
Liver							
Total serum bilirubin 2-4 mg/dL	1.34	.07	−0.11, 2.80	0.67	.004	0.21, 1.12	16
Total serum bilirubin greater than or equal to 5 mg/dL	1.95	.053	−0.02, 3.93	1.32	< .001	0.65, 1.99	26
Upper gut							
Oral caloric intake less than 40% of requirements, with poorly controlled anorexia, nausea, and vomiting	2.64	.008	0.68, 4.61	0.41	.08	−0.05, 0.87	20
Immunosuppressive therapy							
Any prednisone dose or secondary therapy for GVHD	1.13	.21	−0.62, 2.90	0.84	.02	0.12, 1.56	17
Performance							
Unwell, ambulatory, limited in strenuous activity, capable of self-care, and spending more than 50% of waking hours out of bed (scored as CC or DD; see "Patients, materials, and methods")	1.81	.03	0.15, 3.48	0.80	.006	0.24, 2.48	20
Unwell, limited in self-care, and spending more than 50% of time in bed, or worse (scored as EE or FF; see "Patients, materials, and methods")	3.11	.01	0.91, 5.31	1.65	< .001	0.91, 5.31	37

Final scaled weight factors for the aGVHDAI are averages of the weights from the average and current interval models.
To convert bilirubin from milligrams per deciliter to micromoles per liter, multiply milligrams per deciliter by 17.1.

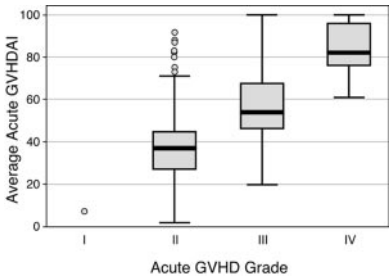

Figure 2. Distribution of values for average aGVHDAI, according to the modified GVHD grading system[4] in 386 patients. The boxes display values for the average aGVHDAI from the 25th to the 75th percentile; the bars within the boxes display the median value; and the vertical bars display the upper and lower adjacent values.[23] Data points outside this range are plotted as individual circles.

Average aGVHDAI scores were below this level in 18 of the 29 (62%) subjects with grade IV GVHD and in 114 of the 123 (92%) subjects with grade III GVHD. The mean aGVHDAI values within each peak GVHD grade, however, were statistically significantly different from one another ($P < .001$ for all comparisons).

We derived overall grades from organ scores according to Przepiorka et al[4] and Rowlings et al[5] without considering mortality and evaluated the true- and false-positive rates associated with these overall grades dichotomized at each level (II-IV and B-D, respectively) to predict day-200 NRM in the validation data set (Table 3).

The points associated with each pair of true positive and false positive rates are shown in Figure 3 along with the ROC curve for the average aGVHDAI in the validation data set. The area under the ROC curve for each of these 2 peak value-based systems is significantly lower than the area under the ROC curve for the average aGVHDAI ($P < .001$ for both comparisons).

An aGVHDAI graphic for clinical use in real time

The validation data set was used to generate contour lines that reflect the predicted NRM at day 200 as a function of the current aGVHDAI scores during each of 3 time intervals after transplantation (Figure 4).

Discussion

This aGVHDAI addresses the major shortcomings of the current acute GVHD grading systems. The aGVHDAI is evidence-based and was developed in one randomly selected cohort of patients and validated in an independent cohort. The aGVHDAI does not depend on peak values and can be used as either a current or average indicator of day-200 NRM risk. Unlike peak GVHD

Table 3. Dichotomized GVHD grades as predictors of day-200 NRM in the validation data set of 193 patients

Grading algorithm	True positive, %	False positive, %
Przepiorka et al[4]		
I vs II-IV	100	99
I-II vs III-IV	61	30
I-III vs IV	19	2
International Bone Marrow Transplant Registry[5,10]		
A vs B-D	100	99
A-B vs C-D	81	74
A-C vs D	21	2

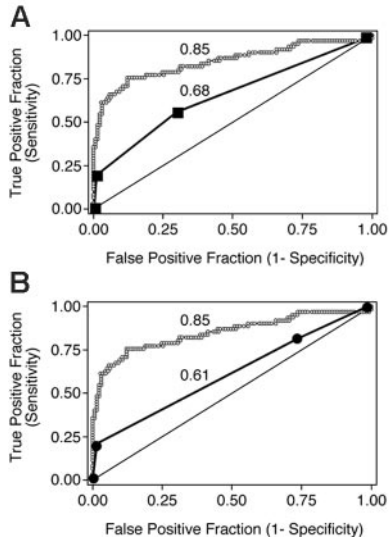

Figure 3. ROC curves depicting the accuracy of 3 GVHD scoring systems for predicting day-200 NRM in 193 patients from the validation data set. The average aGVHDAI line (small ○) is identical to that in Figure 1C. Data points for grading according to Przepiorka et al[4] and the International Bone Marrow Transplant Registry[5,10] are shown in panels A and B, respectively. The numbers adjacent to each ROC curve are the areas under the respective curves.

grades, the average aGVHDAI has the advantage of accounting for the duration of GVHD manifestations. A substantial number of patients whose peak disease severity gave them only a grade II designation died before day 200 because their burden of disease over time gave them the same or greater mortality risk as for patients with a greater peak severity of disease and shorter duration. Conversely, some patients graded as III or IV GVHD had short-lived symptoms (likely a result of prompt and effective therapy) and an average aGVHDAI over time that placed them at lower risk for mortality by day 200. The requirement for continuing immunosuppressive therapy to control signs and symptoms of acute GVHD serves as an important component of the aGVHDAI. This component addresses a shortcoming of the current grading systems, which do not account for response or lack of response after treatment.[3-6] Unlike current descriptive scoring systems that include death as a criterion for grade IV GVHD, the aGVHDAI does not consider death, since this scoring system was developed for the specific purpose of predicting the risk of death within the first 200 days after HCT. The aGVHDAI is a more accurate predictor of NRM than the grading systems according to Przepiorka et al[4] or the International Bone Marrow Transplant Registry.[5,10] Before this index can be widely used to guide clinical decision-making, however, it must be validated at other centers, in a wider range of patients. We do not envision this index as a fixed formula, but rather as the beginning of a process of providing an evidence-based method for determining prognosis in patients with acute GVHD, to be recalibrated over time to account for changes in practice.

Although the components that were considered for the logistic regression models in this study were complex, the final elements of the aGVHDAI are relatively simple. For example, it is not necessary to measure stool volumes or estimate body surface area of skin involved with GVHD because neither parameter entered the model. And because only the most extreme upper gut symptoms entered the model, scoring the upper gut is straightforward if a patient had little oral intake while constantly vomiting during a day

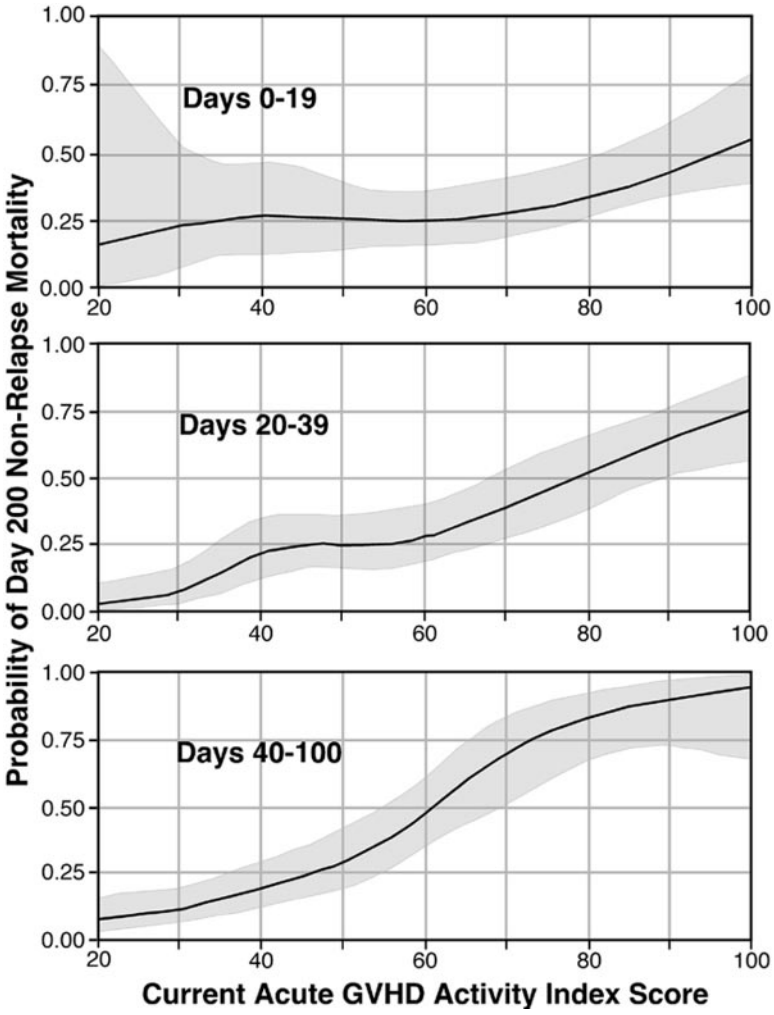

Figure 4. Predicted NRM by day 200 as a function of current aGVHDAI scores at different points in time after transplantation. The top panel shows this relationship for days 0 to 19, the middle panel for days 20 to 39, and the bottom panel for days 40 to 100. Within each panel, the solid line represents the predicted NRM by day 200 across a range of values for current aGVHDAI scores, and the shaded areas represent 95% confidence intervals.

of the interval in question. Calculation of oral intake as a percent of caloric requirements may be needed if patients have only anorexia as a manifestation of GVHD and the oral intake is uncertain.[17] Because liver dysfunction is a well-recognized contributor to mortality after allogeneic transplantation,[15,24] it was not surprising to find total serum bilirubin parameters as components of the aGVHDAI. The doses of immunosuppressive drugs needed to control GVHD did not improve the correlation with the risk of NRM over a binary assessment as to whether any therapeutic dose of immunosuppressive drugs was needed to control symptoms.

We envision that the aGVHDAI could be used for 2 purposes. One would be to assess the risk of a fatal outcome in a patient with acute GVHD in real time, using the graphic in Figure 4. Patients at high risk should be considered for more aggressive treatment or for entry into clinical intervention trials, thereby allowing patients to be stratified according to the risk of mortality. Patients at low risk may be spared the morbidity of unnecessary continued high-dose immunosuppressive therapy. A second purpose would be to measure the burden of acute GVHD across time, as a research tool to assess the efficacy of prevention and treatment in clinical trials. Because standard GVHD grading is an inaccurate predictor of survival,[10] clinical trials that use such grading as the primary end point of efficacy might overestimate the effectiveness of a given intervention or might fail to detect a potential benefit due to lack of sensitivity.

The aGVHDAI described here was inspired by the Crohn's Disease Activity Index (CDAI), a system for scoring the severity of an intestinal disorder that has skin, liver, and extraintestinal manifestations, similar to those of acute GVHD.[25] Whereas the CDAI was developed by using an expert panel's definitions of severity of disease, the aGVHDAI used day-200 NRM for validation. Validation according to mortality avoids using manifestations of disease to predict disease severity. On the other hand, using day-200 mortality to capture the severity of acute GVHD across time has the disadvantage that mortality reflects not only GVHD but also residual toxic effects of the conditioning regimen, immunosuppressive drug treatment for GVHD, and infections. Because GVHD does not develop in a vacuum, we think it reasonable to score the disease by examining the totality of its effects, including those of the drugs used to treat the disease and the emergent infections that arise as a consequence. Like the CDAI, which was recalibrated by an expert panel to be sure that the components of the index and their weightings had not changed over time,[26] the aGVHDAI will have to be recalibrated at time intervals and also examined in cohorts of patients who were not represented in the panel of 386 patients from which the activity index was derived.

In particular, it would be of interest to determine how well the aGVHDAI might apply to a larger cohort of pediatric patients and to those who have HCT after nonmyeloablative conditioning regimens.

In summary, we have developed and validated a disease activity index that measures the burden of acute GVHD across time with day-200 mortality as the end point. The aGVHDAI is easy to calculate at 10-day intervals from the highest total serum bilirubin concentration, severity of upper gut symptoms, need for immunosuppressive drugs, and performance score. This index should be of use both in clinical practice and as a tool in investigation of methods to prevent and treat acute GVHD.

Acknowledgment

We are grateful to the nurses, dietitians, and physicians who cared for the patients in this study for their meticulous attention to detail and complete medical records.

References

1. Martin PJ, McDonald GB, Sanders JE, et al. Increasingly frequent diagnosis of acute graft-versus-host disease after allogeneic hematopoietic cell transplantation. Biol Blood Marrow Transplant. 2004;10:320-327.
2. Sullivan KM. Graft versus host disease. In: Blume K, Forman SJ, Appelbaum F, eds. Thomas' Hematopoietic Cell Transplantation, ed 3. Malden, MA: Blackwell Publishing; 2004:635-664.
3. Glucksberg H, Storb R, Fefer A, et al. Clinical manifestations of graft-versus-host disease in human recipients of marrow from HL-A-matched sibling donors. Transplantation. 1974;18:295-304.
4. Przepiorka D, Weisdorf D, Martin P, et al. 1994 Consensus Conference on Acute GVHD Grading. Bone Marrow Transplant. 1995;15:825-828.
5. Rowlings PA, Przepiorka D, Klein JP, et al. IBMTR Severity Index for grading acute graft-versus-host disease: retrospective comparison with Glucksberg grade. Br J Haematol. 1997;97:855-864.
6. Parrish RS, Hazlett LJ, Bridges KD, Henslee-Downey PJ. A multivariate approach for assessing severity of acute graft-versus-host disease in bone marrow transplantation. Stat Med. 1999;18:423-440.
7. Martin P, Nash R, Sanders J, et al. Reproducibility in retrospective grading of acute graft-versus-host disease after allogeneic marrow transplantation. Bone Marrow Transplant. 1998;21:273-279.
8. Al-Ghamdi H, Leisenring W, Bensinger WI, et al. A proposed objective way to assess results of randomized prospective clinical trials with acute graft-versus-host disease as an outcome of interest. Br J Haematol. 2001;113:461-469.
9. Martin PJ, Schoch G, Gooley T, et al. Methods for assessment of graft-versus-host disease. Blood. 1998;92:3479-3481.
10. 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.
11. Petersdorf EW, Gooley TA, Anasetti C, et al. Optimizing outcome after unrelated marrow transplantation by comprehensive matching of HLA class I and II alleles in the donor and recipient. Blood. 1998;92:3515-3520.
12. Clift RA, Buckner CD, Appelbaum FR, et al. Allogeneic marrow transplantation in patients with chronic myeloid leukemia in the chronic phase: a randomized trial of two irradiation regimens. Blood. 1991;77:1660-1665.
13. Clift RA, Buckner CD, Thomas ED, et al. Marrow transplantation for chronic myeloid leukemia. A randomized study comparing cyclophosphamide and total body irradiation with busulfan and cyclophosphamide. Blood. 1994;84:2036-2043.
14. Storb R, Deeg HJ, Whitehead J, et al. Methotrexate and cyclosporine compared with cyclosporine alone for prophylaxis of acute graft versus host disease after marrow transplantation for leukemia. N Engl J Med. 1986;314:729-735.
15. Gooley TA, Rajvanshi P, Schoch HG, McDonald GB. Serum bilirubin levels and mortality after myeloablative allogeneic hematopoietic cell transplantation. Hepatology. 2005;41:345-352.
16. Weisdorf DJ, Snover DC, Haake R, et al. Acute upper gastrointestinal graft-versus-host disease: clinical significance and response to immunosuppressive therapy. Blood. 1990;76:624-629.
17. Wu D, Hockenbery DM, Brentnall TA, et al. Persistent nausea and anorexia after marrow transplantation: a prospective study of 78 patients. Transplantation. 1998;66:1319-1324.
18. Oken MM, Creech RH, Tormey DC, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol. 1982;5:649-655.
19. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978;8:283-298.
20. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13-22.
21. DeLong ER DD, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating curves: A non-parametric approach. Biometrics. 1988;44:837-845.
22. Petersdorf E, Anasetti C, Martin PJ, et al. Limits of HLA mismatching in unrelated hematopoietic cell transplantation. Blood. 2004;104:2976-2980.
23. Tukey J. Exploratory Data Analysis. Reading, MA: Addison-Wesley; 1977.
24. Martin PJ, Schoch G, Fisher L, et al. A retrospective analysis of therapy for acute graft-versus-host disease: initial treatment. Blood. 1990;76:1464-1472.
25. Best W, Becktel JM, Singleton JW, Kern F. Development of a Crohn's disease activity index: National Cooperative Crohn's Disease Study. Gastroenterology. 1976;70:439-444.
26. Best W, Becktel JM, Singleton JW, Kern F. Rederived values of the eight coefficients of Crohn's Disease Activity Index (CDAI). Gastroenterology. 1979;77:843-846.