Acute GVHD biomarkers predict long-term outcomes after GVHD diagnosis. Our group has validated the MAGIC algorithm which uses the concentrations of 2 GVHD biomarkers, ST2+REG3a, to predict lethal GVHD (defined as death from GVHD without relapse) [Hartwell 2017]. Other published biomarker combinations that predict GVHD outcomes include ST2+REG3a+TNFR1 [Levine 2015], ST2+TIM3 [Abu Zaid 2017], ST2+TNFR1 [McDonald 2015], TIM3+TNFR1+IL6 [McDonald 2015], as well as AREG alone [Holtan 2018]. It is not clear which biomarker combination best predicts lethal GVHD because these algorithms were developed using different patient cohorts with different endpoints. To answer this question, we compared the predictive accuracy of different biomarker combinations as well as these six specific combinations in the same patient cohort. We studied 522 patients with serum samples at GVHD diagnosis who were transplanted at 19 Mount Sinai Acute GVHD International Consortium (MAGIC) centers between January 1, 2004 and April 30, 2017. Patients were divided into training (n = 253) and validation (n = 269) sets; validation patients were transplanted after November 1, 2015 and had not previously been used to generate an algorithm. Biomarkers were measured by ELISA and log-transformed values were used to predict 1-year lethal GVHD by competing risk regression. Four of the 6 biomarkers (ST2, REG3a, TNFR1, and TIM3) independently predicted lethal GVHD in the training set. We then developed algorithms that predicted lethal GVHD using all possible combinations of these 4 biomarkers as well as the published combination of TIM3+TNFR1+IL6 and AREG alone. The best algorithms of 1-4 biomarkers for predicting lethal GVHD were REG3a alone, ST2+REG3a, ST2+REG3a+TNFR1, and ST2+REG3a+TNFR1+TIM3. While ST2+REG3a was the most accurate combination based on the lowest Akaike Information Criterion (AIC), the other published biomarker combinations produced similar AICs. Therefore, we used the independent validation set to compare all six published algorithms. We generated area under the receiver operating characteristic curves (AUC) for each algorithm (FIG 1). We next determined the threshold that maximized sensitivity and specificity for each algorithm and calculated the cumulative incidence of lethal GVHD in the resulting high and low risk strata (FIG 2). Highly similar results were obtained when the cumulative incidence of non-relapse mortality was used as the endpoint. In a validation set of previously unanalyzed patients, several biomarker combinations reproducibly stratified patients with GVHD for risk of death.
The best 4 algorithms, all of which included ST2, produced comparable outcomes. Adding TNFR1 to ST2+REG3a did not improve accuracy. The ST2+REG3a algorithm best identifies patients at onset of GVHD for high risk of lethal GVHD and NRM, and thus remains a standard for biomarker-based risk prediction.

Although 90% of pts developed symptoms at least once, physicians suspected GVHD in only 23% of cases and GVHD was considered the only etiology in 12%. Biopsies (bx) were performed in 40% of pts, most often for upper or lower GI symptoms. Bx results were equivocal in 11% although the range among centers was broad (0-45%). Equivocal results were most likely for skin bx and during the 2nd week after transplant. Systemic steroids were administered to 70% of pts at least once, often for non-GVHD reasons (41% of steroid courses). When GVHD was in the differential diagnosis, systemic steroids were started 81% of the time.

The ERC adjudicated 2,008 cases with symptoms suggestive of GVHD onset. In 12.3% of cases the ERC modified either the center reported diagnosis of GVHD (4.6%) or the automated generated severity grade (7.7%). GVHD was diagnosed in 1025 pts and categorized as confirmed (31%), probable (50%), or possible (19%). The proportion of definitive diagnoses (Confirmed or Negative) increased with longer time post-transplant (Fig 1). ERC adjudication substantially lowered the incidence of GVHD grade 2-4 from 30% as reported by centers to 23% (confirmed or probable) by ERC. An adjudicated diagnosis of GVHD strongly associated with 6 month NRM, adjusted for patient age, gender and HLA matching (HR = 2.0, p < 0.001). Adjudicated GVHD grade was also strongly associated with 6 month NRM (Grade 2, HR = 2.1, p < 0.001; Grade 3-4, HR = 7.1, p < 0.001).

In summary, the largest prospective observational study of GVHD outcomes demonstrated that symptoms in GVHD target organs are almost universal after transplant, their differential diagnosis is broad, tissue biopsies are infrequently done and their results are often equivocal, and steroids are often prescribed for reasons other than acute GVHD. These limitations undermine the current methods of diagnosis and reporting of acute GVHD on clinical trials. A robust reporting and adjudication system has significant value and may improve the chances of success of prospective clinical trials.

Figure 2. Cumulative incidence of lethal GVHD.

Figure 1. ERC confidence level of patients with new onset of GVHD symptoms.