Supplemental Materials

Predicting Acute Graft-versus-Host Disease Using Machine Learning and Longitudinal Vital Sign Data from Electronic Health Records

Details of vital sign features

For each of the six vital signs, four daily summary statistics were computed: mean, standard deviation, minimum and maximum. This transformed the six irregularly-spaced time-series into 24 regularly-spaced time series of length 10. Five trend features were then calculated on each of the 'summary' time series: average value, linear slope, sample entropy,[30] and the absolute value and angle of the first fast Fourier transform coefficient.[31] This produced 120 continuous variables describing each patient's vital trajectories (**Table S2**). Each continuous variable was then discretized with respect to the study population into quintiles[32] and each quintile mapped to a binary feature.[14,15]

Details of model selection

To select the model hyperparameter C (inverse of regularization strength), we performed a grid search from 1e–6 to 1e6 for the C value with 5-fold cross validation (repeated 20 times). The best hyperparameter setting was determined based on maximum area under the receiver operating characteristic curve (AUC) averaged across validation folds.

Alternative evaluation

Given time to event (aGVHD onset), we display the Kaplan Meier plot for high-risk and low-risk patients (from the held-out testing set, N=85) in **Figure S1**. Patients were stratified into two risk groups based on the 30th percentile of the predicted risk scores. We note a good separation between the two groups (high and low risk), where patients in the high risk group exhibited poorer outcomes compared to the low risk group.

Non-linear model

On the task of predicting grade II-IV aGVHD, random forest (a non-linear model) achieved an AUC of 0.651 (95%CI 0.525–0.768); the difference in performance compared to the proposed logistic regression model (AUC=0.659, 95%CI 0.536–0.784) is not statistically significant (p>0.05).

Alternative outcome

When using an alternative outcome definition to predict grade III-IV aGVHD, the fraction of positive cases (over the entire dataset) decreased from 31.8% to 13.6%, leading to a greater class imbalance. On this task, the proposed model achieved an AUC of 0.596 (95%CI 0.423–0.761) on the held-out test set.

Feature importance

Among the models that leveraged a single vital sign (in addition to "baseline features"), the model using temperature achieved a higher AUC of 0.595 (95%CI 0.470–0.720; **Figure S2**). A detailed breakdown of estimated feature importance is provided in **Figure S3**. The model relied most on temperature and SBP, and longitudinal patterns (e.g., increasing/decreasing as characterized by positive/negative slopes) were more important than the average values. We display the row-wise and column-wise average importance in **Figure 4A** and **Figure 4B** of the main text.

Visualization & clustering of important features

In **Figure S4**, we display the trajectories of Temperature and Systolic BP for two patients (from the testing set) with the highest/lowest predicted risk scores. In **Figure S5**, we display the clustering results using the important features pertaining to Temp-slope, Temp-abs(A1), Temp-angle(A1), and SBP-abs(A1).

Supplemental References

30. Richman JS, Moorman JR: Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 278:H2039-49, 2000
31. Fu T-c: A review on time series data mining. Engineering Applications of Artificial Intelligence 24:164-181, 2011
32. Cochran WG: The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24:295-313, 1968

Table S1. The 52 baseline features.

Sex: Female	Disease Code category: Malignant
Sex: Male	Disease Code category: Non-malignant
Age (-0.001, 18.0]	Disease Risk: 0 - Non-malignant
Age (18.0, 45.0]	Disease Risk: 1 - Low
Age (45.0, 75.0]	Disease Risk: 2 - Intermediate
Prophy: ATG	Disease Risk: 3 - High
Prophy: Cyclo	Intensity: 0 - Full
Prophy: Enbrel	Intensity: 1 - Reduced
Prophy: IL-11	Donor source: 0 - Related
Prophy: KGF	Donor source: 1 - Unrelated
Prophy: MMF	Match: No
Prophy: MTX	Match: Yes
Prophy: Siro	Stem cell source: 0 - Bone Marrow
Prophy: Steroid	Stem cell source: 1 - Peripheral Blood
Prophy: Tacro	Stem cell source: 2 - Cord Blood
Race: African American	Season: Spring
Race: Asian	Season: Summer
Race: Bi/Multi Racial	Season: Fall
Race: Black or African American	Season: Winter
Race: Caucasian	Marital Status: Divorced
Race: Declined	Marital Status: Legally Separated
Race: Hispanic	Marital Status: Married
Race: Middle Eastern	Marital Status: Significant Other
Race: Native American	Marital Status: Single
Race: Unknown/Other	Marital Status: Unknown
Race: White	Marital Status: Widowed

Table S2. Details on vital sign features. The following transformations are applied to each of the 6 vital signs (temperature, heart rate, respiratory rate, diastolic blood pressure, systolic blood pressure, and peripheral capillary oxygen saturation). A_1 refers to the first Fourier coefficient of the one-dimensional discrete Fourier transform by the fast Fourier transformation algorithm. Each transformed quantity is then quantized based on quintiles and mapped to five binary features, resulting in a total of 600 vital sign features.

Function	**Daily summary statistics**			
	mean	**std**	**min**	**max**
average value	average value of daily mean	average value of daily std	average value of daily min	average value of daily max
linear slope	linear slope of daily mean	linear slope of daily std	linear slope of daily min	linear slope of daily max
sample entropy	sample entropy of daily mean	sample entropy of daily std	sample entropy of daily min	sample entropy of daily max
abs(A_1)	abs(A_1) of daily mean	abs(A_1) of daily std	abs(A_1) of daily min	abs(A_1) of daily max
angle(A_1)	angle(A_1) of daily mean	angle(A_1) of daily std	angle(A_1) of daily min	angle(A_1) of daily max

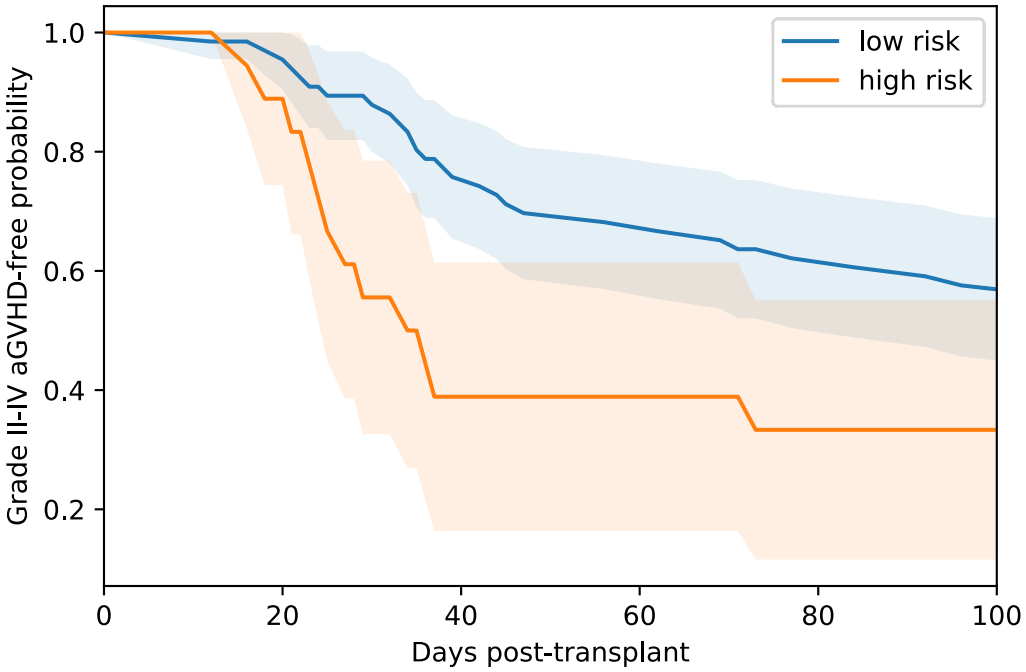

Figure S1. Kaplan Meier plot and 95% confidence intervals for patients in the held-out test set (N=85) within 100 days of transplant. Patients were stratified into two risk groups based on the 30th percentile of the predicted risk scores. Patients in the high risk exhibited poorer outcomes group compared to the low risk group.

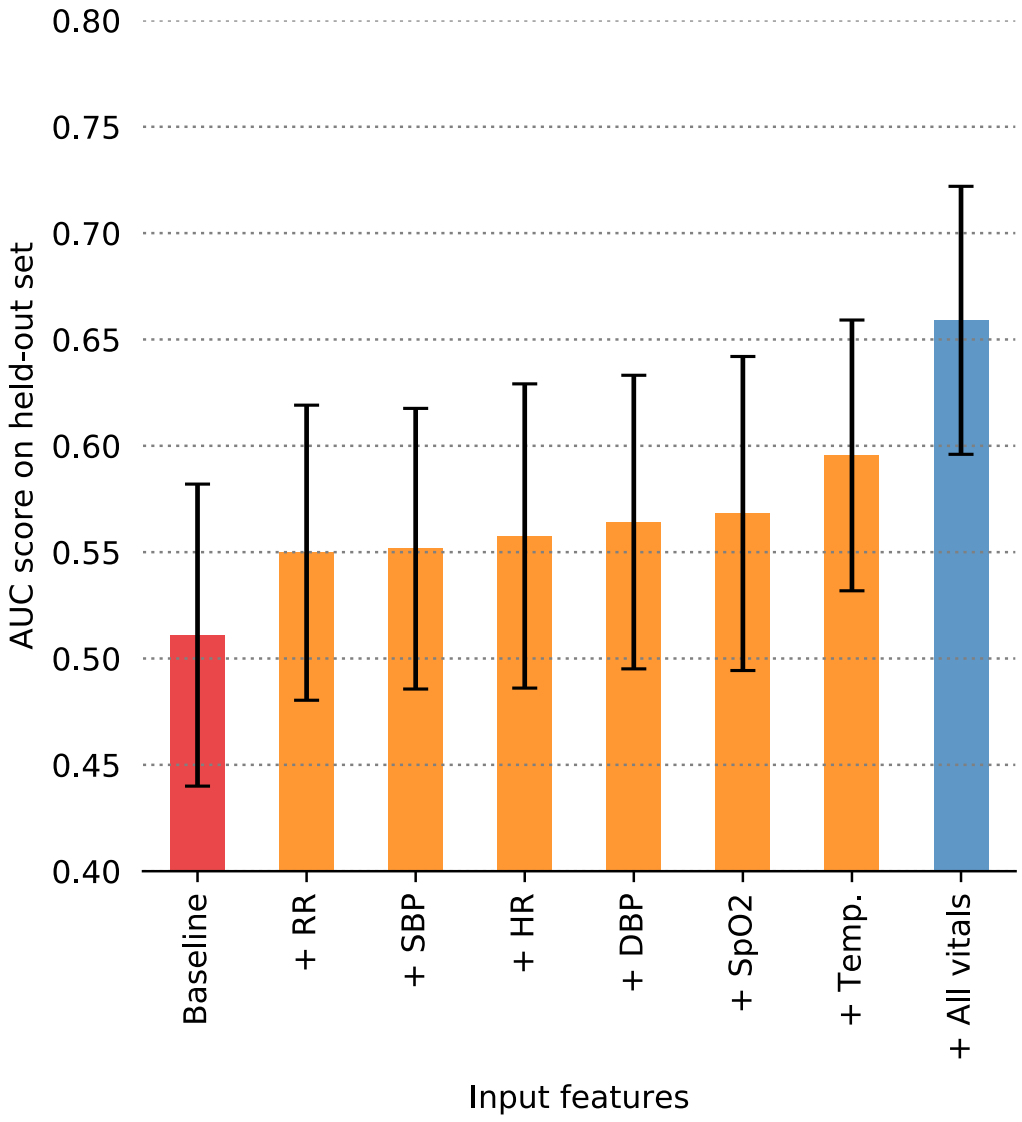

Figure S2. Sensitivity analysis of vital signs. A comparison of the discriminative performance of models trained using features from a single vital sign. Error bars represent plus/minus one standard error calculated on 1,000 bootstrapped samples of the held-out test set (N=85). Among the models that used a single vital sign, the model using temperature led to a higher AUC.

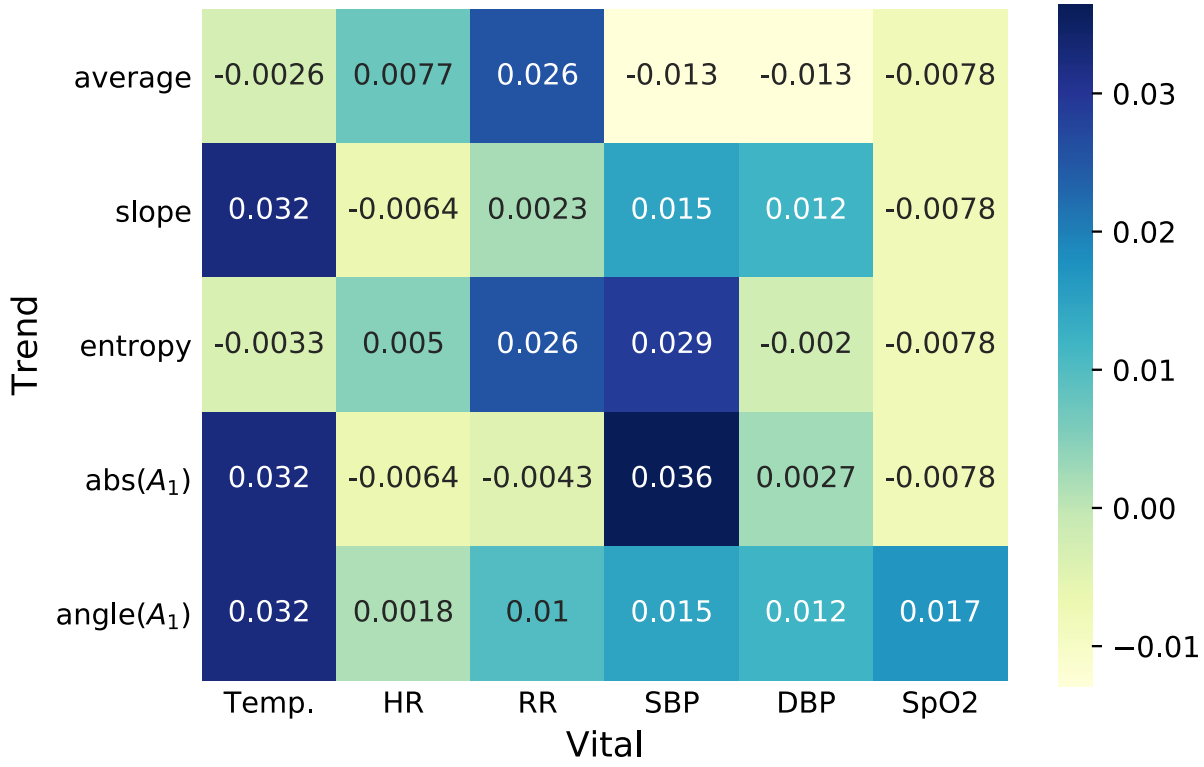

Figure S3. Heatmap of feature importance of different vital signs and different trend features. The importance of a feature group is defined as the decrease in AUC when that group of features are excluded from the model. Darker shades correspond to greater importance. Among vital signs, temperature and systolic blood pressure are important. Among trends, features pertaining to longitudinal patterns (e.g., slope and fast Fourier transform coefficients) are usually more important than the average values.

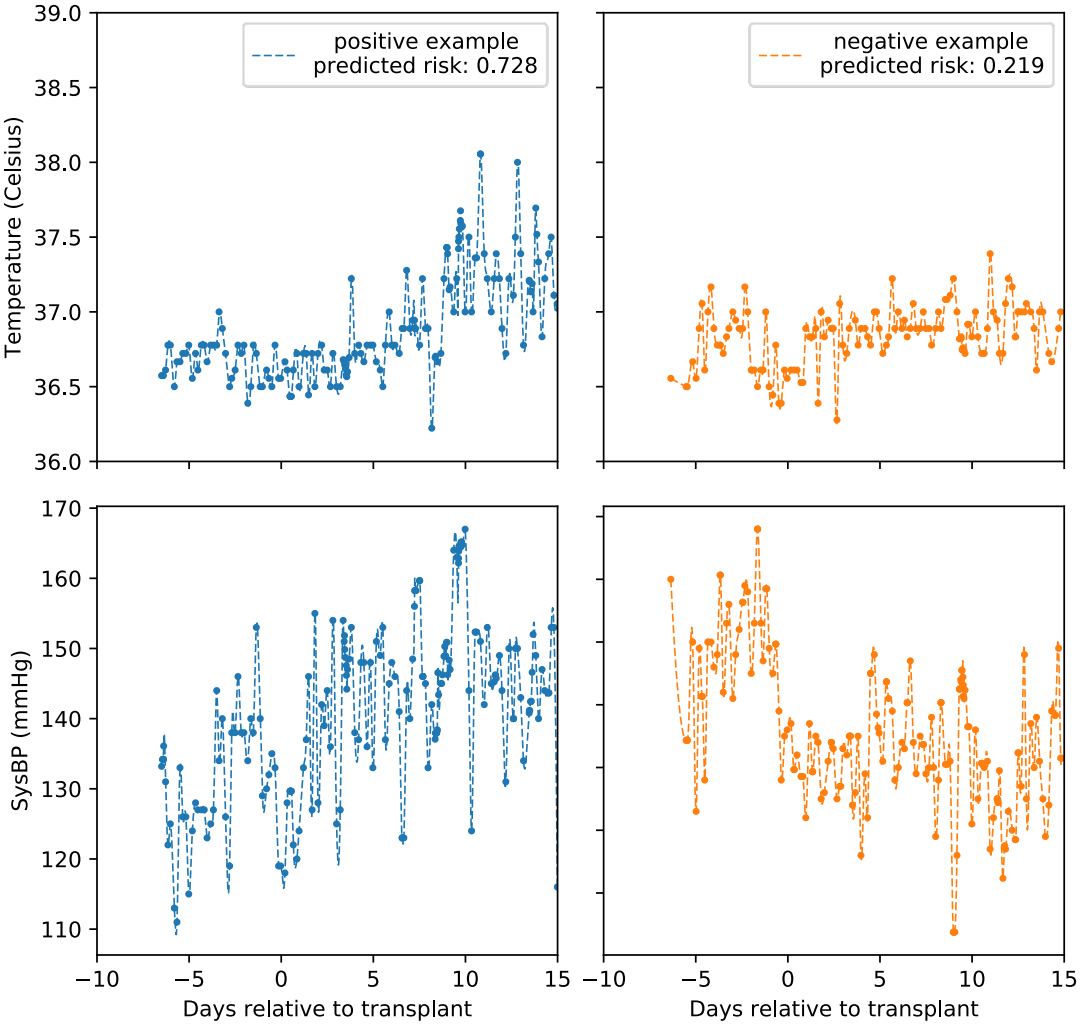

Figure S4. Trajectories of temperature and SBP values for two patients with different outcomes. These two patients are from the held-out test set and had the highest/lowest risk score as predicted by our proposed model. Plotted values are smoothed by moving averages with a window size of 0.1 days.

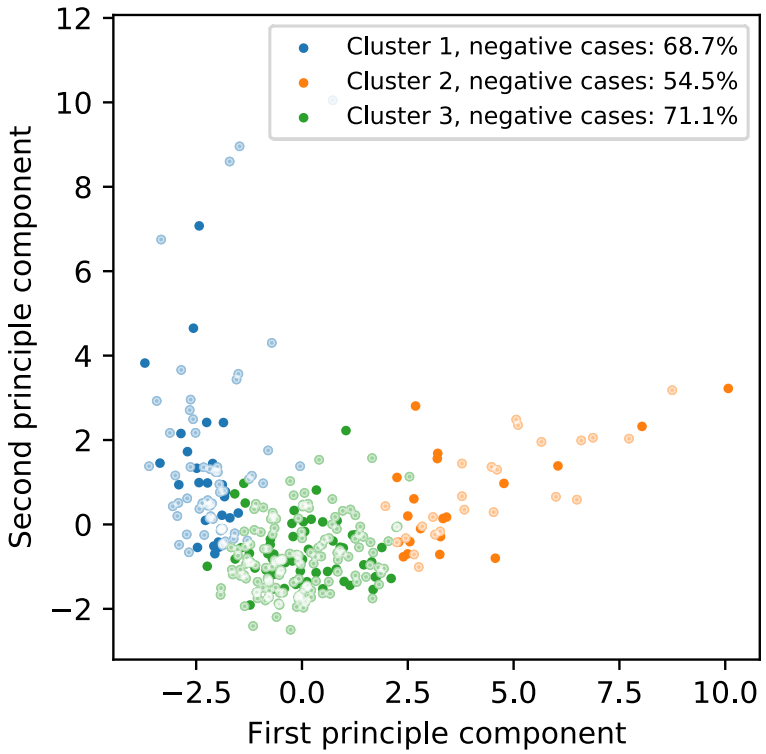

Figure S5. Clustering results of all patients (N=324) using only the important features, plotted in the space of the first two principle components. Input features are 16-dimensional, real-valued, and pertain to Temp-slope, Temp-abs(A1), Temp-angle(A1), and SBP-abs(A1). The clusters have varying class balance (light color - negative examples, dark color - positive examples), suggesting the important features capture part of the differences in patients with different outcomes.