Machine Learning Models to Differentiate the Etiology of Cutaneous Reactions in Post-Stem Cell Transplant Patients

Clara Lemaitre[1], Nimesh V Nagururu[1], Audrey Lacy[1], Vince Wang[1], Nandita Balaji[1], Jonathan Hung[1],
David Weiner MD[2], Austin Burns MD[2], Olivia Pierog MD[2], Casey O. Taylor PhD[1], Joseph L. Greenstein PhD[1], Sima Rozati MD, PhD[2]

1. Johns Hopkins University, Department of Biomedical Engineering, Baltimore, MD, USA
2. Johns Hopkins Medical Institutes (JHMI), Department of Dermatology, Baltimore, MD, USA

Introduction

- Over 25,000 hematopoetic stem cell transplants (HSCT) are performed annually in the United States to treat hematologic conditions and malignancies
- 3 rashes commonly occur as a post-transplant complication: cutaneous Graft-Versus-Host-Disease (GVHD), viral reactivation syndromes, and drug eruptions
- GVHD occurs in 40% of post-HSCT patients and has a 35% mortality rate.
- Differentiating between cutaneous eruption etiologies is difficult given overlapping clinical presentations and the complex medical course of HSCT patients (i.e. immunosuppression and complex drug regimens)
- Accurate identification of rash cause is essential to initiate appropriate and timely treatment.

Objective

Assist dermatologists by employing machine learning models to synthesize diverse, high-dimensional data to aid in differentiating the cause of cutaneous eruptions in post-HSCT patients

Data Overview

Dataset: a retrospective cohort comprised of patients with rash within 1 year of non-allogeneic Bone Marrow Transplant at JHMI between 2015-2021

Available Features: rash characteristics, rash etiology labels, demographics, transplant characteristics, medications received, Shiohara DIHS criteria, Regiscar score, and all laboratory data within one year of transplant.

Feature Subsets

- All Features
- Abbreviated Clinical: Previously identified characteristics associated with GVHD
- Extended Clinical: all available clinical data and viremia labs
- All Labs: laboratory data within one year of transplant

Demographics and Rash Characteristics by Rash Etiology

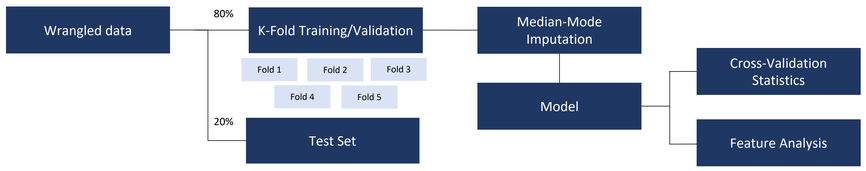
Variable	Overall N = 625[1]	Rash Etiology GVHD N = 471[1]	Non-GVHD N = 154[1]	p-value[2]
Age at transplant (years)	58 (44, 66)	59 (45, 66)	57 (37, 65)	0.13
Female Sex	253 (40%)	194 (41%)	59 (38%)	0.53
Non-White Ethnicity*	169 (27%)	110 (23%)	59 (38%)	0.004
Days from transplant to rash	55 (36, 96)	57 (38, 94)	49 (30, 111)	0.12
Rash duration (days)	42 (17, 99)	46 (21, 108)	20 (11, 61)	<0.001
Rash Location				
Head or neck	401 (64%)	320 (68%)	81 (53%)	<0.001
Extremities	386 (62%)	291 (62%)	95 (62%)	0.98
Acral	17 (2.7%)	13 (2.8%)	4 (2.6%)	>0.99
Trunk	476 (76%)	372 (79%)	104 (68%)	0.004
Rash BSA percentage				0.037
<50%	386 (62%)	277 (59%)	109 (71%)	
50%	32 (5.1%)	27 (5.7%)	5 (3.2%)	
>50%	144 (23%)	113 (24%)	31 (20%)	
100%	63 (10%)	54 (11%)	9 (5.8%)	
Skin biopsy taken	292 (47%)	229 (49%)	63 (41%)	0.10
Associated Pruritis	349 (56%)	287 (61%)	62 (40%)	<0.001
Days from transplant to diarrhea (days)	4 (-1, 11)	4 (-1, 10)	4 (-1, 16)	0.75
Diarrhea duration (days)	9 (-1, 28)	9 (-1, 29)	9 (-1, 26)	0.24
RegiSCAR Excluded category (<2 Points)*	588 (94%)	443 (94%)	145 (94%)	0.96

[1]Median (IQR); n (%)
[2]Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test
* Variable collapsed for presentation, but p-value computed amongst all categories

Table 1. Details patient demographics and rash characteristics of our cohort

Supervised Approach: Methods

Supervised Approach: Binary classification of GVHD and non-GVHD rashes



Supervised Approach: Results

Comparison of Machine Learning Model Performance in Cross-Validation

Model	F1 Score All	F1 Score Abbr. Clinical	F1 Score Ext. Clinical	F1 Score Labs	ROC AUC All	ROC AUC Abbr. Clinical	ROC AUC Ext. Clinical	ROC AUC Labs
Logistic Regression	0.8	0.759	0.797	0.774	0.678	0.662	0.663	0.62
Random Forest	0.871	0.86	0.862	0.857	0.717	0.683	0.757	0.655
XGBoost	**0.874**	0.859	0.855	0.834	0.712	0.706	0.723	0.633
ElasticNet	0.801	0.765	0.805	0.774	0.678	0.659	0.663	0.61
Support Vector Machine	0.86	0.803	0.813	0.848	0.686	0.624	0.669	0.61
AdaBoost	0.806	0.81	0.856	0.758	0.646	0.669	0.709	0.563
Multilayer Perceptron	0.829	0.826	0.84	0.821	0.674	0.651	0.675	0.607

Table 2. Details performance (F1 score & ROC AUC) of 7 different machine learning models on cross-validation set for varying subsets of features (abbr.=abbreviated & ext.=extended). The XGBoost model using all of the features performed the best as indicated by having the highest F1 score of 0.874 and was selected as the final model.

Shapley Additive Explanation Plot for Final Model

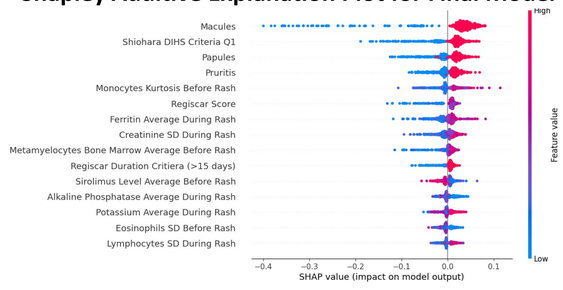


Figure 1. Shapley analysis for the final XGBoost model based on the training & validation data. Shows the impact of the top 15 most important features on GVHD classification.

Performance of Final Model on Held-Out Test Set

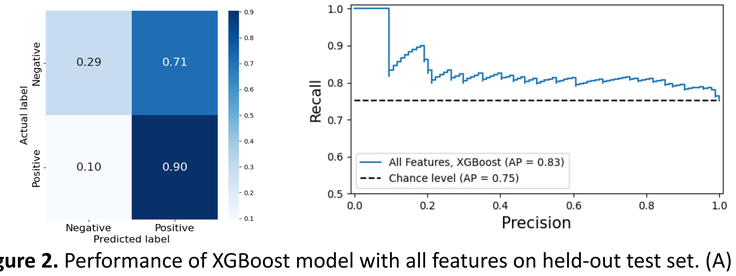
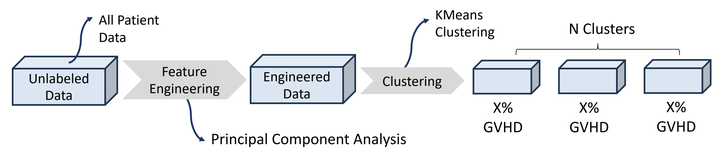


Figure 2. Performance of XGBoost model with all features on held-out test set. (A) Confusion matrix comparing predicted label vs. true label normalized by true label (B) Precision-Recall Curve, where AP indicates Average Precision.

Unsupervised Approach: Methods

Unsupervised Approach: Unsupervised hierarchical clustering models were utilized to identify naturally occurring groups within overall cohort of patients and within the GVHD rash cohort.



Unsupervised Approach: Results

Exploration of Naturally Occurring Patient Cohorts

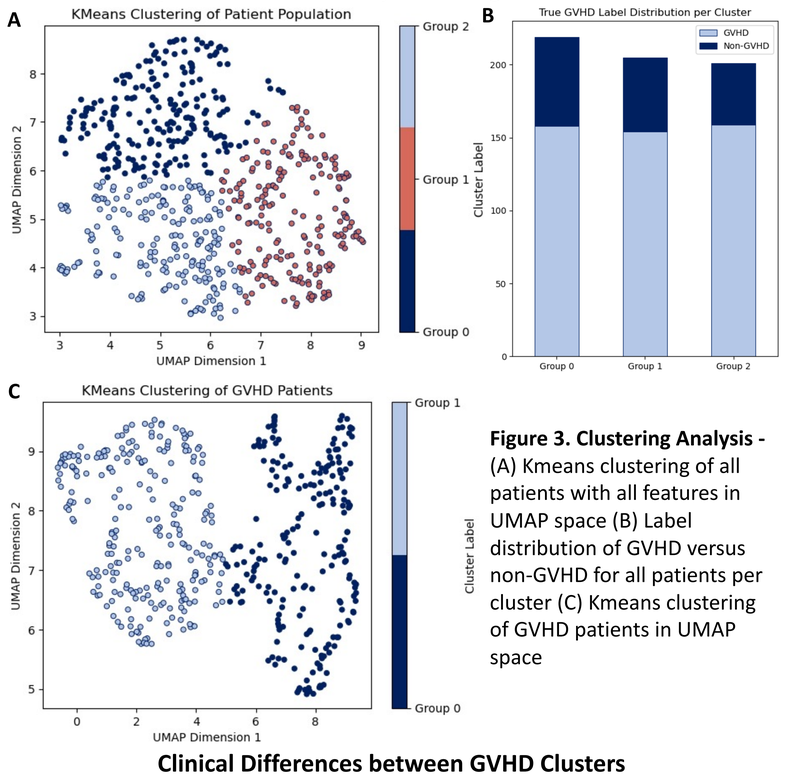


Figure 3. Clustering Analysis - (A) Kmeans clustering of all patients with all features in UMAP space (B) Label distribution of GVHD versus non-GVHD for all patients per cluster (C) Kmeans clustering of GVHD patients in UMAP space

Clinical Differences between GVHD Clusters

Table 3. Details statistically significant extended clinical features between patients when stratifying based on cluster labels from Figure 3C. Additional significant lab features were identified between the clusters.

Variable	Cluster 0 N = 253[1]	Cluster 1 N = 218[1]	p-value[2]
Rash Duration	63 (33, 127)	33 (15, 80.5)	0.007
Rash Site: Head/Neck	156 (73%)	134 (61%)	0.005
Rash Type: Papules	145 (67%)	125 (57%)	0.035
Days: transplant to rash	55 (42, 64)	61 (50, 68)	<0.001
Viremia: HHV6	50 (13%)	43 (20%)	0.039
Viremia: EBV Time	41 (9.8%)	35 (16%)	0.049
Viremia: EBV Present	42 (10%)	36 (17%)	0.049
Regiscar Score	0 (-1, 0)	-1 (-1, 0)	<0.001

[1]Median (IQR); n (%)
[2]Welch's unequal variances t-test

Conclusion & Future Direction

- Supervised learning models offer some integration of high-dimensional data ranging from lab to clinical data as identified by the average precision improvement over random chance, however, the improvement is marginal.
- Unsupervised approaches find no naturally occurring GVHD clusters; however, they identify distinct populations within GVHD patients.
- Future work is focused on improving feature-engineering and employing semi-supervised learning techniques to address label uncertainty