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

Clara Lemaitre1, Nimesh V Nagururu1, Audrey Lacy1, Vince Wang2, Nandita Balaji1, Jonathan Hung1, David Weiner MD2, Austin Burns MD2, Olivia Pierog MD2, Casey O. Taylor PhD1, Joseph L. Greenstein PhD1, Sima Rozati MD, PhD2
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 hematopoietic 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, Shohara 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 viiremia labs
• All Labs: laboratory data within one year of transplant

Demographics and Rash Characteristics by Rash Etiology

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

Supervised Approach: Methods

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

Comparison of Machine Learning Model Performance in Cross-Validation

Table 2. Details performance (F1 score and 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.

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 2. Performance of XGBoost model with all features on hold-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.

Supervised Approach: Results

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.

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.