Lymphoma Development Prediction

Data Science Institute
COLUMBIA UNIVERSITY

**Eric Boxer, Po-Chieh Liu, Costas Vafeades,
Ioannis Tzemos and Peter Grantcharov
Dr. Katherine Bettencourt and Julie Yu
Dr. Adam S. Kelleher**

Data Science Capstone Project with Johnson & Johnson

Introduction

Diffuse large B-cell lymphoma (DLBCL) is the most prevalent form of non-Hodgkin's lymphoma among adults, yet its causes and disease progression factors are poorly understood. Prior work at Johnson & Johnson has indicated that patients with mood disorders are less prone to develop DLBCL.

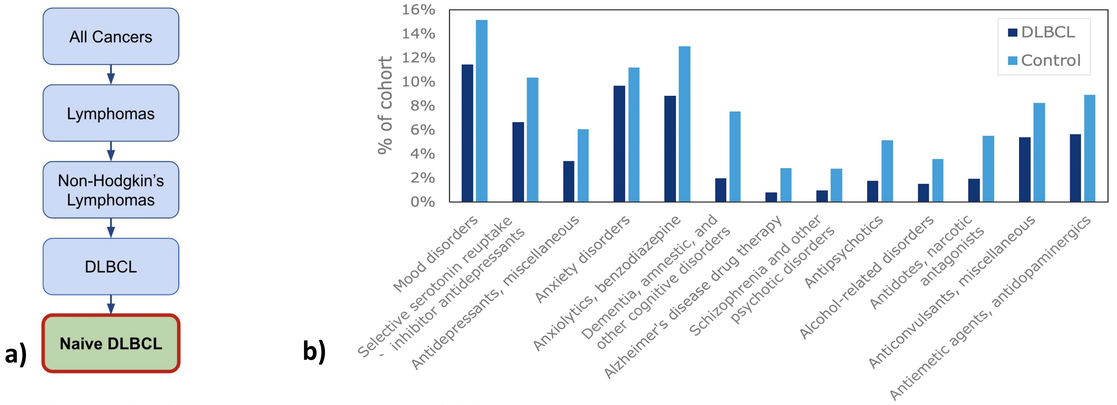

Figure 1. a) Tumor subtype tree b) Mental illness event code prevalence in cohorts of interest

NLP Approach - Contextual Representation Learning

We matched patients on demographic features and their length of time in the dataset without an event that could leak information about a DLBCL diagnosis. Word2Vec embeddings are trained for events and used as input into the Patien2Vec framework to learn interpretable, longitudinal patient representations and predict DLBCL outcomes.

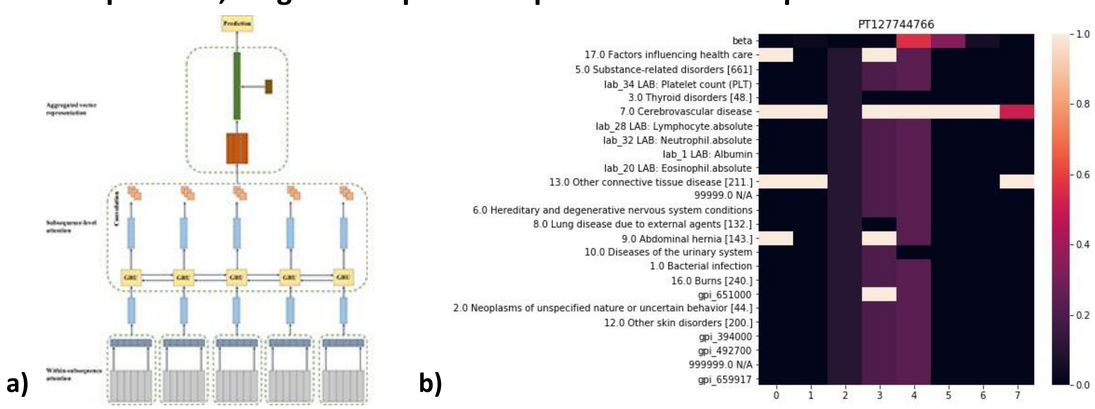

Figure 2. a) Patient2Vec architecture b) Attention weights from a correctly classified patient

Event Code Classification

We incorporated all event code types into a single model. Control and DLBCL cohorts were one-to-one matched on demographic variables and diagnosis time period. Univariate analysis was applied to reduce feature space dimension from 15,519 to 87. 79 features were found in significantly greater proportions for the DLBCL cohort (patients with mood disorders who go on to develop DLBCL in the dataset). The features with the greatest disparities in cohort proportions are shown in Figure 3.

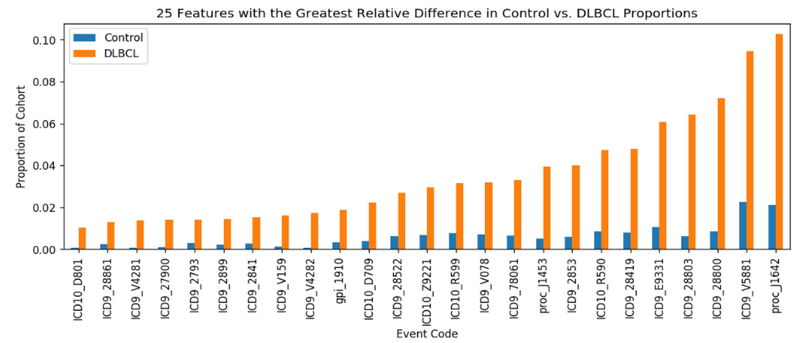

Figure 3. Event code proportions in cohorts of 25 features with greatest proportion disparities

Medicine Focused - GPI Codes Approach

We identified three criteria for our matching of the DLBCL and Non-DLBCL cohorts: patient-related information (date of first mood diagnosis, birth year, first year active in the healthcare system), biological characteristics of each patient (gender, race, ethnicity) and geographic information (region, division, average household income). The first two remained intact during our matching process and the last category was recursively loosened.

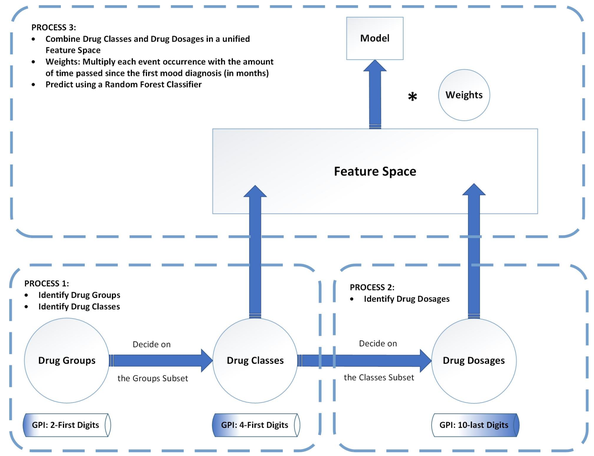

Figure 4. Three-step process to create the reduced feature space for the model

Ensemble Classifier

We used a voting classifier on top of our models' outputs to predict DLBCL outcomes for a hold-out test set. The model is able to adequately differentiate between non-DLBCL patients. Future work can focus on improving sensitivity of the predictions, to capture DLBCL patients.

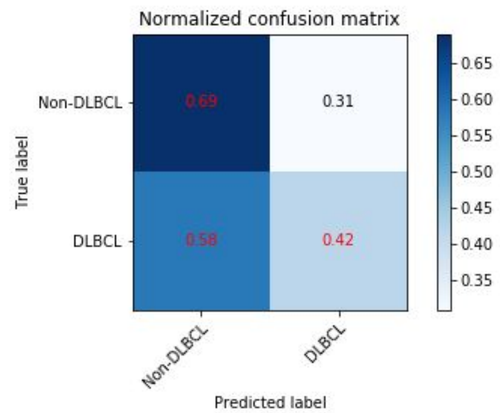

Figure 5. Classification results from an ensemble of our three methods

Acknowledgments

This work was supported by the Janssen. The views expressed in this poster are those of the authors and do not necessarily represent the views or policies of the Janssen.

References

Jinghe Zhang et al., Patient2Vec: A Personalized Interpretable Representation of the Longitudinal Electronic Health Records. arxiv: 1810.04793 [q-bio.QM] 25 Oct 2018.