Predicting immunotherapy response of cancer patients using clinical and genomic attributes

Zhiyue Zhang, Edward Zheng, David White, Matthew Bare

COMP562 Fall 2022

1 Introduction

Immunotherapy is a novel cancer treatment that stimulates the immune system to fight cancer. One major category of immunotherapy are drugs called Immune Checkpoint Inhibitors (ICI). Currently, the only FDA-approved biomarker for ICI outcome is Tumor Mutation Burden (TMB), which measures the number of mutations a tumor contains in a given length of genome, or the foreignness of a tumor compared to normal tissues. It is widely accepted that the more foreign a tumor is (i.e. higher TMB), the easier it is for the immune system to recognize it. Because of this, patients with higher TMB should respond better to ICI.

However, literature has suggested that TMB is not the only factor contributing to the treatment outcome of ICI. Here, we analyzed the data from 1661 patients who received ICIs at the Memorial Sloan Kettering Cancer Center (MSKCC), accessed through cBioPortal. We present machine learning models that predict tumor-specific ICI survival outcome using not only TMB, but also the clinical attributes (age and sex) and genomic information (clinically-enriched mutations) of patients.

2 Data

Our project was based on a publicly-available dataset of patient mutatons that we processed in R using the code found in the drive-download-20221210T233345Z-001/ directory. This dataset included basic information about each patient's status: their biological sex (, the type of cancer they were suffering from, their age, and whether they were still alive. These data come from a study published by Samstein et al. (2019). Our hope was that we could build a model capable of predicting a patient's survival based on the particular mutation rates displayed in their genome. To this end, we separated the data into groups according to the type of cancer suffered by the patient for some of our tests, as we hoped more specific models would be able to achieve higher accuracy rates.

3 Methods

3.1 Logistic Regression with TMB as Predictor

Our first approach involved using TMB alone to predict OS_STATUS (patient survival status). We extracted TMB_NONSYNONYMOUS and OS_STATUS from the cancer_data.csv data set. We then applied an 80:20 train:test split on the data. We implemented logistic regression using the LogisticRegression() object from sklearn along with the .fit() function.

3.2 Multiple Logistic Regression with Cancer Type Subsetting

Our second approach used sex, age, and clinically enriched genes together to predict survival status. We created a function, multivar_regression(df, var_arr), that takes input data and a list of the predictors to train on. We again used sklearn's LogisticRegression() to train the model. We performed a separate regression for each cancer type.

4 Results

Our inital general logistic regression model was able to achieve an accuracy of 0.60241 when trained on the dataset. This model was a single-varable logistic regression based on the TMB value observed for each patient, and was not designed to work on any specific cancer type (and, in fact, did not take cancer type directly into account at all). While not a particularly impressive result itself, this indicated that there was promise in the concept that machine learning models could predict a patient's survival based on the cancer-related mutations that were obsereved in their genome. We also worked on a multivariable logistic regression model, incorporating all the values of the data set except the patient's survival, but were unable to directly achieve better results than the first approach. In fact, the result was actually slightly worse, with an accuracy of only 0.54217, so we decided instead to investigate in another direction.

We found more success by dividing the dataset between different types of cancer. We looked at four types: non-small cell lung cancer, melanoma, bladder cancer, and renal cell carcinoma. Running the single-variable model on each of these categories intially gave similar results to our previous attempts, with accuracies ranging from o.51163 to 0.65093. However, we were able to improve some of the categories significantly by utilizing a multivariable logistic model that was sspecalzed for each type of cancer by incorporating only the specific mutations known to be associated with that cancer type. Melanomas showed a marked increase, going from an accuracy of 0.60937 to 0.67188. Lung cancer stayed roughly the same as the single-variable model, as did renal cell carcinoma. Bladder cancer, however, performed notably better with the new model, going from an accuracy of only 0.51163 to an accuracy of 0.79070. This implies that

TMB is a poor indicator of bladder cancer survivability specifically, but bladder cancer is significantly better modelled based on specific mutations associated with that disease.

5 Conclusion

Given the developing nature of cancer treatment, it's difficult to ascertain in each case whether the accuracy of the model is a true reflection of the model's predictive capability for future patients or whether it's a better reflection of the effectiveness of the cancer treatment. There is the possibility that higher accuracy in the model may simply reflect that a type of cancer is particularly unresponsive to treatment (many deaths and easy to predict) or particularly responsive (many survivals and easy to predict). Further work would seek a way to mitigate this uncertainty, perhaps by generalizing the prediction across cancer types.

6 Bibliography

Samstein et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet. 2019 Feb;51(2):202-206. doi: 10.1038/s41588-018-0312-8. Epub 2019 Jan 14. PMID: 30643254; PMCID: PMC6365097.