Predicting a Decline in Patient Reported Outcomes for Cancer Patients on Chemotherapy

CS 229, Spring 2020, Final Project Write Up
Nicolai Ostberg (nostberg) & Dylan Peterson (dpeters9)

1 Introduction

Cancer is a major cause of global morbidity and mortality, and was the second leading cause of death worldwide in 2018.[1] Unlike other common diseases, such as heart disease, treatment options for cancer are frequently associated with significant adverse effects. This is especially true with respect to chemotherapy, which has drug-dependent side-effects that can range from severe fatigue and hair loss to cognitive slowing and skin discoloration. The effects of chemotherapy are not limited to physical symptoms and can persist for years after treatment: one landmark study even showed that breast cancer survivors who received chemotherapy had worse reported quality of life years after treatment.[2]

Given the impact chemotherapy can have on patients, understanding the factors associated with outcomes and being able to predict worse outcomes during and following treatment is important, as it could allow for additional, targeted interventions that seek to mitigate these adverse outcomes. While there has been much work to date on predicting outcomes related to cancer and treatment with chemotherapy (see Section 2: Related Work), few studies have focused specifically focused on predicting the impact of chemotherapy on Patient Reported Outcomes (PROs), such as a patient's subjective assessment of their physical or mental health. For instance, querying PUBMED for *"machine learning"* AND *"cancer"* AND *"patient reported outcomes"* returns 5 articles.

Despite their subjective nature, routine collection of PROs in oncologic care has been shown to improve patient satisfaction and even survival.[3] One well validated tool for capturing PROs are PROMIS surveys, which can be used to calculate global mental health (GMH) and global physical health (GPH) scores. These scores are designed to have a population mean of 50 points (SD of 10 points) and can be used by clinicians to get a more objective snapshot of a patient's subjective health status (lower scores relate to worse health). We hypothesized that being able to predict if a patient is at risk of poor or declining PROMIS scores while on chemotherapy, prior to chemotherapy initiation, would be beneficial to both clinicians and patients, allowing for personalized treatment plans and more focused interventions for patients at increased risk. To this aim, we decided to train several common machine learning algorithms to predict poor or declining GPH scores while on chemotherapy using electronic health record (EHR) data that would be available to clinicians prior to a patient's first day of chemotherapy.

2 Related Work

As mentioned, much prior work has been done on predicting aspects of cancer care. Searching for PUBMED for *"machine learning"* AND *"cancer"* returns 4,252 articles and the same search on Google Scholar returns over 1.25M results. Broadly, the studies have sought to predict (1) susceptibility to developing cancer; (2) cancer detection; (3) response to treatment, cancer survival and prognostication; and (4) susceptibility to cancer recurrence. These generally have used features derived from some combination of clinical, genomic, and radiologic data to train a wide variety of models (from logistic regression to deep neural nets) to predict on cohorts ranging in size from tens of patients to more than tens of thousands of patients. Kourou et al. present a strong review of the applications of machine learning for cancer care.[4]

Looking specifically at predicting PROs for patients undergoing cancer treatment, however, reveals far fewer studies. In fact, we had difficulty finding publications in on this topic. This is perhaps because PROs are not as routinely captured as other data types (such as clinical labs). Ranashinghe et al. circumvented this through by scraping free text from online canceer support groups and applying NLP to identify features associated with patient-reported outcomes; however, this technique did not use previously validated PRO metrics, such as PROMIS scores, so its clinical utility is questionable.[5] Studies that use ML techniques on validated PRO measures for cancer patients are even more infrequent. Pan et al report on using ensemble methods to predict PROs in response to radiotherapy with promising results (AUROCs of 0.79-0.87), however, this was not aligned with our treatment of interest (chemotherapy).[6] We identified one study, by Wang et al., that sought to cluster cancer patients on chemotherapy by PROMIS scores however, this study was severely limited, including only 96 pediatric patients and 4 features.[7] Additional unpublished work by the Boussard Lab at Stanford showed 743 patients' PROMIS scores clustered into unique trajectories while on chemotherapy. However, this work failed to find any meaningful predictors of the changes in PROMIS scores in response to chemotherapy using demographic and clinical data ($\sim$ 20 features) with linear models. Therefore, we thought using more advanced machine learning models on a high dimensional feature set could potentially yield improved predictive capabilities.

3 Dataset and Features

Data was sourced from the Stanford Research Repository (STARR) data warehouse, which contains data from nearly three million de-identified patient records. Relevant clinical data captured includes demographic information, diagnoses, vital signs, lab values, orders, and procedure histories. Data is imported from the physician-facing EHR (EPIC) and transformed into tables that can be queried by researchers.

This project focused on cancer patients present in the STARR database. Inclusion criteria for this study included (1) cancer diagnosis after January 1, 2013 (date of integration of PROMIS surveys into clinic workflow), (2) received chemotherapy, and (3) completion of at least one pre-treatment and one on-treatment PROMIS survey. These inclusion criteria resulted in a cohort of 1252 patients.

Next, we queried the STARR database to derive feature vectors for each patient. Specifically, we included vital signs, diagnoses, medications, procedures, lab values, healthcare utilization and demographic data. In order to prevent our models from learning from "future" events, we restricted our data to only include information captured on or prior to the patient's first date of chemotherapy administration. To ensure our data best reflected the current health of each patient, except for diagnoses (which tend to be longstanding), we removed any data collected more than 6 months prior to the start of chemotherapy. Furthermore, to deal with sparsity in our data, we excluded diagnoses, lab tests, and procedures that were not present for or performed on at least 2.5% of our cohort. In addition, we removed physiologically impossible vital signs based on clinician input.

Vitals and labs vary temporally, yet the model architectures employed were invariant to time. In addition, sample frequency was low so sequence based models couldn't be trained. However, we constructed a featurization scheme that allowed us to capture some of this temporal variation in the feature vectors for each patient. Specifically, an exponentially weighted average was calculated for each time varying feature f according to the following formula:

$$f = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \text{ where } w_i = e^{-t/365}$$

where $t =$ the number of days from the start of chemotherapy. We used this weighting scheme because it weighted recent values more heavily while still accounting for historical trends.

Additional processing was performed on lab values. In order to deal with missing values, we employed two featuriztation techniques. In one, we binned patients' time-weighted average values into tertiles and explicitly encoded missingness, resulting in four features per lab test (binned labs). This scheme is particularly important in healthcare settings, as the absence of a test can be informative, perhaps suggesting the clinician did not feel it was necessary to obtain. We also constructed an additional featurization scheme that imputed missing values using a patient's K-Nearest Neighbors (KNN labs).

Procedures were also encoded using the exponential weighed averaging scheme in order to more heavily weight recent procedures that might have a larger impact on quality of life and mental health. Diagnoses recorded were featurized into a sparse binary matrix, with a 1 indicating that a patient had been diagnosed with the condition. Similarly, binary features were made for classes of medications (e.g. the class "analgesic" would include aspirin, acetaminophen, etc). A patient received a 1 for a given medication class feature if they had been prescribed a medication within that class in the 6 months prior to their first date of chemotherapy. We included additional features for the total number of diagnoses, procedures, and medication classes that each patient had in order to further quantify the patient's overall health.

Health care utilization data was featurized into the number of ED visits, proportion of ED visits resulting in hospitalizations, the total number of hospitalizations, the average length of hospital admissions, and the number of psychiatrist appointments (given our interest in subjective measures) in the 6 months before chemotherapy. Additional demographic information, such as race, gender, and age were also featurized. Categorical variables with multiple options (e.g. insurance status) were expanded to multiple binary features per variable. Additionally, patients' pre-treatment (baseline) PROMIS scores were included as additional features.

Finally, we had two binary output labels to identify patients who do poorly on chemotherapy. The first was positive if a patient's GPH score dropped by one standard deviation (-10 points) while on chemotherapy (drop outcome). The second was positive if a patent's on chemotherapy score was less than or equal to one standard deviation below the population mean (i.e. ≤ 40 points, threshold outcome). These were explicitly calculated as:

$$y^{(i)}_{\text{drop}} = 1\{\text{score}^{(i)}_{\text{on chemo}} - \text{score}^{(i)}_{\text{at baseline}} \leq 10\} \quad \text{and} \quad y^{(i)}_{\text{threshold}} = 1\{\text{score}^{(i)}_{\text{on chemo}} \leq 40\}$$

Ultimately, we had 7 vital sign features, 582 binned lab features, 148 KNN lab features 291 procedural features, 266 diagnostic features, 121 medication features, 5 utilization features, 88 demographic features, and 6 baseline score features. These were combined into two "full" feature matrices: one with KNN labs and one with binned labs, with 933 and 1367 features, respectively.

4 Methods

We explored a variety of models with varying architectures to best capture the resulting feature matrices and outcomes. Specifically, we tested a logistic regression model with LASSO, Ridge, or Elastic Net penalties, ensemble models such as random forest and gradient boosted trees, neural networks in the form of a multi-layer perceptron, a support vector machine (SVM), K Nearest Neighbors classifier, and an ensemble voting model.

For each model, we tuned hyperparameters in order to optimize model performance. Specifically, we performed 10 fold cross validation on the training set and evaluated performance using AUROC. We employed a grid search to discover the best hyperparmeters for each model, which increased the computational complexity of training but ensured that the optimal set of hyperparameters was found. We then selected the best performing model and evaluated on the test set. The details of what hyperparameters were tuned for each model is described below. All code was written in Python 3.6 and all models were implemented with scikit-learn 0.22.1 with the exception of the gradient boosted, which was implemented with LightGBM 2.3.1.[8]

Logistic Regression:

Because the number of features was similar to the number of samples, regularization was needed in order to achieve high performance. For both the LASSO model and the Ridge model, we tuned the regularization parameter λ between 10^{-2} and 10^{8} in increments of powers of 10, resulting in 11 potential models. we noted that the model failed to converge with regularization values less than one due to the regularization increasing the weights of the model.

For the Elastic net model, we tuned both the regularization parameter λ between 10^{-2} and 10^{8} and the fraction of L1 (LASSO) regularization compared to L2 (Ridge) between 0.1 and 0.9 in increments of 0.1. Between tuning both parameters, we trained a total of 99 elastic net models.

Tree Based ensemble models:

We tested two tree models - random forest and gradient boosted trees. In our random forest model, we kept the number of estimators constant at 1000 and tuned the following hyperparameters: the minimum number of samples per tree split (2,5,10, and 100) and the number of features used (either all features or the square root of the number of features).

The learning rate and the number of leaves in the tree for the gradient boosted tree were tuned. The test set was used to evaluate each iteration and early stopping was employed with a tolerance of 2, in order to prevent overfitting.

Neural Network:

The neural network required the most exhaustive hyperparameter searching. For each model, the regularization parameters for the weights of the network was tuned (between 10^{-4} and 10^{4}). The number of nodes in the first layer was either 100, 325, 550, 775, or 1000, the number of nodes in the second layer was one of 50, 162, 275, 387, or 500, and the third hidden layer was fixed at 10 nodes. Due to the combinatorics of all of these hyperparameters, there were 225 possible models. ReLu activation was used for all nodes except for a final sigmoid node to produce a classification. The model used an Adam optimizer and used the test set to evaluate the model after each epoch to perform early stopping if there was no improvement in 5 epochs.

Support Vector Machine:

Two hyperparamters were tuned for the Support Vector Machine (SVM) model: the regularization parameter which was tuned between 10^{-4} and 10^{4} in and the type of kernel used: linear, polynomial, radial basis function, or sigmoid. A total of 36 models were trained for each trial.

Voting-based Models:

Finally, we employed two voting based models. One was simple: a K-nearest neighbor classifier. We tuned the number of neighbors k used to make a classification between 1 and 100 on a log scale. In addition, the weight the each neighbor had on the resulting classification was either scaled based off of the L2 distance or uniform across all neighbors.

The final voting based model was an ensemble model that averaged the predicted probabilities of a positive label for the three best logistic models, the two best tree based models, and the best SVM (6 constituents total), to output a final label probability. There was no additional hyperparameter tuning employed for this model.

5 Experiments, Results, and Discussion

Experiments:

We had two primary goals for this project: first, to optimize predictive performance as measured by AUROC; and second, to determine the relative impact of each feature set (e.g. vitals) on the predictive performance on our best model. To accomplish this, we devised two experiments. First we trained each of our 9 models to predict

each of the two outcomes using the two final feature matrices (the 933 feature matrix with KNN labs and the 1367 feature matrix with the binned labs) according to the details outlined in Section 4. This entailed 36 (9 x 2 x 2) trainings and evaluations. Next, we selected the best performing model-feature matrix-outcome combination and performed an ablation analysis, training and testing the model using reduced feature matrices that did not contain various feature types, (e.g. removing demographic features), and analyzed how performance changed as we "hid" these data in order to better understand the relative importance of each data type for predictions.

Results:

Experiment 1: Best Model

Our models, when trained on the full feature sets, predicted with AUROCs ranging from 0.5413 to 0.7713 (see Table 2). For both outcomes, models trained on the feature matrix that used the KNN labs generally outperformed those that were trained on the feature matrix that used the binned labs (hereafter, all models should be assumed to be trained using KNN labs, unless otherwise specified). The models for the drop outcome performed generally better than those for the threshold outcome (mean ± SD AUROC scores of 0.7166 ± 0.0582 and 0.7039 ± 0.0550, respectively). For both outcomes, the Ensemble Voting model performed the best, with an AUROC of 0.7713 for the drop outcome and 0.7521 for the threshold outcome (see Figure 1); and an overall AUROC across all four tasks of 0.7570 ± 0.0116. The KNN models performed the worst for both outcomes, and the FFNN models performed poorly as well, while the SVM and LightGBM models had mixed performance. Finally, the logistic models generally performed fairly well on all tests in the experiment.

Table 1: AUROC Scores for all models using all features

	KNN Labs		Binned Labs		Overall	
Model	10 point drop in GPH	GPH below 40	10 point drop in GPH	GPH below 40	Mean	SD
LASSO	0.7636	0.7309	0.7636	0.7309	0.7472	0.0189
Ridge	0.6942	0.7345	0.6932	0.7105	0.7081	0.0193
Elastic Net	0.7672	0.7352	0.7696	0.7352	0.7518	0.0192
KNN	0.5981	0.5730	0.5666	0.5413	0.5697	0.0233
LightGBM	0.7602	0.6967	0.7342	0.7203	0.7278	0.0265
SVM	0.6773	0.7271	0.6947	0.6900	0.6973	0.0212
Random Forest	0.7340	0.7184	0.7071	0.7203	0.7199	0.0110
FFNN	0.6837	0.6676	0.6854	0.6527	0.6723	0.0154
Ensemble Voting	**0.7713***	**0.7521**	**0.7603**	**0.7442**	**0.7570**	0.0116
Outcome Mean	**0.7166**	0.7039	0.7083	0.6939	0.7057	0.0579
Outcome SD	0.0582	0.0550	0.0625	0.0635		

** denotes the optimal model-feature set-outcome combination used in the ablation analysis*

Experiment 2: Ablation Analysis

Table 2: Ablation results

Feature set ablated	AUROC	95% CI
Diagnoses	0.790	0.727-0.854
Demographics	0.781	0.718-0.846
Utilization	0.776	0.711-0.842
Prescriptions	0.775	0.711-0.840
Vitals	0.768	0.703-0.835
Procedures	0.755	0.689-0.821
Demographics and baseline scores	0.599	0.515-0.679
All features except baseline scores	0.770	0.707-0.832

Next we performed an ablation analysis for the drop outcome where we removed a single set of features available to the ensemble voting model and then evaluated the test performance. Intuitively, larger drops in AUROC indicate that the feature was important for the performance of the full model for the prediction task. Results are shown in Table 2. A confidence interval was constructed by bootstrapping the test set. We note that the pre-treatment PROMIS scores alone showed remarkable predictive power in this task (AUROC 0.770), and additional features perhaps add limited benefit to the prediction task.

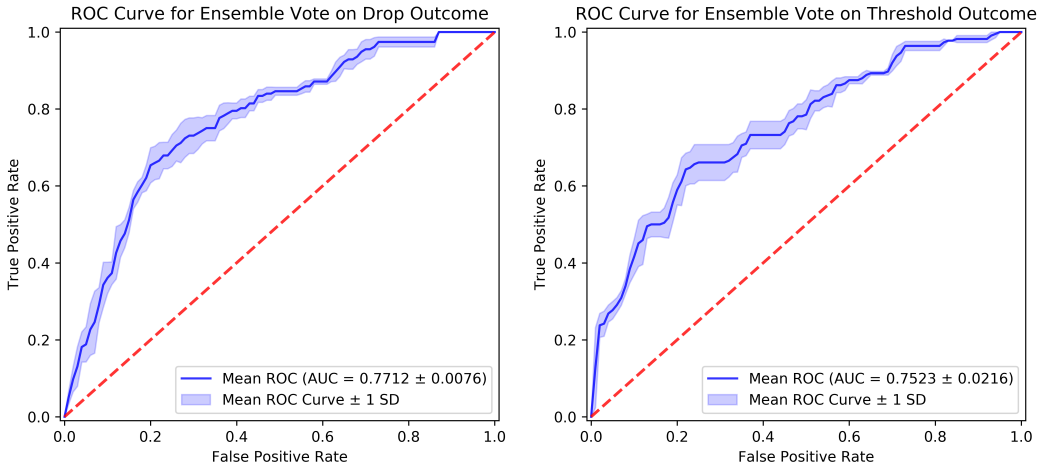

Figure 1: ROC curves for the Ensemble Voting models for the two outcomes. After being fit on the training data, the models predicted on 5 folds of the testing data (i.e. 80% of the testing data per fold). The mean ROC curve of those predictions (± 1 SD) is plotted above.

Discussion:

Ultimately, we were able to generate a model that achieved a reasonable AUROC (0.7713 ± 0.0076). To find the superior model in terms of AUROC performance, we extensively tested a variety of feature embeddings and model architectures. We found that training using imputed labs performed better than binning. This is perhaps because the imputed labs features contained continuous data rather than the binary bins, and therefore provided more information to the models. However, this method has not traditionally been employed in health care predictions as clinically it does not "make sense" to think about certain lab values for certain patients (e.g. women should not have Prostate Specific Antigen levels as they do not have prostates). Furthermore, as mentioned previously, additional information can be gleaned from a variable being missing. As such, perhaps a combined approach, providing the value when present and indicating its missingness when not, could outperform both of our lab featurization schemes.

For both full feature sets, an ensemble voting model performed best for both outcomes. This has been seen in other applications and is likely due to the ensemble voting being able to leverage and combine the orthogonal advantages of each constituent model to produce a superior estimate.

In terms of the relative importance of the feature types for our superior model, we saw AUROCs that tended to increase from the full feature set as we ablated features. This was contrary to what we were expecting, and perhaps raises the concern that our models were overfit (with the number of features approaching the number of observations), despite regularization. However, the confidence intervals for the ablated AUROCs overlap with the full feature AUROC. To investigate this further, we could perform additional ablation testing, such as randomly limiting to some subset of features, but this would be computationally quite expensive. Finally, it is important to note much of the predictive power of our model is found in a few features - the baseline PROMIS scores of the patient. This raises fears of data leakage; however, clinicians have confirmed that these data are routinely available at our time of prediction, the first date of chemotherapy administration.

6 Conclusion and Future Work

In this project, we were able to use electronic health record data generated before a patient's first administration of chemotherapy to predict if they would experience a decline in their subjectively reported physical health after starting treatment. We were able to develop a model that performed well enough to potentially have clinically utility. Our work expands the limited study of predicting changes in PROs in response to chemotherapy.

In the future, we are going to attempt to increase the sample size by relaxing some of the inclusion criteria, with physician input to capture a larger and more diverse cohort. We will further continue to refine our featurization and data cleaning methods. We will also try to further refine the hyperparameter searches for our voting model's constituents, and may try adjusting the relative weight that each constituent model has in the final vote. Finally, we hope to continue to collaborate with Stanford oncologists and potentially deploy this model in the clinic in the future to test the utility of the predictions in a real-world setting.

7 Contributions

This project was performed with the help of Dr. Hernandez-Boussard (School of Medicine, BMI) and Dr. James Brooks (School of Medicine, Urology). The two assisted with defining the prediction task as well as obtaining access to the data and as the Nero computational servers on which the project was run. They approved of our featurization schemes, but had no role in our modeling approaches or experimental design.

Both partners were involved in the project design, with the input of the aforementioned faculty. Dylan took the lead in querying the STARR database for our data and designing the experiments, while Nicolai took the lead in developing our predictive models. They both were involved with data cleaning: Nicolai led the cleaning and featurization of the vitals, diagnoses, medication, procedures, and binned lab data; while Dylan led the cleaning and featurization of the demographic, utilization, KNN imputed lab, and outcome data. Each ran experiments on the Nero servers and assisted with making figures and tables, with Dylan leading experiment 1 and Nicolai leading experiment 2. Finally, both members were involved with writing the project report.

References

[1] "Cancer - world health organization fact sheet," (2018). `https://www.who.int/news-room/fact-sheets/detail/cancer`.

[2] P. A. Ganz, K. A. Desmond, B. Leedham, *et al.*, "Quality of life in long-term, disease-free survivors of breast cancer: a follow-up study," *Journal of the National Cancer institute* **94**(1), 39–49 (2002).

[3] E. Basch, A. M. Deal, M. G. Kris, *et al.*, "Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial," *Journal of Clinical Oncology* **34**(6), 557 (2016).

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, *et al.*, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal* **13**, 8–17 (2015).

[5] W. Ranasinghe, D. de Silva, T. Bandaragoda, *et al.*, "Robotic-assisted vs. open radical prostatectomy: A machine learning framework for intelligent analysis of patient-reported outcomes from online cancer support groups," in *Urologic Oncology: Seminars and Original Investigations*, **36**(12), 529–e1, Elsevier (2018).

[6] X. Pan, R. Levin-Epstein, J. Huang, *et al.*, "Dosimetric predictors of patient-reported toxicity after prostate stereotactic body radiotherapy: Analysis of full range of the dose-volume histogram using ensemble machine learning," *Radiotherapy and Oncology* (2020).

[7] J. Wang, S. Jacobs, D. A. Dewalt, *et al.*, "A longitudinal study of promis pediatric symptom clusters in children undergoing chemotherapy," *Journal of pain and symptom management* **55**(2), 359–367 (2018).

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).