Longitudinal Quality of Life Analysis of Cancer Survivors

Hallie Ertman, Naman Sharma, Adrian Tullock, Riley Waters

INTRODUCTION

Fred Hutchinson Cancer Research Center has created a centralized cancer survivorship data system called SIMS. The system compiles **clinical diagnoses** and **survey data** from survivors collected prior to their first clinic visit and during annual follow ups.

Clinicians actively use SIMS for patient evaluation needs, but no major statistical analyses have been performed on it. The purpose of this project is to gain analytical insights that may improve the clinical process for cancer patients and better understand how cancer affects their lives. Specifically, we addressed the following questions:

> *Can quality of life indicators be accurately predicted using the survey data?*

> *What associations can be found between patient characteristics and health behaviors?*

> *What common trends exist when comparing the baseline and follow-up surveys?*

DATASET OVERVIEW

> **934** Clinical Patients

> **756** Baseline Survey Responses

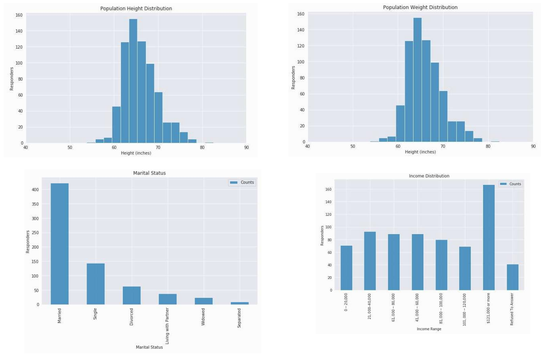
> **463** Follow-up Responses

> **800+** Baseline Survey Features

> **400+** Follow-up Survey Features

> **3000+** Clinical Features

EXPLORATION



Population overview of demographic variables

METHODOLOGY

MODEL CHOICES

> Random forests were chosen for their ability to handle thousands of input variables, extract the importance of features, and handle missing data estimation.

> Linear models were used to inspect the direct relationships of highly correlated variables.

> To construct linear models, the correlation of all variables in the dataset to each quality of life variable was captured (using a simple linear regression, chi square, PPMC, or multiple tests as appropriate). The highly correlated variables were inspected and some highly related variables (e.g. multiple measures of the same mental health symptom) were removed. Models were then constructed using forward selection.

PIPELINE PROCESS

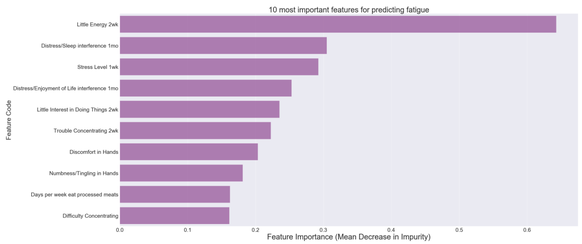
Merge raw datasets on patient id → **Map** feature codes to question text → **Drop** low-fill and zero-variabilty features → **Encode** categorical data → **Remove** outliers above 3 SD of the mean → **Standardize** numerical data after log-transform

Build other models with most important features ← **Build** Random Forest to extract feature importance ← **Balance** data with SMOTE ← **Identify** interaction terms with correlation matrix ← **Collapse** some features to categories ← **Record** new features of interest

> Different models were made for response variables relating to restlessness, feelings of depression, moods and worries, lifestyles, fatigue, and difficulty of concentration

> Data balancing was only performed in cases where the response variable was highly skewed

> Two sets of features were tested to construct the random forest - one detailed set that included most of the preprocessed data and one summary set that included categorized features and aggregated statistics (e.g. number of total medications)



An example of the important features found in predicting fatigue with random forest

RESULTS

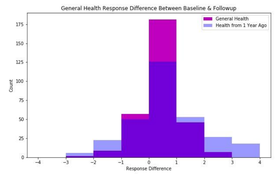
KEY FINDINGS

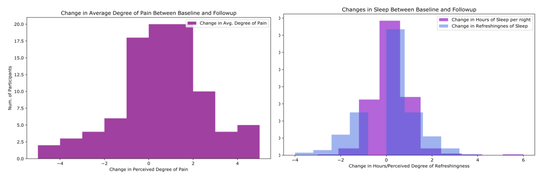
> Some surprising categories proved to be generally poor predictors of quality of life, including sexual functioning and pain.

> Unsurprisingly, many mental health-related symptoms are highly correlated; those related to depression and anxiety.

> Several predictor categories, including sleep and reported moods, were similarly predictive of many quality of life measures.

> Interestingly, most general health responses did not change by much, if at all.





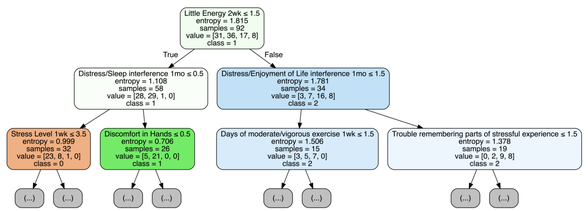
Changes in Selected Pain and Sleep Related Indicators Between Baseline and followup

MODEL EVALUATION

> The random forest model achieved ~80% accuracy in predicting categorical indicators of restlessness, depression, moods, lifestyles, fatigue, and concentration

> The detailed feature set did not increase the accuracy substantially over the summarized feature set

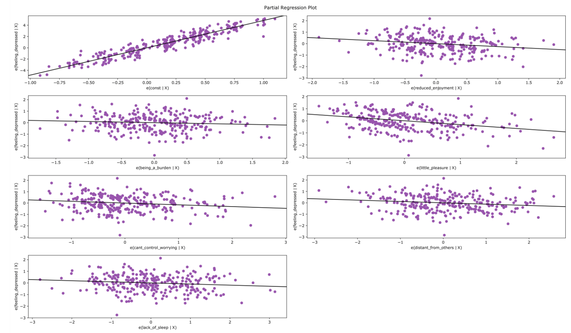
> All models depended primarily on a small handful of highly important variables that were intuitively correlated with the indicator



The first two levels of a decision tree used in the random forest for predicting fatigue.

> The most successful models in terms of prediction tended to use mental health-related predictors. However, models incorporating sleep and other physical predictors had arguably more utility, as the frequent comorbidity of many mental health symptoms is well established.

> One OLS model using a patient's feeling that he or she is a burden, feeling distant from others, reported inability to control worrying, taking little pleasure in general life experiences, lack of sleep, and reduced enjoyment of previously enjoyed activities was able to predict approximately 60% of the variance in a patient's likelihood of feeling depressed.



Partial regression plots from a model predicting a patient's liklihood of feeling depressed

CONCLUSIONS

> Quality of life indicators can be accurately predicted with the survey data, but the models rely on features that are sometimes obviously correlated.

> Many correlations were uncovered, but most of them are well established. Interestingly, some domains of responses such as pain were not found to be predictive of quality of life indicators

> Clinical data was not used extensively due to data sparsity and lack of strong predictors. It contains patients with many highly distinct clinical histories, such that binning them results in very small groups. As this project continues, several clinical categories of interest, such as those based on cancer type, are likely to grow. Revisiting the clinical data is recommended.

> Several other limitations of this data set are likely to be ameliorated by time and continued data gathering. The follow-up survey data, currently a small number of observations, should grow.

FRED HUTCH