

# Feature Engineering and Outcome Prediction in Multiple Myeloma Patients

MA 679 Final | Trinity Partners Team Three

Albert Ding, Emma Zhang, Xiangliang Liu

# Table of Contents

Abstract.....	3
Introduction.....	3
Methods.....	4
Data.....	5
EDA.....	7
Model.....	10
Discussion.....	16
Limitations.....	17
Future Direction.....	18
Reference.....	19
Appendix.....	20

# Abstract

Multiple myeloma (MM) is a form of cancer diagnosed in 30,000 new patients in the US annually with a five year mortality rate of almost 50%. Researchers have yet to find a cure, but recent advances in cancer research have made it a treatable disease. To better understand these treatments, we modeled and predicted the outcome of remission or relapse in MM patients based on a feature set that included both demographic factors and specific treatment regimens. Using methods of interpretable AI, we framed our classification analysis to examine the efficacy of various treatments. We also compared two approaches of feature selection/engineering: 1) manually constructed treatment variables drawn from medical literature and 2) automatic feature selection of treatments using the deep learning implementation from Tensorflow on a combined feature set inclusive of both the engineered feature treatments and the raw HCPCS codes. Our key findings were three-folds:

- 1) When comparing our model with manually constructed feature sets against a model with automatically selected features, we found that the model with manually constructed features outperformed the one with automatically selected features.
- 2) Bortezomib, a component treatment within induction therapy, had by far the highest variable importance from a treatment feature perspective, which is consistent with both the existing literature and the findings of other groups' projects within MSSP; Bortezomib was sometimes referenced in its raw Healthcare Common Procedure Coding System (HCPCS) code form as "J9041" in those projects. <sup>(1)</sup>
- 3) Random forest model had the highest test accuracy (72%) of the three models we built, possibly due to treatment algorithm structure and outperformed our null model based on the high proportion class (68% accuracy), our L1 regularized logistic regression (70%), and our deep learning model (57% accuracy after specifying 10 epochs due to memory constraints on the computing server).

## Introduction

Multiple myeloma (MM) is a cancer that forms in white blood cells known as plasma cells, which recognize and attack germs by forming antibodies. This process causes cancer cells to accumulate in the bone marrow, where they crowd out healthy blood cells and produce abnormal proteins that cause complications.

Currently, a dearth of understanding exists for selecting the right regimen for patients with MM. Choosing the right treatments involves factoring in a holistic consideration of attributes such as disease dynamics, patient comorbidities, experience on first line treatment, and other idiosyncrasies that vary from subpopulation to subpopulation or even patient to patient. The purpose of our analysis was to shed light on this topic by building models to predict whether a

patient with MM will go into remission or relapse. We have applied this analysis on a sample of patients records from Medicare data between 2013 to 2017.

This dataset was chosen because Medicare is part of the largest insurance payer in the US and covers in some form everyone 65 or older, those younger than 65 with certain disabilities, and those with end-stage renal disease (ESRD). Though it is not a perfect proxy of the general American populace, Medicare claims are likely the largest and most representative source of healthcare outcomes and treatments available. Medicare along with other electronic health records (EHR) and have spawn a growing community of both researchers in academics and industry.

## Methods

Using data supplied by Trinity Partners, we sought to build a predictive model that could determine whether a patient would go into relapse or remission. After merging a series of four tables, each with different claims information, we filtered down the dataset to include only patients who had MM including the entire claims history of patients in remission and the treatment history prior to relapse for those in relapse.

Once we produced our final dataframe of non-relapsed MM patients, we narrowed down the feature set by excluding redundant and irrelevant variables such as revenue information from diagnostic test claims. We noticed that the features we were interested could be bifurcated into two broad categories: demographic data and treatment data with the latter shown as either an HCPCS codes or CPT codes.

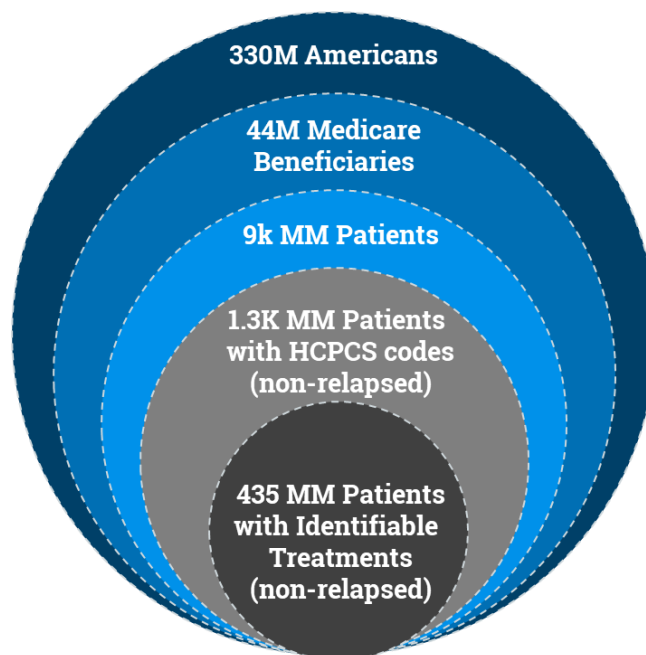
Next, we visualized the distribution of the HCPCS codes and saw that many of the most common ones were irrelevant ones such as a routine venipuncture. In order to measure how effective different treatment regimens were, we would have to construct the treatment variables themselves out of the HCPCS codes, which function as the components of the treatment. To find the right groupings of HCPCS codes, we employed a domain-research driven method where we manually constructed the variables and an automatic machine learning method which involved funneling the raw HCPCS values, as well as the engineered features, into a deep learning model to see how the model would use that data.

From there, we visualized the proportions of patients in relapse and remission under each treatment and under different demographic variables such as age, gender, and sex. Using both demographic from the dataset and treatment-related features we engineered, we built three predictive models using L1 regularized logistic regression, random forest, and deep learning. With the aid of our model results and follow-up analysis, we interpreted the results and synthesized our three main finding mentioned in the abstract.

# Data

This initial dataset was obtained from Trinity partners who supplied us with a set of approximately 9000 MM patients and a random sample of 5,000 non-MM patients. This dataset was derived from a five percent random sample of all Medicare patients covered under Part A and Part B but not Part C Coverage which is administered by private insurance companies contracted with Medicare. These dual models typically cover everything that the Original Medicare (Part A and Part B) cover and may offer extra benefits as well.

We noted that this dataset is likely not representative of the general US population due to the multiple layers of nested sampling that occurred to produce it. We have illustrated below this multi-step sampling process with choices we made in gray and choices outside of our control in blue <sup>(2)</sup>:



Upon further investigation, we realized that CPT codes only contained broad categories of treatment procedures. As a result of this constraint, we had to narrow the dataset to only MM patients who had at least one HCPCS code within his or her claim observations. Further complicating the issue was the fact that even within these patients, treatments regimens might not be identifiable due to the exclusive presence of claims that are broadly prevalent such as blood tests. We have prepared a brief table below illustrating this distribution to contextualize why we were forced to further pare down the set from 1305 patients to 435:

	1-5 claims	5-10 claims	10+ claims
Number of patient IDs	403	145	760
Percent of patient IDs	31%	11%	58%

From there, we filtered the data down to include only patients with MM and created variables for each treatment regimen based on a research methodology supplied to us by Philip Coombs, a research librarian specializing in data management and literature searching at the Boston University Alumni Medical Library. Philip helped guide us to several valuable resources including BMJ Best Practices, one of the premier clinical decision support tools for healthcare professionals. Within their database, we found a treatment algorithm that healthcare practitioners use in the context of treatment formulation. These different options were presented in a format akin to a decision tree with splits depending on the patient profile and previous health history. Using the components described in each treatment, we were able to map out the treatment algorithm to different HCPCS codes and used that reference to engineer our own set of treatment features derived from those codes:

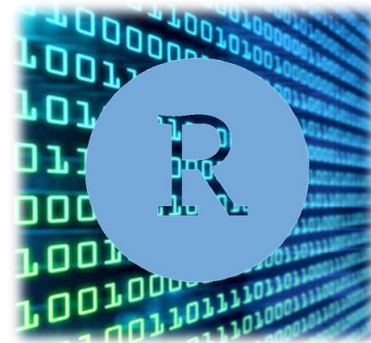
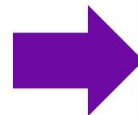
### Treatment algorithm

Please note that formulations/routes and doses may differ between drug names and brands, locations. Treatment recommendations are specific to patient groups: [see disclaimer](#)

**ACUTE**

newly diagnosed transplant candidates (<65-70 years, good performance status)

- 1st line** ✓ induction therapy
- Plus** ✓ deep vein thrombosis prophylaxis
- Adjunct** ✓ stem cell mobilization
- Adjunct** ✓ conditioning regimen
- Plus** ✓ stem cell transplant
- Adjunct** ✓ supportive care
- bone disease
  - Plus** ✓ bisphosphonates or denosumab



Below is a snippet of a matrix we created by hand, converting 98 different treatment variables we encountered in the treatment algorithm to corresponding HCPCS code to approximate the information we gained from the treatment algorithm (presented in its entirety in the appendix):

Method	Primary?	Option # Name	Index	First Line	Plus	Adjunct	All	Category	Codes	Patient
induction therapy	primary	1 thalidomide	1	1	0	0	0	Acute	None	newly diagnosed trar
induction therapy	primary	1 dexamethasone	2	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
induction therapy	primary	2 lenalidomide	3	1	0	0	0	Acute	None	newly diagnosed trar
induction therapy	primary	2 dexamethasone	4	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
induction therapy	primary	3 bortezomib	5	1	0	0	0	Acute	J9041	newly diagnosed trar
induction therapy	primary	3 dexamethasone	6	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
induction therapy	primary	4 bortezomib	7	1	0	0	0	Acute	J9041	newly diagnosed trar
induction therapy	primary	4 thalidomide	8	1	0	0	0	Acute	None	newly diagnosed trar
induction therapy	primary	4 dexamethasone	9	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
induction therapy	primary	5 lenalidomide	10	1	0	0	0	Acute	None	newly diagnosed trar
induction therapy	primary	5 bortezomib	11	1	0	0	0	Acute	J9041	newly diagnosed trar
induction therapy	primary	5 dexamethasone	12	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
induction therapy	primary	6 doxorubicin	13	1	0	0	0	Acute	None	newly diagnosed trar
induction therapy	primary	6 bortezomib	14	1	0	0	0	Acute	J9041	newly diagnosed trar
induction therapy	primary	6 dexamethasone	15	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
induction therapy	primary	7 bortezomib	16	1	0	0	0	Acute	J9041	newly diagnosed trar
induction therapy	primary	7 cyclophosphamide	17	1	0	0	0	Acute	J9070, J85	newly diagnosed trar
induction therapy	primary	7 dexamethasone	18	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
induction therapy	primary	8 vincristine	19	1	0	0	0	Acute	None	newly diagnosed trar
induction therapy	primary	8 doxorubicin	20	1	0	0	0	Acute	None	newly diagnosed trar
induction therapy	primary	8 dexamethasone	21	1	0	0	0	Acute	J8540, J111	newly diagnosed trar
deep vein thrombosis prophylaxis	primary	9 aspirin	22	0	1	0	1	Acute	None	newly diagnosed trar
deep vein thrombosis prophylaxis	secondary	9 enoxaparin	23	0	1	0	1	Acute	J1650	newly diagnosed trar
stem cell mobilization	primary	10 filgrastim	24	0	0	1	1	Acute	J1442, Q51	newly diagnosed trar
stem cell mobilization	primary	10 plerixafor	25	0	0	1	1	Acute	J2562	newly diagnosed trar
conditioning regimen	primary	11 melphalan	26	0	0	1	1	Acute	J9245	newly diagnosed trar
stem cell transplant	primary	12 stem cell transplant	27	0	1	0	1	Acute	38241	newly diagnosed trar
supportive care	primary	13 supportive care	28	0	0	1	1	Acute	None	newly diagnosed trar
bisphosphonates or denosumab	primary	14 pamidronate	29	0	1	0	1	Acute	J2430	newly diagnosed trar
bisphosphonates or denosumab	primary	15 zoledronic acid	30	0	1	0	1	Acute	J3489, J34	newly diagnosed trar
bisphosphonates or denosumab	secondary	16 denosumab	31	0	1	0	1	Acute	J0897	newly diagnosed trar
analgesics	primary	17 acetaminophen	32	0	0	1	1	Acute	None	newly diagnosed trar
analgesics	primary	18 codeine sulfate	33	0	0	1	1	Acute	None	newly diagnosed trar
analgesics	primary	19 morphine sulfate	34	0	0	1	1	Acute	J2270, J22	newly diagnosed trar
nontransplant therapy	primary	20 melphalan	35	1	0	0	0	Acute	J9245	newly diagnosed nor
nontransplant therapy	primary	20 prednisone	36	1	0	0	0	Acute	J7512	newly diagnosed nor
nontransplant therapy	primary	20 thalidomide	37	1	0	0	0	Acute	None	newly diagnosed nor
nontransplant therapy	primary	21 melphalan	38	1	0	0	0	Acute	J9245	newly diagnosed nor
nontransplant therapy	primary	21 prednisone	39	1	0	0	0	Acute	J7512	newly diagnosed nor

We used this data structure to creates a new feature space in our dataframe in R that would count the number of instances each patient was exposed to each treatment regimen identified; later, we would experiment with simple indicator variables for each as well (a binary treatment administered at least once or not) for each patient but found that this did not perform quite as well. This process finally helped us get to the final data frame which had a far wider structure that included all of the relevant demographic and contextual data from the original dataset as well as counts for the subset of the 98 treatments in the table for which we found a corresponding HCPCS code.

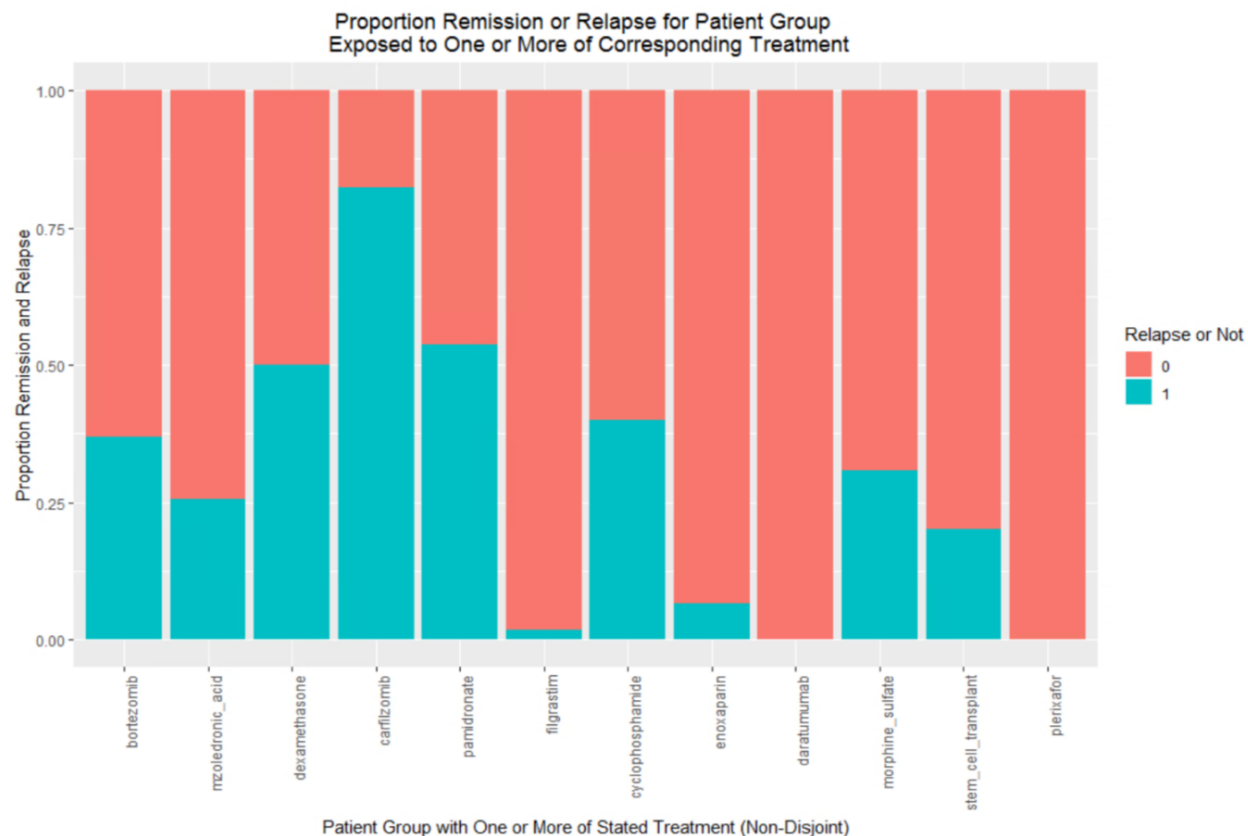
## Results

### EDA

In this section, we have focused on visualizing the associations between different treatments and demographic factors on the outcome of variable of relapse or remission, which can be visualized

as a sample proportion. The initial univariate plots intended to understand feature distributions can be found in the appendix.

Below, we have produced a visualization of the proportions remission and relapse for each patient exposed to at least one dosage of medication. Because some patients were exposed to multiple items within the treatments we engineered variables for, a patient's outcome may be included in the calculation of more than one proportion. Finally, we have been forced to exclude information in these summaries with less than 11 patients due to compliance guidelines, even though these patients were not removed from the observation pool used for modeling later/

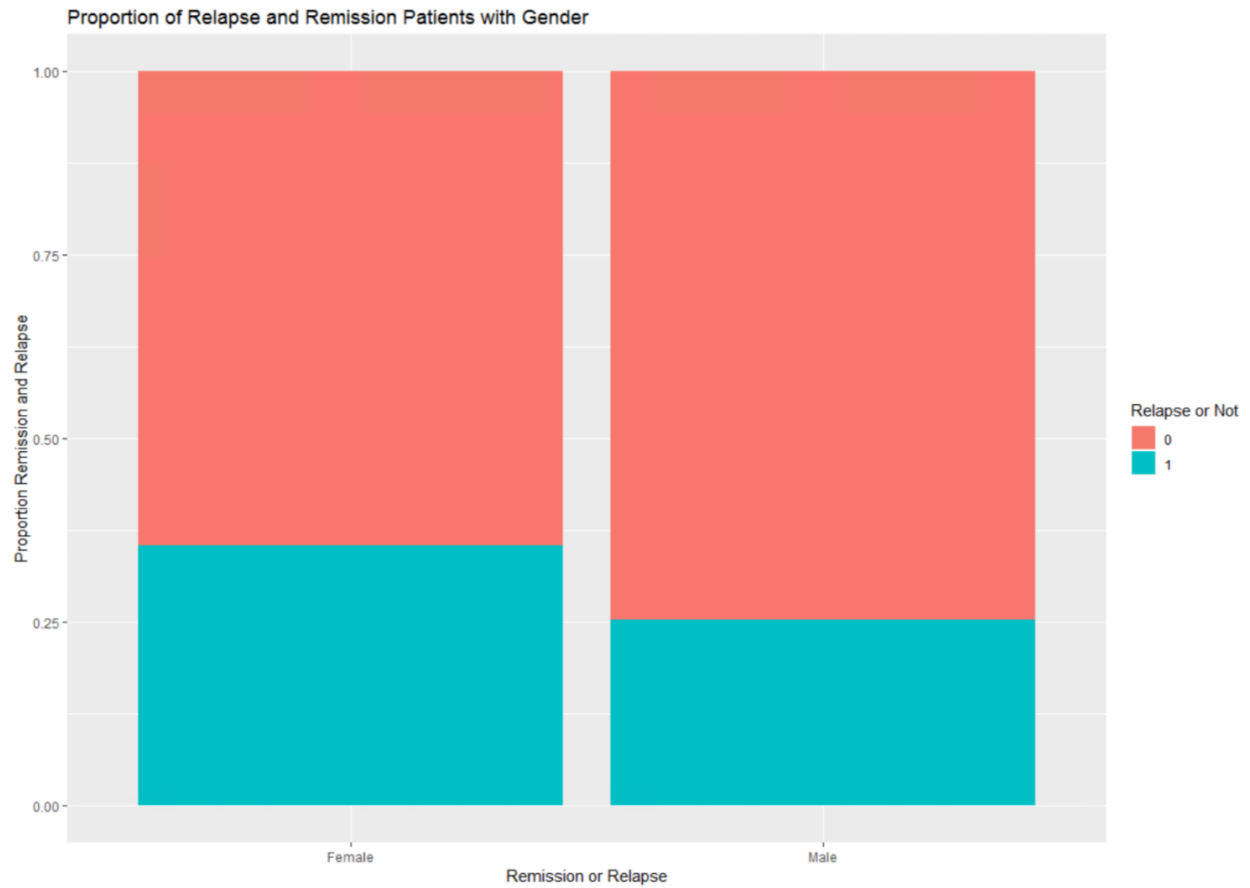


Treatment	Bortezomib	Zoledronic acid	Dexamethasone	Carfilzomib	Pamidronate	Filgrastim	Cyclophosphamide	Enoxaparin	Daratumumab	Morphine Sulfate
Instances of Treatment	1,279	708	200	169	119	54	25	15	14	13
Number of Patients*	114	251	44	16	43	25	16	15	14	12

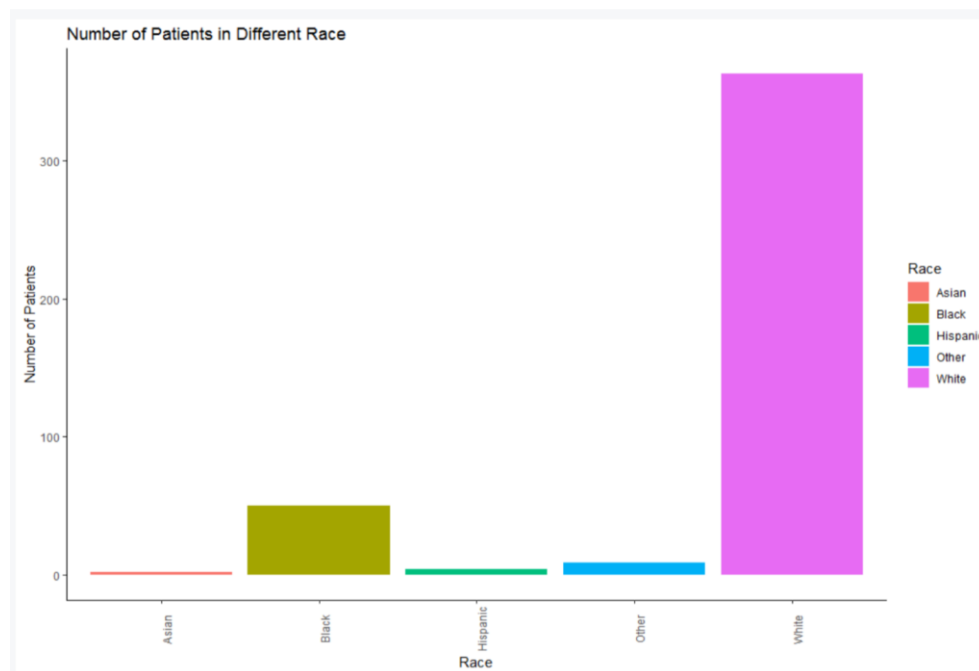
\*One or More Treatment Administered (Non-Nested)

Next, we turn to gender. We can see the female relapse proportion is slightly higher within the sample we ended up with:

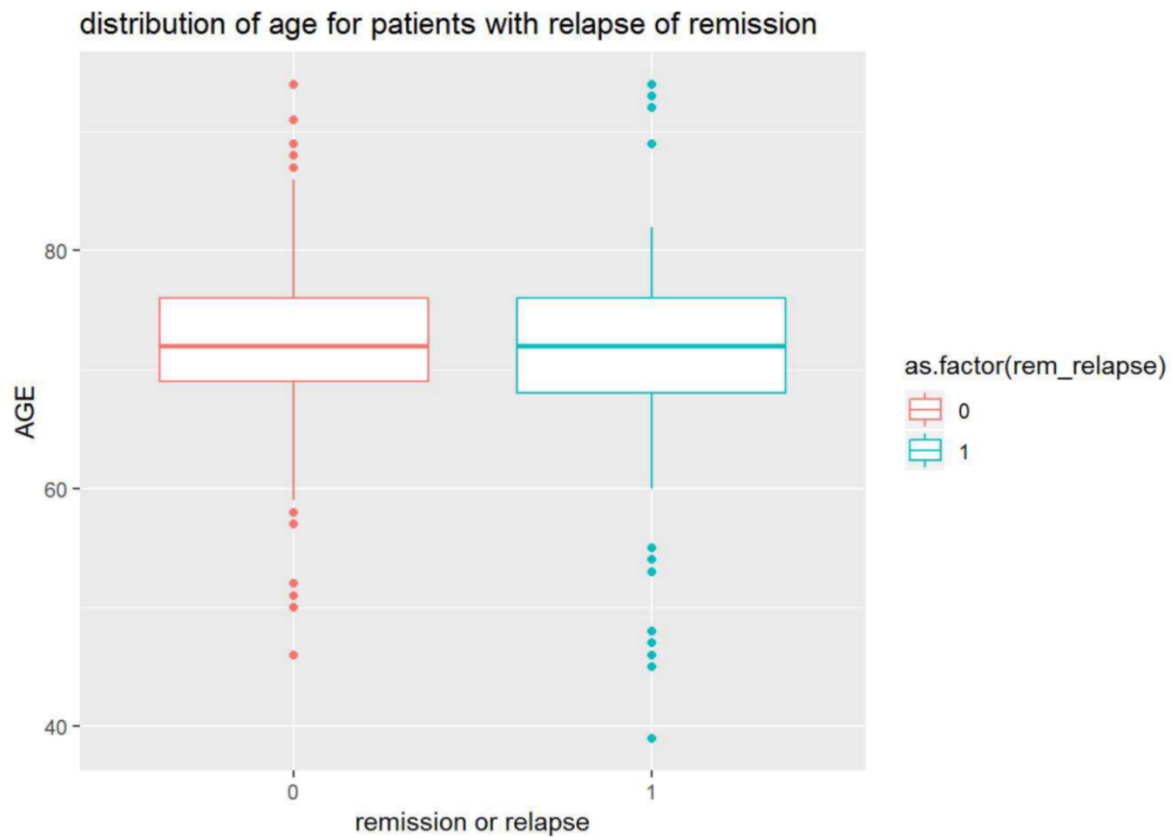




There is a lack of racial diversity in our sample as most patients are white, notably forming a disproportionate component of the sample relative to that of the overall population:



The patient age distributions between remission and relapse looks fairly similar based on an initial box plot visualization:



The boxplot indicates that MM patients who relapsed tend to have broader age range. And most patients have the MM disease around 70s. Using a regularized logistic regression model and random forest classifier, we predicted the outcome of treatment: relapse or remission and found that random forest performed better with a test set accuracy of 72%, which was relative improvement over the null model rate of 68%.

## Model:

Within this section, we used three methods to fit models and compare the result. All the classification models used whether MM patients will relapse as response variable with a 1 indicating a relapse and a 0 remission. In addition, we used treatment and demographic information about patients as predictors. For each treatment predictor, there's an indicator variable that summarized how many times the treatment was conducted to the same patients. For each model, we split the dataset into training set and testing set. We used training set to fit the model, and use testing set to test the accuracy of the model. There are 3 models in this section: Regularized Logistic Regression Model, random forest model, and deep learning method.

## L1 Penalized Logistic Regression (Lasso) Model

In this section, we used lasso logistic regression model to fit the model. We used 10-fold cross validation to tune the penalization hyperparameter represented by  $w$  below which is also the vector of coefficient values:

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^n \log(\exp(-y_i(X_i^T w + c)) + 1).$$

As a reminder to the reader, the  $\|w\|_1$  penalization term refers to least absolute deviations (LAD) or least absolute errors (LAE). This minimizes the sum of the absolute differences ( $w$ ) between the target value ( $Y_i$ ) and the estimated values ( $f(x_i)$ ):

$$\|w\| = \sum_{i=1}^n |y_i - f(x_i)|$$

The model gives us around 70% accuracy rate. The following are the model summary confusion matrix and accuracy rate below the model summary:

### Model Summary

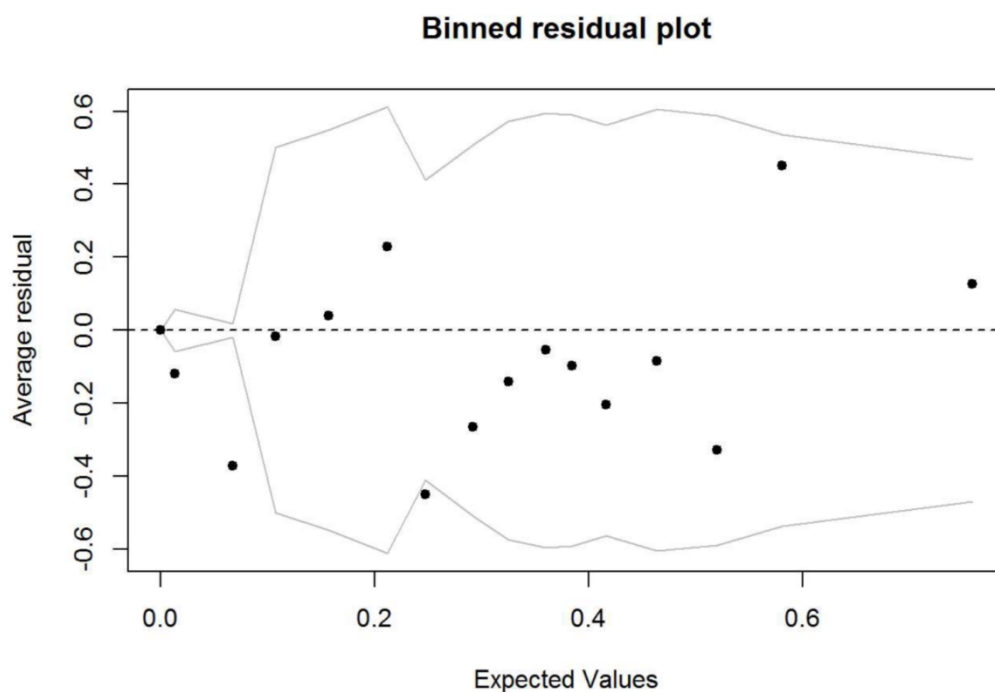
```
(Intercept) -0.67904096
Service_Year.x2014 0.15138260
Service_Year.x2015 .
Service_Year.x2016 -0.03392253
Service_Year.x2017 -1.09711927
Data_SetOP .
Data_SetPO .
AGE .
RaceAsian 1.00947300
RaceBlack .
RaceHispanic -0.14298976
RaceNorth American Native .
RaceOther 0.46422298
RaceWhite .
SexMale -0.02353906
bortezomib 0.01831362
carfilzomib .
cyclophosphamide .
daratumumab .
denosumab 0.82416709
dexamethasone .
elotuzumab .
enoxaparin .
filgrastim -0.26510270
melphalan .
morphine_sulfate .
pamidronate 0.25600464
plerixafor -0.31222100
prednisone .
stem_cell_transplant .
zoledronic_acid .
observed
predicted 0 1
0 118 49
1 2 3
[1] 0.7034884
```

Looking at the confusion matrix the model completely failed to pick up relapse patients, incorrectly predicting 49 of 52 cases for a true negative rate, or specificity, of 0.057. This

amounts to basically using the null model, predicting all observations as remission except for three patients.

We believe the vector of coefficient values given by the output can not be trusted due to the performance of the model. Our intuition and follow-up analysis indicate that limited sample size against which to measure contrast contributed to the unreliable and missing coefficient values provided by the model. For this reason, we don't believe that follow-up model visualizations such as partial dependence plots would yield insight, especially given the shoddy parameter estimation process resulting from the size and nature of the dataset.

The lambda used to test the model is 0.1244988, which maximized the accuracy. As we can see this is a small penalty term.



Though the binned residual plot indicates most points are inside the black lines, these estimates are misleading. The sparsity inside the treatment data is pretty significant because most of the treatments have very small sample size. The regularized logistics model may fail to capture those treatments features in the model.

## Random Forest

Given the decision tree structure of the treatment algorithm, we were optimistic that random forest might be a stronger model choice. Though this method would not have the advantage of

providing coefficient estimates or values, we could implement a variable importance plot to diagnose the effect of removing each variable on the analysis.

For our treatment, we used the remission as the outcome variable and imported other demographic features of patients and top 15 treatments as our predictors.

Random Forest

432 samples

21 predictor

2 classes: '0', '1'

No pre-processing

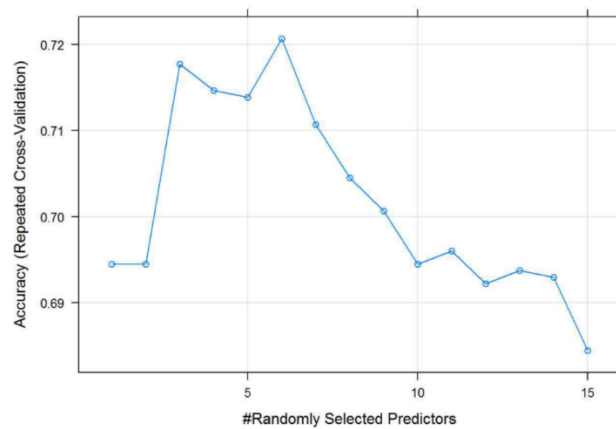
Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 389, 389, 389, 389, 389, 388, ...

Resampling results across tuning parameters:

mtry	Accuracy	Kappa
1	0.6945032	0.0000000
2	0.6945032	0.0000000
3	0.7177237	0.1055283
4	0.7146582	0.1427093
5	0.7138830	0.1846671
6	0.7207188	0.2287094
7	0.7106942	0.2116529
8	0.7045455	0.2099852
9	0.7006519	0.2006011
10	0.6945032	0.1944359
11	0.6960359	0.1975933
12	0.6921776	0.1913422
13	0.6937104	0.1967740
14	0.6929352	0.1955635
15	0.6844080	0.1739597

We also used repeated cross validation method to find the best number of predictors with random forest. The following plot indicates that result. As we can see, random forest model achieved the highest accurate rate at 72% when 6 predictors were chosen.



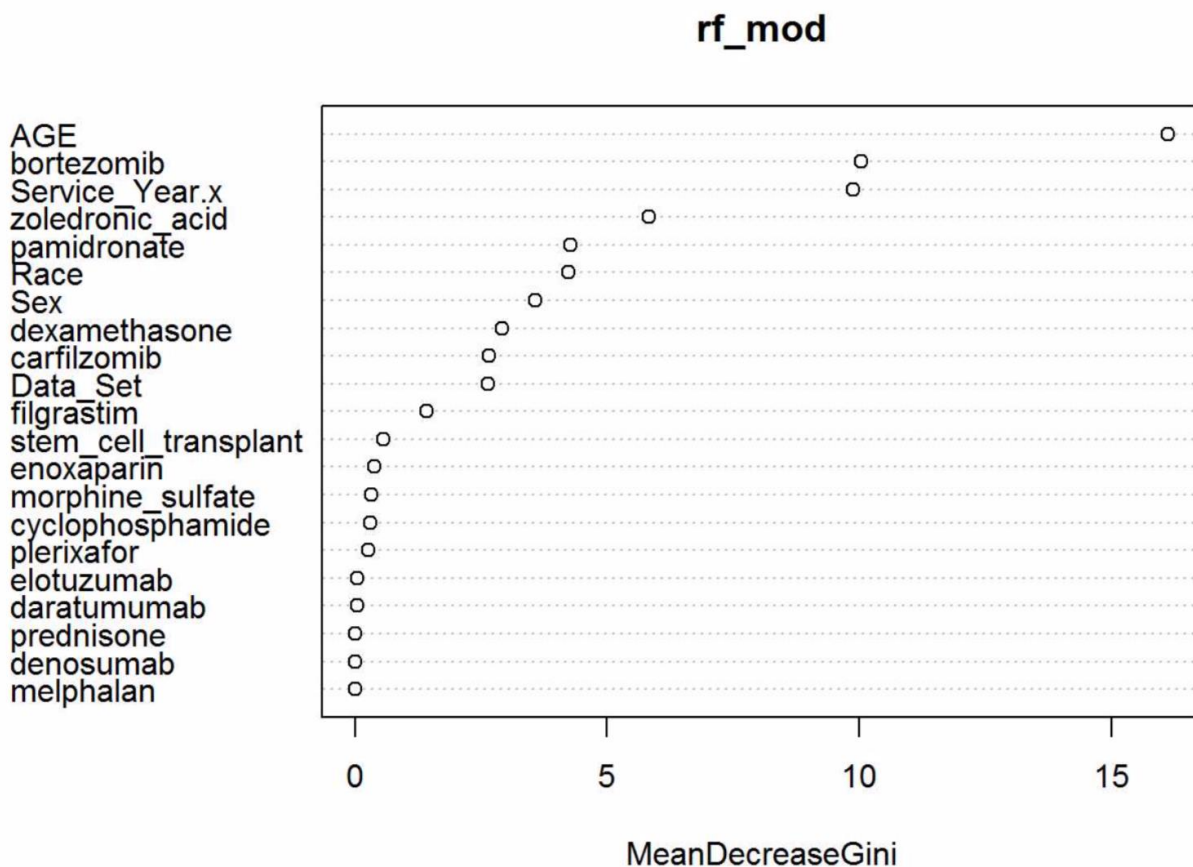
We also made the variable importance plot to check which predictors are the most influential ones. Given an impurity function such as the Gini index, we split at some node  $t$  if the change in the index  $i(\cdot)$  such as Gini index we split at some node  $t$  if the change in the index:

$$\Delta i(t) = i(t) - \frac{N_L}{N} i(t_L) - \frac{N_R}{N} i(t_R)$$

is significant, where  $t_L$  is the node on the left, and  $t_R$  the node on the right. It is possible to evaluate the importance of some variable  $X_k$  when predicting  $Y_k$  by adding up the weighted impurity decreases for all nodes  $t$  where  $X_k$  is used, averaged over all trees:

$$I(X_k) = \frac{1}{M} \sum_m \sum_t \frac{N_t}{N} \Delta i(t)$$

From the from the following plots we can see Age and service year are very important. This is because most patients have MM disease in their 70s. Also 2017 have more patient than the other years. Another very important feature is bortezomib. This is the most frequently used treatment within manually selected features.



We also run cross validation to confirm our result, the confusion matrix of Random Forest displays as follows. This represents a significant improvement over the lasso logistic regression and in particular generates a meaningful and large improvement in specificity:

```

      predicted
observed  0   1
      0 108  12
      1  37  15

```

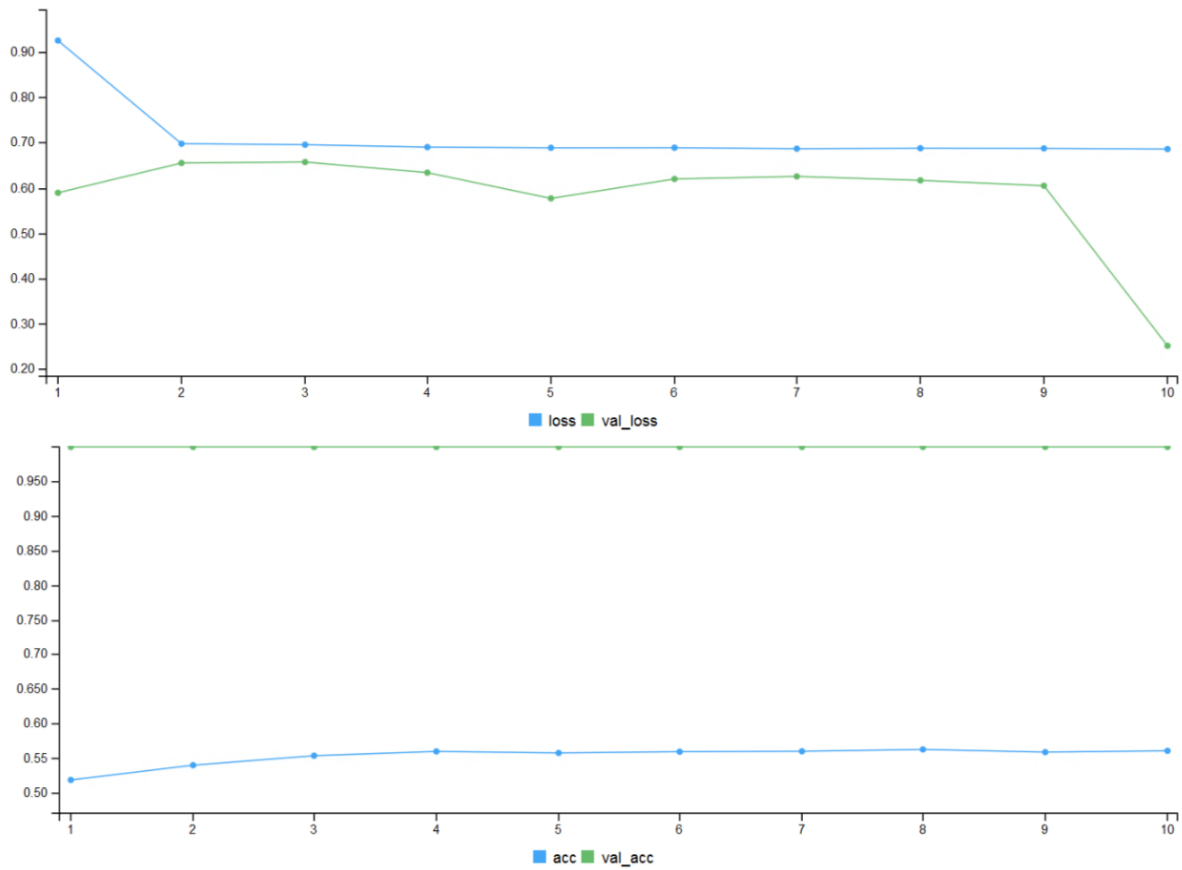
## Deep Learning Model

We ran a deep neural network using the Keras and Tensorflow packages and the full matrix of raw HCPCS codes as well as our engineered features. This amounts to giving the model the choice of selecting our engineered features or our raw features or any combination of both. In theory, a more powerful and flexible algorithm with a larger feature space should outperform in many situations but it does not here.

Due to time and computing limitations, we were only able to run ten epochs on the computing server the data was located on. We note that the predictive accuracy is perfect on the training set given the flexibility of the model but does terribly on the test set due to an excess of variance in parameter estimation.

We will refrain from extensive analysis or interpretation given that the main purpose of the model was just a strawman with which to measure our random forest with manual feature engineering against:

Layer (type)	Output Shape	Param #
=====		
dense (Dense)	(None, 256)	6912
-----		
dropout (Dropout)	(None, 256)	0
-----		
dense_1 (Dense)	(None, 128)	32896
-----		
dropout_1 (Dropout)	(None, 128)	0
-----		
dense_2 (Dense)	(None, 2)	258
=====		
Total params: 40,066		
Trainable params: 40,066		
Non-trainable params: 0		



## Discussion

When comparing our model with manually constructed feature sets against our model with automatically selected features, we found that the manual performed better. To further solidify this result, we note that random forest performed substantially more poorly when we were just as indiscriminate in including variables as we were with deep learning. This suggests its not the algorithm but the feature engineering that made a meaningful difference. We have prepared a summary table below:

	Null Model	L1 Logistic Regression	Random Forest	Deep Learning
<b>Accuracy</b>	68%	69%	72%	57%
<b>Feature Construction</b>	None	Manual	Manual	Automatic
<b>Challenges</b>	N/A	Sparsity in treatment of predictor space impedes estimation of reliable effect size interval	No estimate of effect size or confidence interval	Difficult to tune and slow / computationally expensive to run



From a treatment standpoint, Bortezomib, a component within induction therapy, had the highest variable importance of any identified measured treatment variables. This appears to validate our model in some sense given that “Bortezomib plus melphalan plus prednisone improved response rates up to 71% with significantly prolonged overall survival compared with melphalan plus prednisone alone in nontransplant candidates.”<sup>(3)</sup>

Finally, we found that our random forest model had the highest test accuracy (72%) and find that particularly interesting because the treatment algorithm structure doctors follow resembles a decision tree. We can compare this to our null model based on the high proportion class (68% accuracy), our L1 regularized logistic regression (69%), and our deep learning model (57% accuracy after specifying 10 epochs due to memory constraints on the computing server). We attribute the outperformance of the random forest model to a feature design matrix that was constructed only with meaningful treatment values within the HCPCS codes as well as the analytical similarities between the decision tree process of treatment selection and how each tree is grown within random forest.

Given the results of our initial analysis, we propose the creation of a cross-platform package or function that will automatically convert HCPCS code values to corresponding treatment variables. If such a package, were developed our intuition is that it would save both time spend on data cleaning and provide a code base upon which to build other ideas about how treatment values could be measured.

## Limitations

Although we have established a preliminary result that manual feature engineering may outperform the automatic feature selection, we cannot say this generalizes to every case. Applying deep learning models to a dataset with only 435 observations is, in itself, a bit absurd but meant as a proxy strawman test to challenge the notion that using numerical methods dissociated from the actual domain topic of study does not guarantee outperformance.

Also problematic are the different periods from which patient outcome measurements are taken. Each patient has a unique length of claims history and taken to the extreme, one might surmise that everyone might relapse should they live long enough. Typically, when academics and practitioners attempt this type of outcome inference or prediction from Medicare claims data, they would enrich the data with a range of other pertinent details from electronic health records, adding both breadth and depth longitudinally. Within the HCPCS codes we have captured as treatments, we also do not understand how the intensity or variation within each treatment path should be measured; for instance, inherent in the way we have coded the matrix, we assume that there is a linear relationship between the number of instances of receiving a treatment and the probability of remission or relapse on some scale.

Overall, comparisons and measurements of treatment effects on a population or sub-population level may not apply on a personal level. Finally, we cannot ignore the inherent bias of repeated nested sampling. The obtained sample is likely not representative of any population or sub-population of interest.

## Future Direction

If given more time, we would have liked to ideally reexamined data visualizations after each iteration of the model. This may have enabled us to further finetune our predictive accuracy. Of course, our manual feature engineering process could have benefitted from a more robust treatment algorithm as the input.

If given access to the expertise of licensed doctors, we could have enriched our codification of the predictor design matrix and may have had a more nuanced understanding of how to build the variables. Taking it one step further, we might propose that a platform-agnostic package that would automatically turn a set of HCPCS codes into the matching treatment regimens would likely reduce workload and processing time for many healthcare professionals touching data analysis.

Another area of further exploration is simply a more rigorously and formalized testing framework. We realized that our testing setup was quite informal and that an untuned deep learning model cannot be truly held as a proxy for the gold standard for what is possible, as far as feature selection and engineering goes.

One final detail we wish we had thought of while having access to the data was to quantify our variance explained and out-of-bag error for our chosen random forest model. Alternatively, another potential exploration would have been to implement a jackknife estimate of the standard error based on leave-one-out-cross-validation principles, which is already intricately connected with the two aforementioned principles anyways. This final option appeals to us in particular due to its direct methodology and novelty as a lesser used cousin to the bootstrap.

We present our understanding of the implementation of such a method while deferring to the expertise of Efron, the pioneer behind the bootstrap, and his accomplished pupil and fellow Stanford professor, Hastie in *Computer Age Statistical Inference* <sup>(4)</sup>:

If we let  $\mathbf{x}_{(i)}$  be the sample with  $\mathbf{x}_i$  removed,

$$\mathbf{x}_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)',$$

For any parameter  $\hat{\theta}$  the standard error of the jackknife estimate is

$$\widehat{\text{se}}_{\text{jack}} = \left[ \sum_{i=1}^n (x_i - \bar{x})^2 / (n(n-1)) \right]^{1/2},$$

The advantage of  $\widehat{\text{se}}_{\text{jack}}$  is that the definition can be applied in an automatic way to any statistic  $\hat{\theta} = s(x)$ . All that is needed is an algorithm that computes  $s(\cdot)$  for the deleted data sets  $x_{(i)}$ . We note the following features of the jackknife:

- It is nonparametric with no special form of the underlying distribution  $F$  need be assumed.
- It is completely automatic: a single algorithm can be that inputs the data set  $x$  and function,  $s(x)$  and outputs  $\widehat{\text{se}}_{\text{jack}}$ .
- The algorithm works with data sets of size  $n - 1$ , not  $n$ . There is a hidden assumption of smooth behavior across sample sizes.
- The jackknife standard error is upwardly biased as an estimate of the true standard error.
- The connection of the jackknife formula with Taylor series methods is closer than it appears. We can write:

$$\widehat{\text{se}}_{\text{jack}} = \left[ \frac{\sum_1^n D_i^2}{n^2} \right]^{1/2}$$

where  $D_i$  is the approximate directional derivatives, measures of how fast the statistic  $s(x)$  is changing as we decrease the weight on data point  $x_{(i)}$ :

$$D_i = \frac{\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}}{1/\sqrt{n(n-1)}}$$

So  $\widehat{\text{se}}_{\text{jack}}$  is proportional to the sum of squared derivatives of  $s(x)$  in the  $n$  component directions and the corresponding Taylor series expressions amount to doing the derivatives by formula rather than numerically.

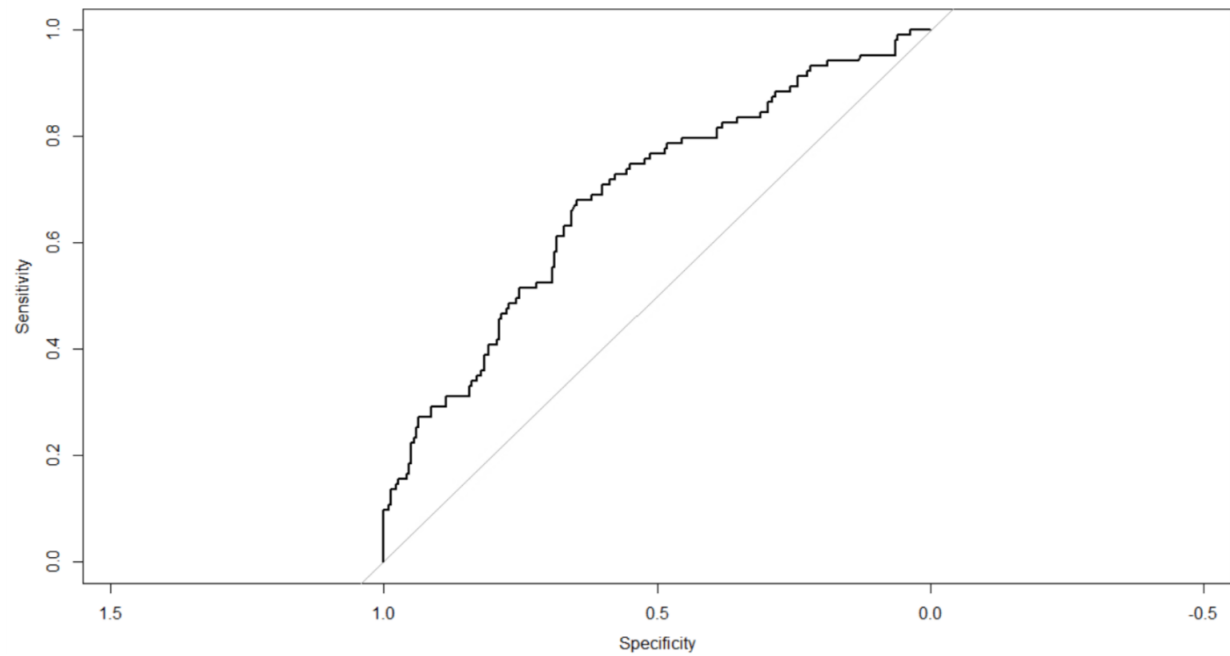
## Reference

- 1) “Inj., Velcade 0.1 Mg.” Healthcare Common Procedure Coding System, [hcpcs.codes/j-codes/J9041/](https://www.cms.gov/medicare/coding/hcpcs/codes/j-codes/J9041/).
- 2) “Philip Espinola Coombs.” Philip Espinola Coombs | Alumni Medical Library, [www.bumc.bu.edu/medlib/profile/philip-espinola-coombs/](http://www.bumc.bu.edu/medlib/profile/philip-espinola-coombs/).
- 3) “Bortezomib plus Melphalan and Prednisone for Initial Treatment of Multiple Myeloma | NEJM.” New England Journal of Medicine, [www.nejm.org/doi/pdf/10.1056/NEJMoa0801479](https://www.nejm.org/doi/pdf/10.1056/NEJMoa0801479).
- 4) Efron, Bradley, and Trevor Hastie. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press, 2017.

# Appendix

## Random Forest Model ROC curve

The following plot is the ROC curve with area under curve (AUC) 0.68:



# Treatment Design Matrix

Below is the design matrix we coded using the BMJ Journal database of treatment algorithms:

Method	Primary2	Option #	Name	Index	First Line	Plus	Adjunct	All	Category	Codes	Patient
induction therapy	primary	1	thalidomide	1	1	0	0	0	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	1	dexamethasone	2	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	2	lenalidomide	3	1	0	0	0	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	2	dexamethasone	4	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	3	bortezomib	5	1	0	0	0	Acute	J9041	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	3	dexamethasone	6	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	4	bortezomib	7	1	0	0	0	Acute	J9041	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	4	thalidomide	8	1	0	0	0	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	4	dexamethasone	9	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	5	lenalidomide	10	1	0	0	0	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	5	bortezomib	11	1	0	0	0	Acute	J9041	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	5	dexamethasone	12	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	6	doxorubicin	13	1	0	0	0	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	6	bortezomib	14	1	0	0	0	Acute	J9041	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	6	dexamethasone	15	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	7	bortezomib	16	1	0	0	0	Acute	J9041	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	7	cyclophosphamide	17	1	0	0	0	Acute	J9070, J8530	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	7	dexamethasone	18	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	8	vincristine	19	1	0	0	0	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	8	doxorubicin	20	1	0	0	0	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
induction therapy	primary	8	dexamethasone	21	1	0	0	0	Acute	J8540, J1100	newly diagnosed transplant candidates (<65-70 years, good performance status)
deep vein thrombosis prophylaxis	primary	9	aspirin	22	0	1	0	1	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
deep vein thrombosis prophylaxis	secondary	9	enoxaparin	23	0	1	0	1	Acute	J1650	newly diagnosed transplant candidates (<65-70 years, good performance status)
stem cell mobilization	primary	10	filgrastim	24	0	0	1	1	Acute	J1442, Q5101, J1442, J1441, J2505, J1440	newly diagnosed transplant candidates (<65-70 years, good performance status)
stem cell mobilization	primary	10	plerixafor	25	0	0	1	1	Acute	J2562	newly diagnosed transplant candidates (<65-70 years, good performance status)
conditioning regimen	primary	11	melfalan	26	0	0	1	1	Acute	J9245	newly diagnosed transplant candidates (<65-70 years, good performance status)
stem cell transplant	primary	12	stem cell transplant	27	0	1	0	1	Acute	J8241	newly diagnosed transplant candidates (<65-70 years, good performance status)
supportive care	primary	13	supportive care	28	0	0	1	1	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
bisphosphonates or denosumab	primary	14	pamidronate	29	0	1	0	1	Acute	J2430	newly diagnosed transplant candidates (<65-70 years, good performance status)
bisphosphonates or denosumab	primary	15	zoledronic acid	30	0	1	0	1	Acute	J3489, J3487	newly diagnosed transplant candidates (<65-70 years, good performance status)
bisphosphonates or denosumab	secondary	16	denosumab	31	0	1	0	1	Acute	J0897	newly diagnosed transplant candidates (<65-70 years, good performance status)
analgesics	primary	17	acetaminophen	32	0	0	1	1	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
analgesics	primary	18	codeine sulfate	33	0	0	1	1	Acute	None	newly diagnosed transplant candidates (<65-70 years, good performance status)
analgesics	primary	19	morphine sulfate	34	0	0	1	1	Acute	J2270, J2275	newly diagnosed transplant candidates (<65-70 years, good performance status)
nontransplant therapy	primary	20	melfalan	35	1	0	0	0	Acute	J9245	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	20	prednisone	36	1	0	0	0	Acute	J7512	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	20	thalidomide	37	1	0	0	0	Acute	None	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	21	melfalan	38	1	0	0	0	Acute	J9245	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	21	prednisone	39	1	0	0	0	Acute	J7512	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	21	bortezomib	40	1	0	0	0	Acute	J9041	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	22	melfalan	41	1	0	0	0	Acute	J9245	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	22	prednisone	42	1	0	0	0	Acute	J7512	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	22	bortezomib	43	1	0	0	0	Acute	J9041	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	22	daratumumab	44	1	0	0	0	Acute	J9145	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	23	bortezomib	45	1	0	0	0	Acute	J9041	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	23	lenalidomide	46	1	0	0	0	Acute	None	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	23	dexamethasone	47	1	0	0	0	Acute	J8540, J1100	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	24	thalidomide	48	1	0	0	0	Acute	None	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	25	lenalidomide	49	1	0	0	0	Acute	None	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
nontransplant therapy	primary	25	dexamethasone	50	1	0	0	0	Acute	J8540, J1100	newly diagnosed nontransplant candidates (>65-70 years and/or poor performance status)
maintenance strategies	primary	26	thalidomide	51	1	0	0	0	Ongoing	None	previously diagnosed patients responding to nontransplant therapy and/or transplant therapy
maintenance strategies	primary	27	bortezomib	52	1	0	0	0	Ongoing	J9041	previously diagnosed patients responding to nontransplant therapy and/or transplant therapy
maintenance strategies	primary	28	lenalidomide	53	1	0	0	0	Ongoing	None	previously diagnosed patients responding to nontransplant therapy and/or transplant therapy
maintenance strategies	primary	29	thalidomide	54	1	0	0	0	Ongoing	None	previously diagnosed patients responding to nontransplant therapy and/or transplant therapy
maintenance strategies	primary	30	bortezomib	55	1	0	0	0	Ongoing	J9041	previously diagnosed patients responding to nontransplant therapy and/or transplant therapy
maintenance strategies	primary	31	dexamethasone	56	1	0	0	0	Ongoing	J8540, J1100	previously diagnosed patients responding to nontransplant therapy and/or transplant therapy
maintenance strategies	primary	32	lenalidomide	57	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	primary	32	dexamethasone	58	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	primary	33	bortezomib	59	1	0	0	0	Ongoing	J9041	relapsing or refractory patients
maintenance strategies	primary	33	dexamethasone	60	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	primary	34	thalidomide	61	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	primary	34	dexamethasone	62	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	primary	35	carfilzomib	63	1	0	0	0	Ongoing	J9047	relapsing or refractory patients
maintenance strategies	primary	35	dexamethasone	64	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	primary	36	bortezomib	65	1	0	0	0	Ongoing	J9041	relapsing or refractory patients
maintenance strategies	primary	36	doxorubicin liposomal	66	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	primary	37	bortezomib	67	1	0	0	0	Ongoing	J9041	relapsing or refractory patients
maintenance strategies	primary	37	thalidomide	68	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	primary	37	dexamethasone	69	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	primary	38	carfilzomib	70	1	0	0	0	Ongoing	J9047	relapsing or refractory patients
maintenance strategies	primary	38	lenalidomide	71	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	primary	38	dexamethasone	72	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	secondary	39	pomalidomide	73	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	secondary	39	dexamethasone	74	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	secondary	40	pomalidomide	75	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	secondary	40	dexamethasone	76	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	secondary	40	cyclophosphamide	77	1	0	0	0	Ongoing	J9070, J8530	relapsing or refractory patients
maintenance strategies	secondary	41	carfilzomib	78	1	0	0	0	Ongoing	J9047	relapsing or refractory patients
maintenance strategies	tertiary	42	panobinostat	79	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	Tertiary options	42	bortezomib	80	1	0	0	0	Ongoing	J9041	relapsing or refractory patients
maintenance strategies	Tertiary options	42	dexamethasone	81	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	Tertiary options	43	daratumumab	82	1	0	0	0	Ongoing	J9145	relapsing or refractory patients
maintenance strategies	Tertiary options	44	daratumumab	83	1	0	0	0	Ongoing	J9145	relapsing or refractory patients
maintenance strategies	Tertiary options	44	lenalidomide	84	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	Tertiary options	44	dexamethasone	85	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	Tertiary options	45	daratumumab	86	1	0	0	0	Ongoing	J9145	relapsing or refractory patients
maintenance strategies	Tertiary options	45	bortezomib	87	1	0	0	0	Ongoing	J9041	relapsing or refractory patients
maintenance strategies	Tertiary options	45	dexamethasone	88	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	Tertiary options	46	daratumumab	89	1	0	0	0	Ongoing	J9145	relapsing or refractory patients
maintenance strategies	Tertiary options	46	pomalidomide	90	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	Tertiary options	46	dexamethasone	91	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	Tertiary options	47	elotuzumab	92	1	0	0	0	Ongoing	C9477	relapsing or refractory patients
maintenance strategies	Tertiary options	47	lenalidomide	93	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	Tertiary options	47	dexamethasone	94	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients
maintenance strategies	Tertiary options	48	kazomib	95	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	Tertiary options	48	lenalidomide	96	1	0	0	0	Ongoing	None	relapsing or refractory patients
maintenance strategies	Tertiary options	48	dexamethasone	97	1	0	0	0	Ongoing	J8540, J1100	relapsing or refractory patients

## Distribution of MM Patient in the United States

Texas, California and Florida have the greatest number of MM patients (and also the largest state populations):

