Outcome Prediction in Multiple Myeloma (MM) Patients

MA679 | May 2019

Albert Ding, Emma Zhang, Ian Liu

BOSTON UNIVERSITY | TRINITY PARTNERS

We employed methods from interpretable AI to measure efficacy of different treatment options for MM

Compared models with manually constructed feature sets against models with automatically selected features and **found that manual outperformed automatic features**

Bortezomib (induction therapy), zoledronic acid (bone disease patients), and **pamidronate (bone disease patients)** were treatments with high variable importance [1]

Random forest model had the highest test accuracy (72%), possibly due to treatment algorithm structure

(1) 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: J Clin Oncol. 2010 May 1;28(13):2259-66. doi: 10.1200/JCO.2009.26.0638. Epub 2010 Apr 5.

Multiple Myeloma

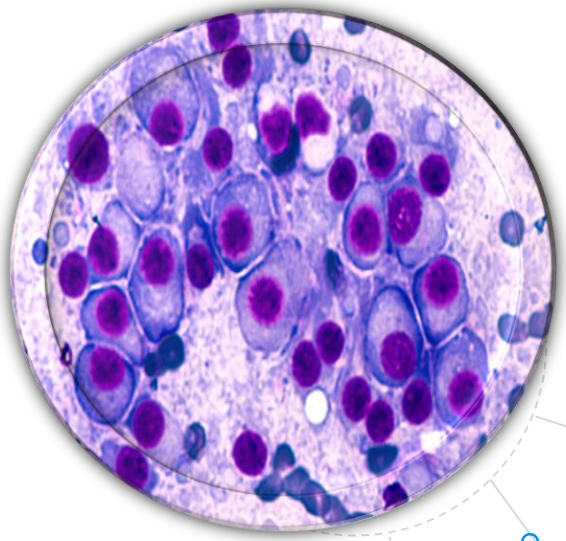
Cancer of White Blood Cells (Plasma)

Symptoms include:

loss of appetite, bone pain, fever

Treatments include:

medications, chemotherapy, corticosteroids, radiation, stem-cell transplant



30,000/year

Newly Diagnosed US Patients

>50%

Five year mortality rate

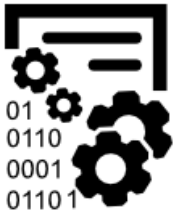
0

Cures for MM

Which treatment regimens work?

...But first how do we measure that?

Executed the full data science lifecycle to explore how to measure treatment efficacy



Data Cleaning

EDA

Domain Research

Feature Engineering

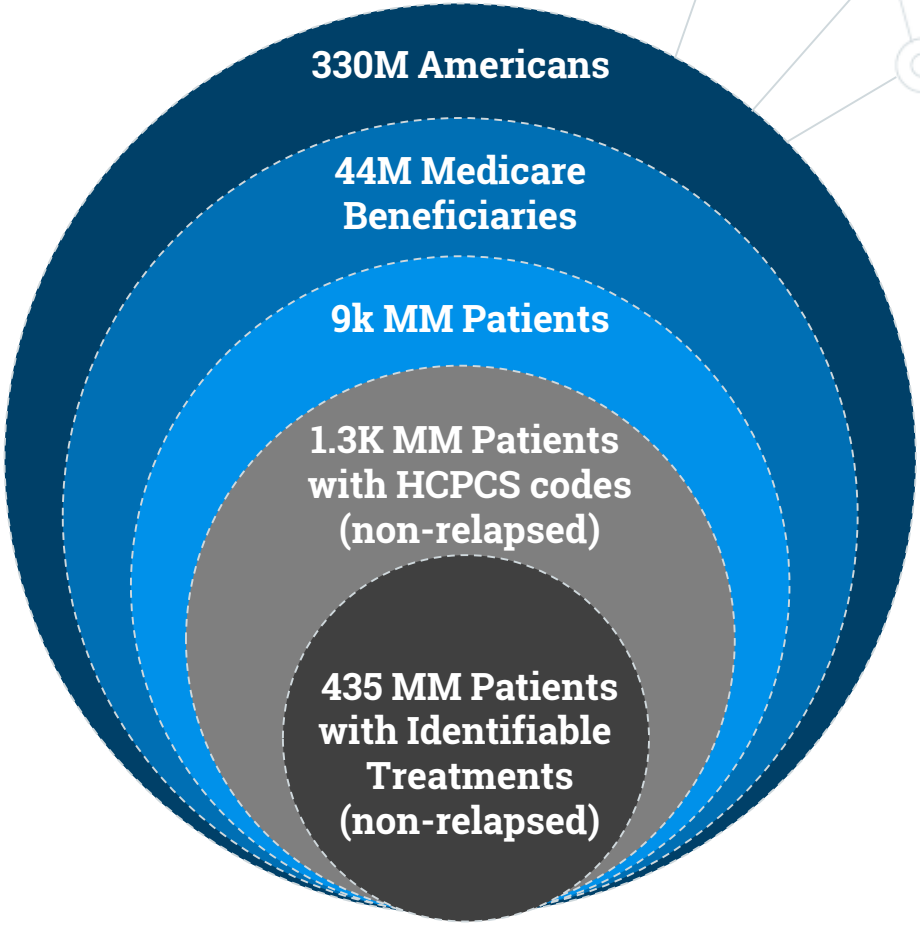
Modeling

Interpretation of Results

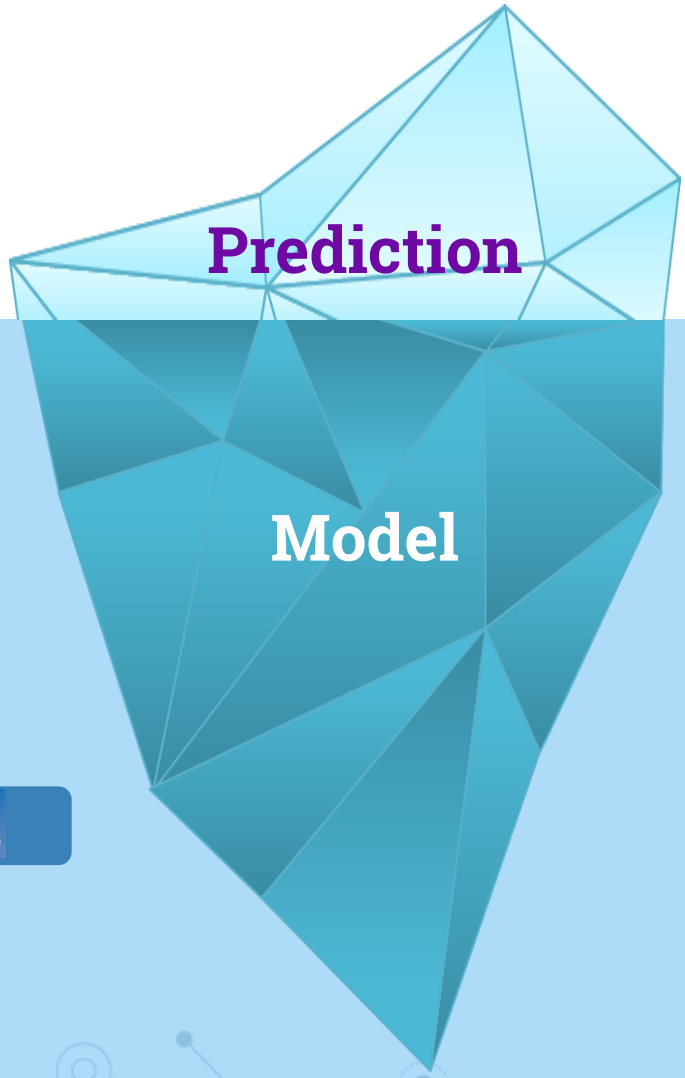
We used a Medicare dataset created through multiple layers of nested sampling processes

Blue: sampling decisions **outside of our control**

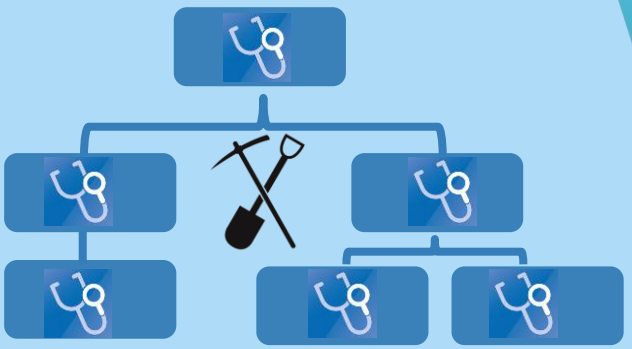
Gray: sampling decisions we **could control**



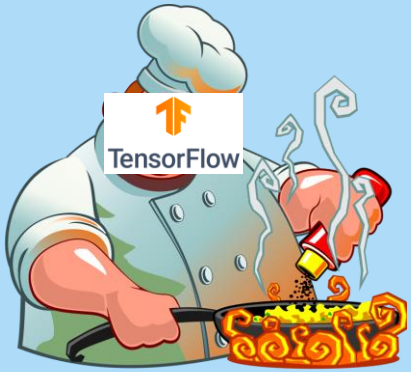
We compared two methods of selecting and engineering model features in an exercise of man versus machine



Manual Features



Automatic Features



Researched treatment algorithms to engineer features, converting sets of HCPCS codes to treatment regimens

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

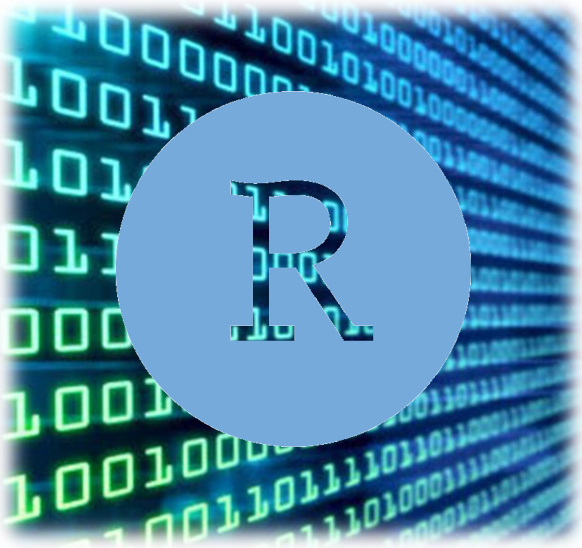
ACUTE

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**



Random forest was the best performing model with 72% accuracy

	Null Model	L1 Logistic Regression	Random Forest	Deep Learning
Accuracy	68%	69%	72%	57%
Feature Construction	None	Manual	Manual	Automatic
Method for Interpretability	N/A	Effect-size model coefficient intervals	Variable importance plot	Model run for illustrative purposes
Challenges	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

Despite reasonable accuracy, model performs poorly in certain regions of confusion matrix, i.e. 0.29 true positive rate

Actual \ Predicted	Remission	Relapse
Remission	108	12
Remission	37	15

	Metric	Synonyms
Accuracy	**0.72**	1 – error rate
False Positive Rate (FP/N)	**0.10**	Type I error, 1- specificity
True Positive Rate (TP/P)	**0.29**	1-Type II error, power, sensitivity, recall
Positive Predictive Value (TP/P*)	**0.56**	Precision, 1 – false discovery rate
Negative Predictive Value (TP/N*)	**0.74**	N/A

Manually constructed features performed better...

But what else should we consider?

Our analysis suffers from limitations out of our control but there's much we can do to improve our initial results

Limitations

Small sample size

No reliable effect size estimates

Outcome periods different for patients

Layers of nested sampling resulting in non-representative sample

Data not rich enough longitudinally and lacks context from sources like EHR

Treatment efficacy on (sub)population basis is distinct from personal level

Effect size inference versus causal inference

Future Work

Iterative EDA and model fitting/tuning to improve predictive accuracy

Synthesize more detailed and robust treatment algorithm

Develop platform-agnostic package to process patient HCPCS codes and match basket of codes to treatment regimens

Deeper dive including acquiring additional data sources on a patient level

Q&A