Treelet Transform of ICD-9-CM Diagnosis Codes

Graduate School of Public Health, Department of Biostatistics

Masters Thesis Defense

Dominic DiSanto

Defended on DATE

Introduction & Background

Objectives

- **Primary Objective**: Transform a large number of ICD-9-CM diagnosis codes into a sparse set of features, using treelet dimension reduction, and apply this new feature space towards the prediction of clinical outcomes of in-hospital mortality, unplanned hospital re-admission, and hospital length of stay.

- **Public Health Significance**: The presented work leverages a large, publicly accessible database of critical care admissions and generate useful predictive models of clinical outcomes using only patient demographic and comorbidity diagnosis information.

Modern Health Data

- Digitization of clinical data (such as in an electronic healthcare record) has led to large volumes of patient-level data

- These data sets commonly contain large patient populations *and* robust data elements for each respective patient

- Large, publicly available data sets are a growing resource of clinical data

Clinical Prediction Models

- Present useful, and ideally generalizable, methods to measure patient risk of adverse, clinical outcomes
- Current prediction models of mortality, length of stay, and unplanned re-admission have limited performance and utility
- Ideal models demonstrate high prediction accuracy with few and easily collected data elements

Dimension Reduction

- Models that allow a number of data elements[1] to be represented by a smaller number of inputs

- Methods often use the correlation structure to represent “similar” covariates in a reduced number of inputs

- Commonly discussed in the context of high-dimensional biological data (e.g. genomic, metabilomic)

[1]: Also commonly referred to as inputs, covariates, features, etc.

Treelet

- A novel dimension reduction method proposed by Ann Lee, Boaz Nadler, and Larry Wasserman in 2008

- Previously improved performance of regression and classification models compared to “raw” input data

- Has yet to be applied in high-dimensional diagnosis data or in fitting of clinical prediction models

Data

MIMIC-III

- A publicily available[2] database of critical care admissions
- Propspetive cohort study of Beth Israel Deaconess Medical Center from 2001 to 2012
- Contains diagnosis, lab, and demographic information from 60,000 admissions in over 45,000 patients

[2]: MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635

ICD-9-CM Diagnosis Codes

- International Classification of Disease, 9th Version
- Coding system of disease and injury diagnosis used in hospital billing
- Over 17,000 unique codes describing various patient diagnoses

Outcomes

- In-hospital mortality

- Unplanned hospital re-admission
  - Captured within year of hospital discharge
  - *Analysis excluded patients who died post-discharge with no hospital re-admission*

- Total hospital length of stay
  - Measured in days

Covariates

- Primary focus on ICD-9-CM diagnosis codes (following treelet dimension reduction)
- Models controlled for patient demographic variables
    - Age
    - Sex
        - Genotypical sex of patient (Male, Female)
    - Insurance
        - Categorized as Medicare, Medicaid, Private Insurance, or Self-Pay

Analytic Cohort

- Final analysis of mortality and hospital length of stay included 38,554 patients
- Hospital readmission analysis included 28,894
    - *Excluding 9,660 patients who died within one-year of discharge without re-admission*
- Mortality and length of stay analytic cohort presented mortality rate of 14.49% (n=5,586)
- 2,153 (7.45%) of patients experienced unplanned re-admission
- Patients had median hospital length of stay of 7 days (and interquartile range of 4 to 12 days)
    - Values ranged from 1 to 295 days

Statistical Analyses

Treelet (1/2)

- Proposed by Lee, Nadler, and Wasserman in 2007 (“*Treelets – An Adaptive Multi-Scale Basis for Sparse Unordered Data*”)

- Inspired by existing dimension reduction methods of principal components analysis and hierarchical clustering

- Aims to represent an input set with reduced dimensionality *and* requiring only a subset of the input information provided

Treelet (2/2)

- For p input predictors, treelet constructs $p-1$ basis matrices (or $B_{L_1}, B_{L_2}, \ldots . B_{L_{p-1}}$)

- The final representation requires identifying a value for the the K parameter (for K retained inputs in the Lth basis matrix)
  - For a given K, there is an identifiable cut-off ($L^* | K$) and respective basis ($B_{L^*|K}$) using the normalized energy score proposed by Lee et al.

- Cross-validation can be used to identify the outcome-specific, optimal K^* (and resulting $B_{L^*|K}$)

Cross-Validation

- Involves random splitting of data into “training” and “test” sets
- Models are fit to “training” sets and performance assessed on “test” sets
- The presented analyses used 5-fold cross-validation to select K and $L|K$ parameters for treelet models
- Final model performance was assessed on a holdout test data set that was *not* used in cross-validation or model fitting | *20% of each outcome’s respective analytic cohort*

Logistic Regression

- Generalized linear model (GLM) that extends ordinary least squares linear regression to model *probabilities* of a binomially distributed outcome

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x_i}\boldsymbol{\beta}$$

- Used in modeling binary outcomes of in-hospital mortality and unplanned re-admission

Negative Binomial Regression

- Poisson regression is the most common GLM fit for count or rate data

- Negative binomial is an extension of Poisson regression, when the outcome of interest is *overdispersed*, using probability mass function:

$$P\left(y_i\right) = \frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\left(y_i!\right)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

for $\mu_i = exp(\mathbf{x}_i\boldsymbol{\beta})$

- Used in the presented work to model hospital length of stay

Model Fit

- Logistic regression classification accuracy was assessed by Brier's Score

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$$

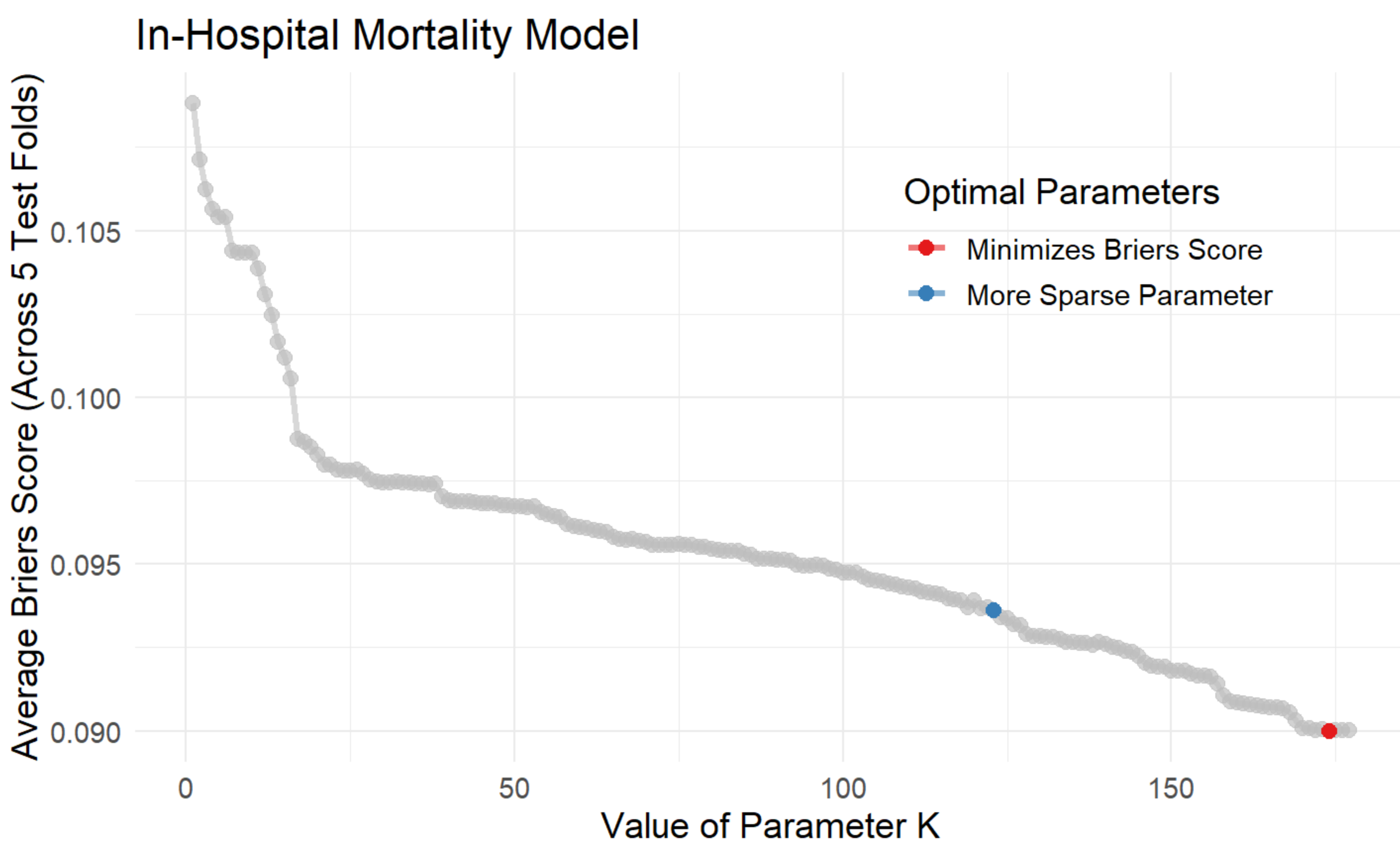
  - *Area under receiver operating characteristic curve is additionally presented for final logistic regression models*

- Negative binomial fit by root-mean-square error

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Results

Mortality (Cross-Validation)

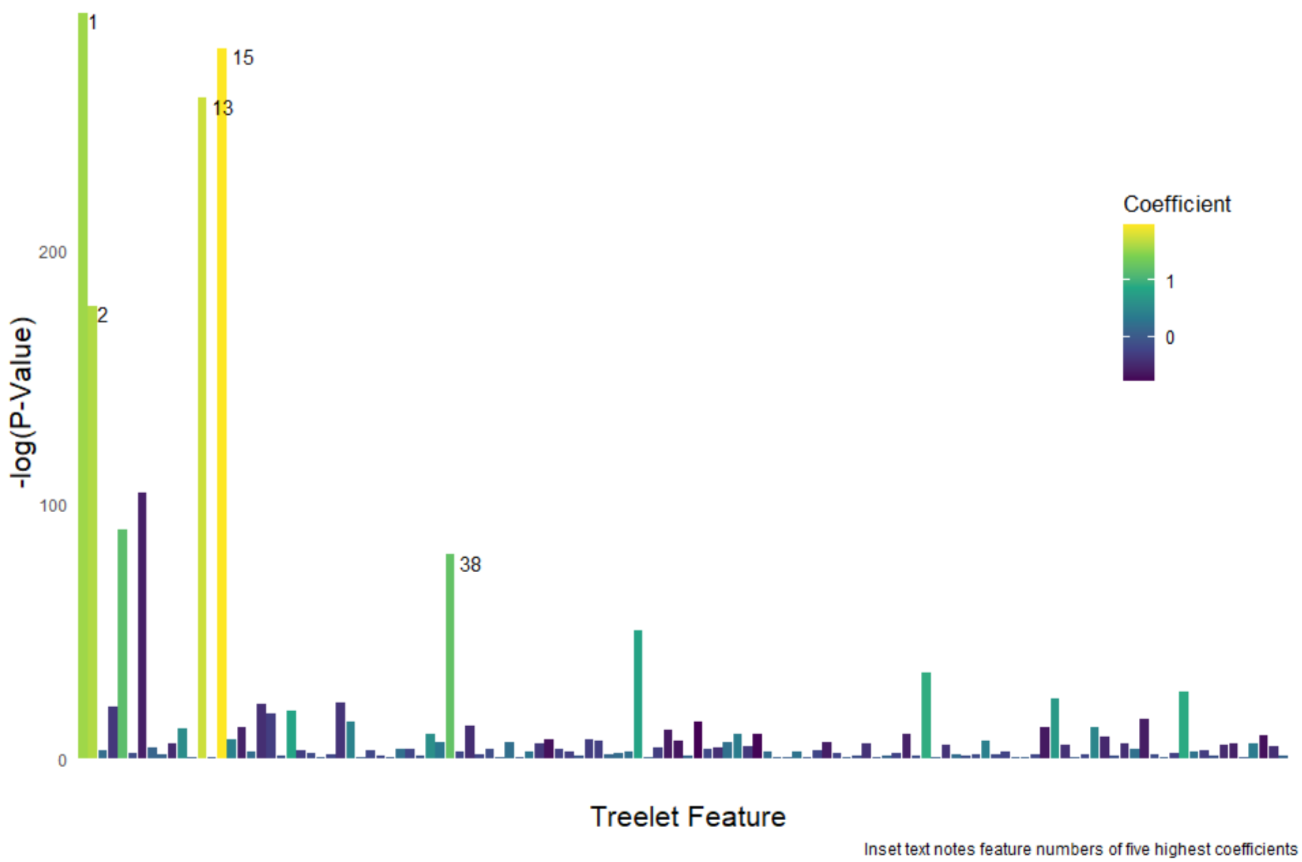


Mortality (Final Model Results)

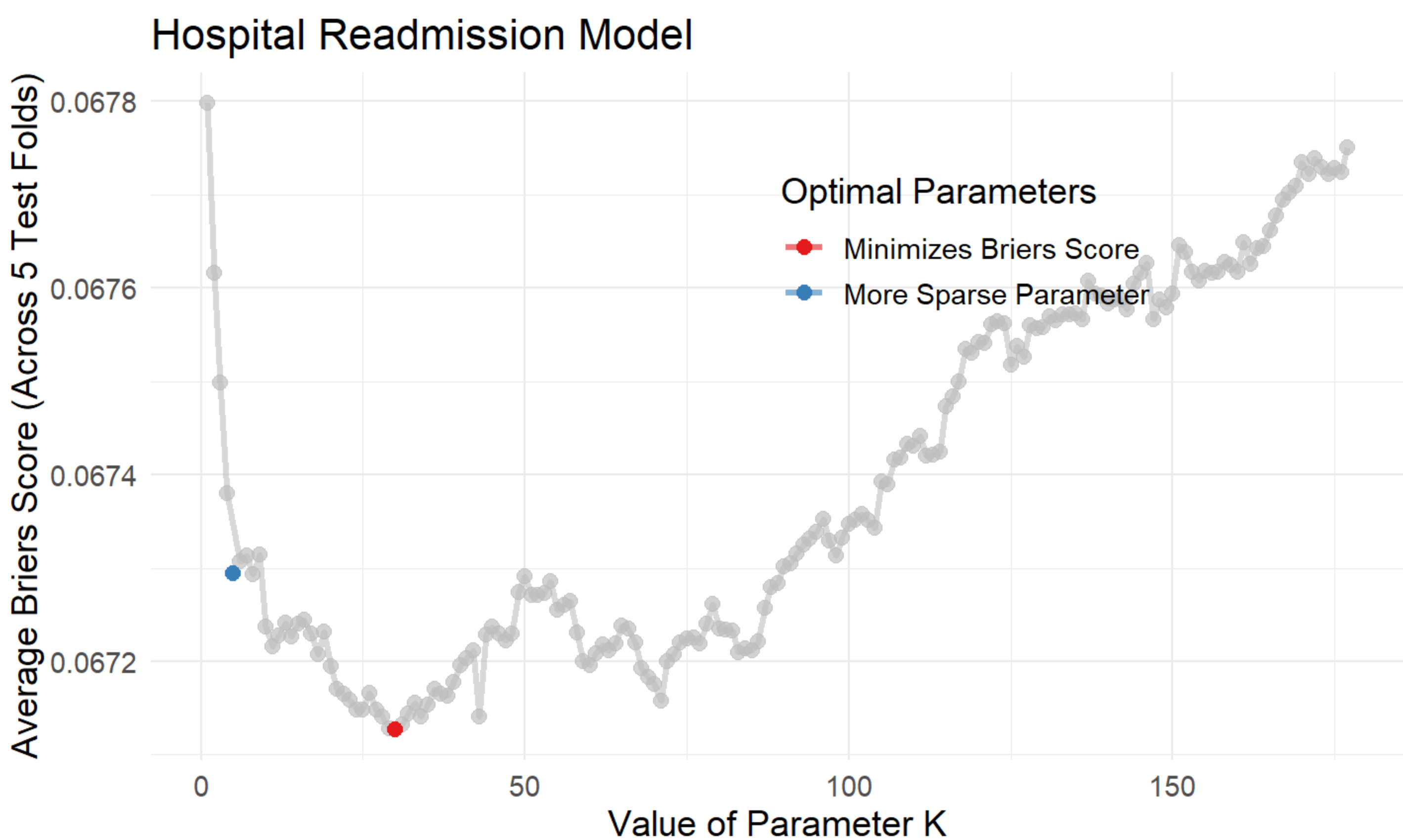
Predictor	β	95% Confidence Interval	P-Value
Intercept Term	-5.021	[-5.371, -4.671]	<0.001
Age	0.038	[0.035, 0.042]	<0.001
Sex (Male)	-0.118	[-0.198, -0.037]	0.004
Insurance			
Medicaid	0.178	[-0.140, 0.497]	0.273
Medicare	0.328	[0.029, 0.627]	0.032
Private Insurance	0.103	[-0.191, 0.397]	0.491
Self-Pay	1.174	[0.762, 1.586]	<0.001

Test Model Performance: Brier Score = 0.0917; AUC = 0.858

Mortality (Covariate Importance)



Readmission (Cross-Validation)

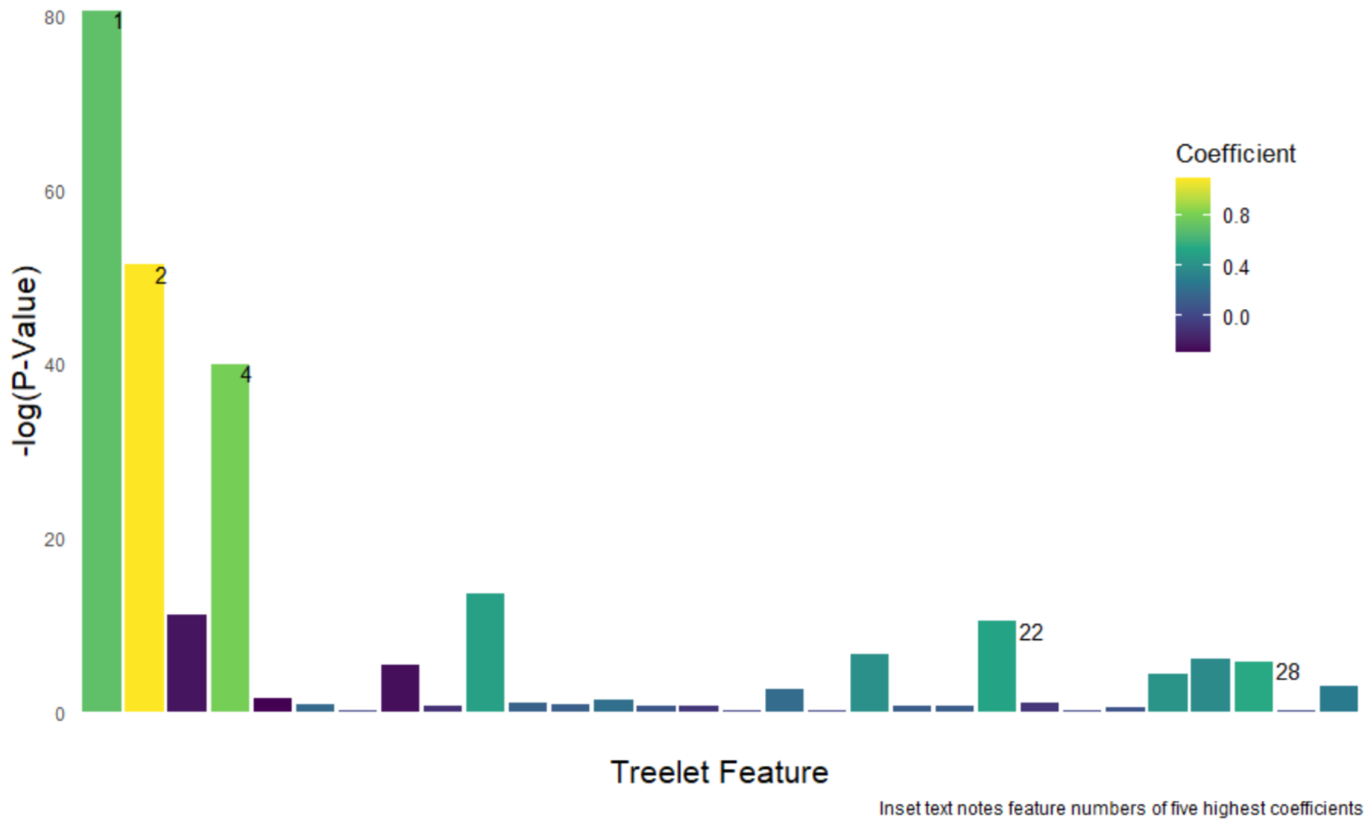


Readmission (Final Model Results)

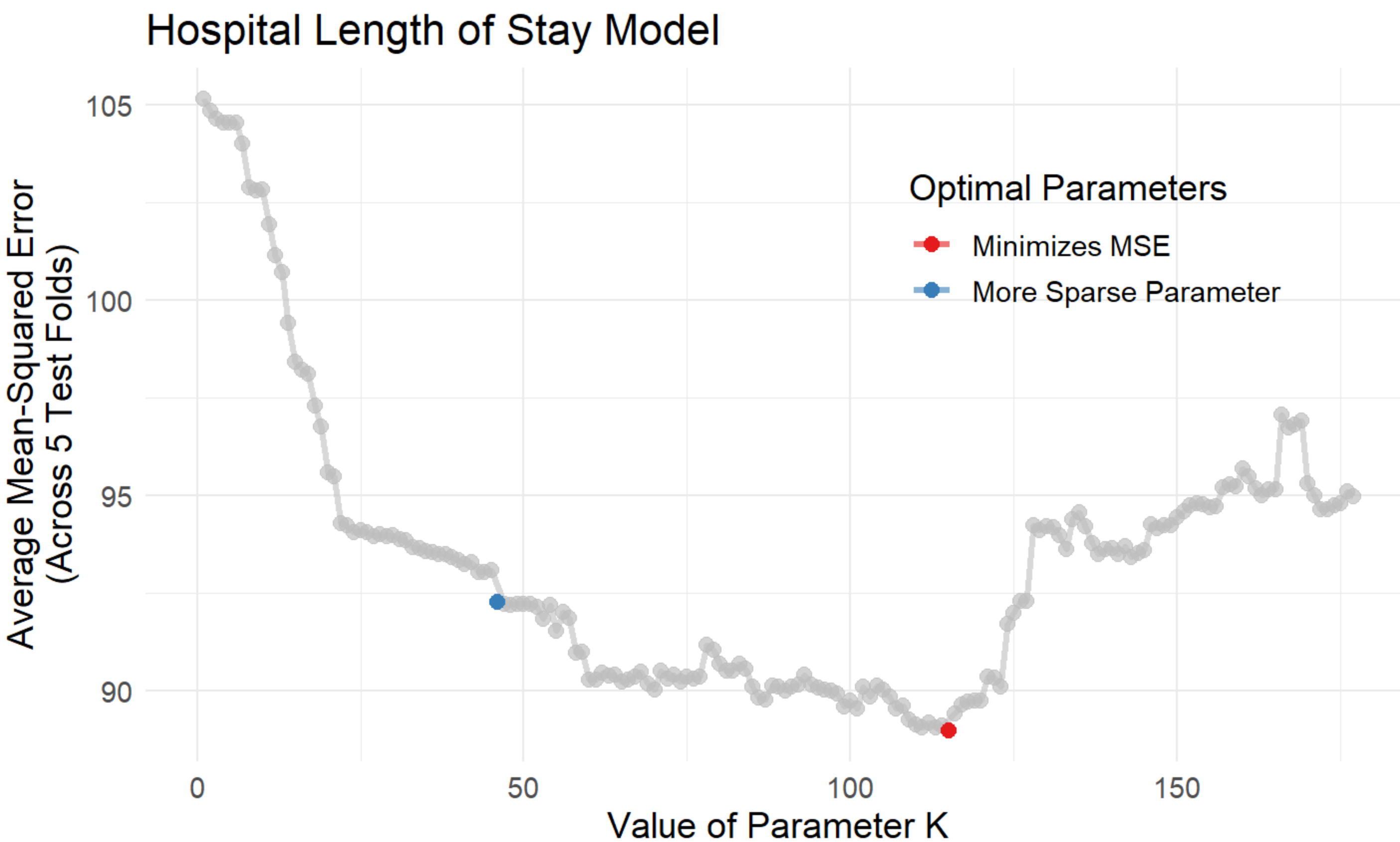
Predictor	β	95% Confidence Interval	P-Value
Intercept Term	-3.137	[-3.490, 2.783]	<0.001
Age	0.002	[-0.002, 0.007]	0.455
Sex (Male)	0.039	[-0.142, 0.064]	0.281
Insurance			
Medicaid	0.484	[0.162, 0.806]	0.003
Medicare	0.310	[0.005, 0.625]	0.053
Private Insurance	0.033	[-0.336, 0.271]	0.833
Self-Pay	-0.608	[-1.278, 0.061]	0.075

Test Model Performance: Brier Score = 0.0681; AUC = 0.661

Readmission (Covariate Importance)



Length of Stay (Cross-Validation)

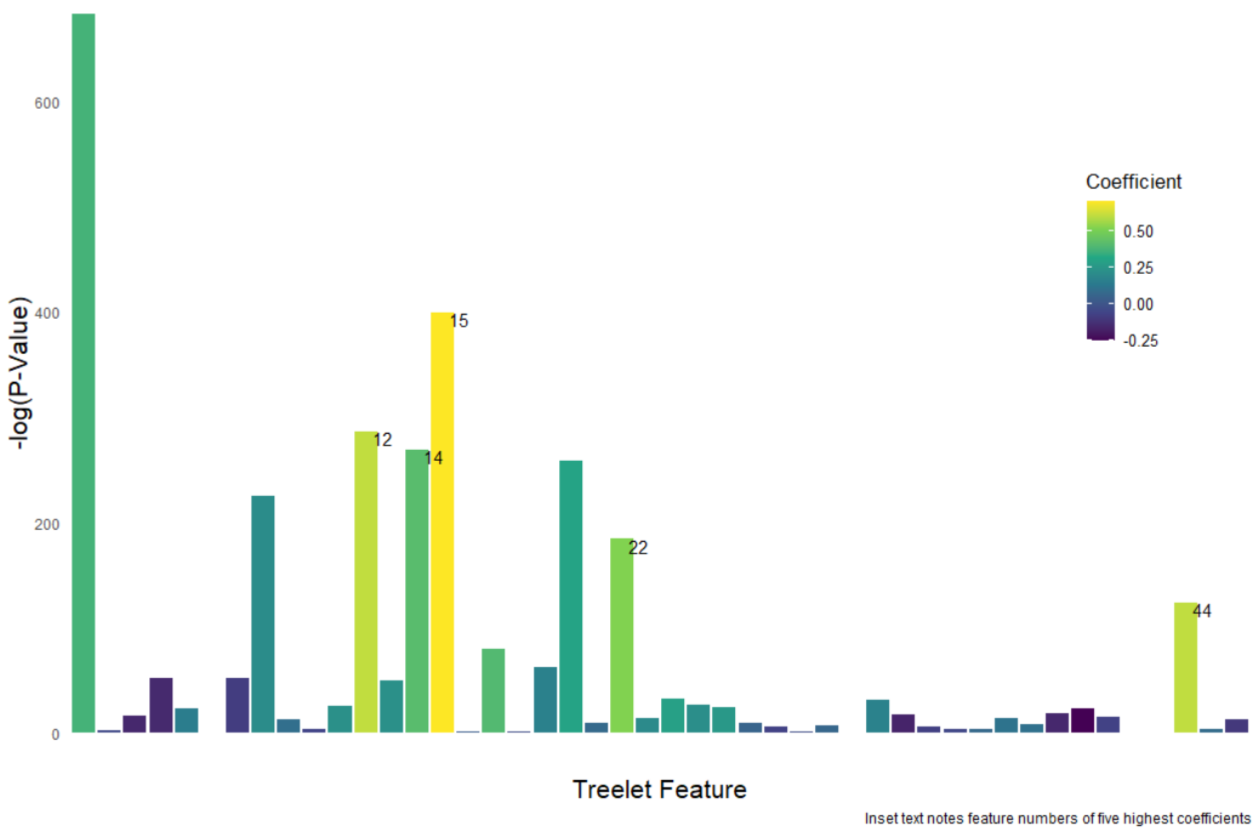


Length of Stay (Final Model Results)

Predictor	β	95% Confidence Interval	P-Value
Intercept Term	2.001	[1.942, 2.061]	<0.001
Age	-0.002	[-0.003, 0.002]	<0.001
Sex (Male)	0.053	[0.035, 0.071]	<0.001
Insurance			
Medicaid	0.114	[0.058, 0.171]	<0.001
Medicare	0.048	[-0.006, 0.101]	0.079
Private Insurance	0.039	[-0.12, 0.090]	0.133
Self-Pay	-0.318	[-0.407, -0.229]	<0.001

Test Model Performance: RMSE = 10.29

Length of Stay (Covariate Importance)



Implications & Conclusions

Model Summaries

- Final model or mortality demonstrates good predictive performance

- Models of re-admission and length of stay demonstrate limited prediction performance

- Treelet reduced dimensions for the number of inputs from our 178 ICD-9-CM diagnosis codes
  - Only the parameters identified for our model of re-admission yielded a sparse feature space

Comparison to Existing Models

- The presented model of mortality out-performs previously published models[3] of in-hospital mortality

- Our results corroborate previous publications, where diagnosis-data alone failed to adequately predict hospital re-admission and length of stay

[3]: Awad et al (2017).

Objectives (Revisited)

- **Primary Objective**: Transform a large number of ICD-9-CM diagnosis codes into a sparse set of features, using treelet dimension reduction, and apply this new feature space towards the prediction of clinical outcomes of in-hospital mortality, unplanned hospital re-admission, and hospital length of stay.

- **Public Health Significance**: The presented work leverages a large, publicly accessible database of critical care admissions and generate useful predictive models of clinical outcomes using only patient demographic and comorbidity diagnosis information.

Summary

- The presented work leverages a large, publicly available data set of critical care admissions and a novel dimension reduction method to build predictive models of hospital mortality, readmission, and length of stay
- When paired with patient age, sex, and payment method data, ICD-9-CM diagnosis codes demonstrate good predictive performance of in-hospital mortality, but remain limited in their ability to predict hospital length of stay and re-admission
- Additional information (e.g. patient discharge disposition, social determinants of health, patient environment data) may be necessary to adequately predict post-discharge outcomes

References

- Awad, A., Bader–El–Den, M., & McNicholas, J. (2017). Patient length of stay and mortality prediction: A survey. Health Services Management Research, 30(2), 105–120. https://doi.org/10.1177/0951484817696212
- Lee, A. B., Nadler, B., & Wasserman, L. (2008). Treelets—An adaptive multi-scale basis for sparse unordered data. The Annals of Applied Statistics, 2(2), 435–471. https://doi.org/10.1214/07-AOAS137
- Harrell, F. E. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis (Updated September 4, 2020). Springer Science & Business Media.
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition). Springer.
- MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635