Treelet Dimension Reduction of ICD-9-CM Diagnosis Codes

Dominic DiSanto, Masters Thesis

Graduate School of Public Health, Department of Biostatistics

Defended on December 7th, 2020

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 Healthcare Data

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

- Large, publicily available data sets are growing source of clinical reserach data, including both:
    - Diverse patient populations
    - Robust data elements for each respective patient

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

- Useful models not only demonstrate high prediction accuracy but ideally require *“feasible”* data
  - Inexpensive
  - Non-invasive
  - Standardized

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[2]

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

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

[2]: 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

Data

MIMIC-III

- A publicily available[3] database of critical care admissions

- Prospective cohort study of Beth Israel Deaconess Medical Center critical care admissions from 2001 to 2012

- Contains diagnosis, lab, and demographic information from 60,000 admissions in over 45,000 patients

[3]: 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
- The presented analysis included only ICD-9-CM codes with $\geq 1\%$ prevalence in our full, analytic cohort

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
  - Mortality rate of 14.49% (n=5,586)
  - Median length of stay was 7 days (range of 1-295 days)

- Hospital readmission analysis included 28,894
  - Excluding 9,660 patients who died within one-year of discharge without re-admission
  - 2,153 (7.45%) of patients experienced unplanned re-admission

Statistical Analyses

Overview

- Applied treelet dimension reduction to ICD-9-CM diagnosis codes

- Used cross-validation of GLMs to identify values of treelet parameters K-dimensionality and $L|K$-basis matrix
  - Logistic regression for in-hospital mortality, hospital-readmission
  - Negative binomial regression for hospital length of stay

- Final model fit measures were assessed on our hold-out test data-set (20% of each analytic cohort)

Treelet

- Using the covariance matrix of our input data, performs a series of rotations[4], grouping together features of high covariance
- For p input predictors, treelet constructs $p-1$ basis matrices (or $B_{L_1}, B_{L_2}, \ldots. B_{L_{p-1}}$) of dimensions $p \times k$
- 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.

[4]: Equivalent to fitting local PCA on the input features of highest covariance

Cross-Validation

- Models are fit to “training” sets and performance assessed on “test” sets
  - Logistic regression classification accuracy was assessed by Brier’s Score $\frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}_i - y_i\right)^2$
  - Negative binomial fit by root-mean-square error $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}$

- 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

Overview (revisited)

- Applied treelet dimension reduction to ICD-9-CM diagnosis codes

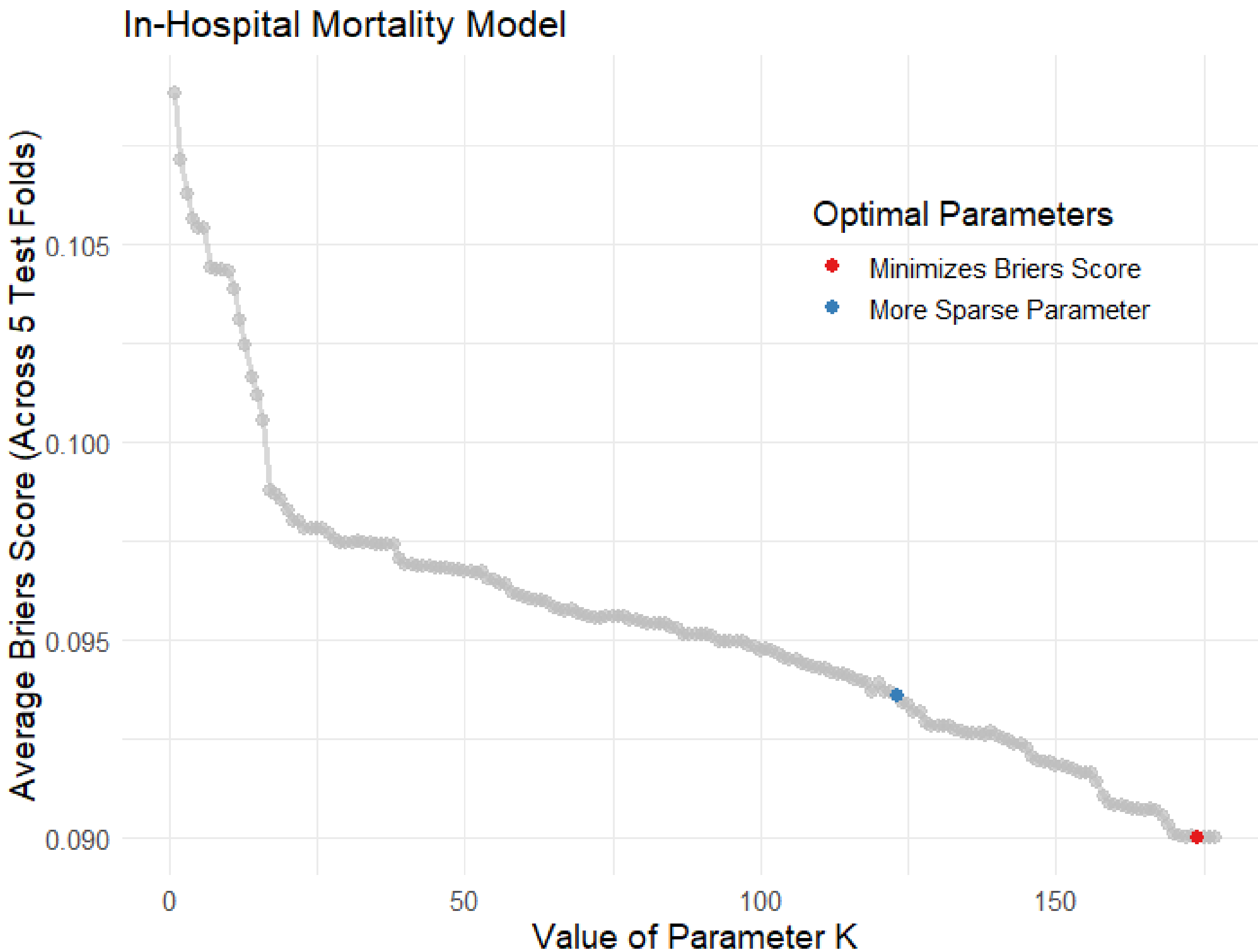
- Used cross-validation of GLMs to identify K-dimensionality and $L|K$ basis matrix parameters for each outcome
    - Logistic regression for in-hospital mortality, hospital-readmission
    - Negative binomial regression for hospital length of stay

- Final model fit measures were assessed on our hold-out test data-set (20% of each analytic cohort)

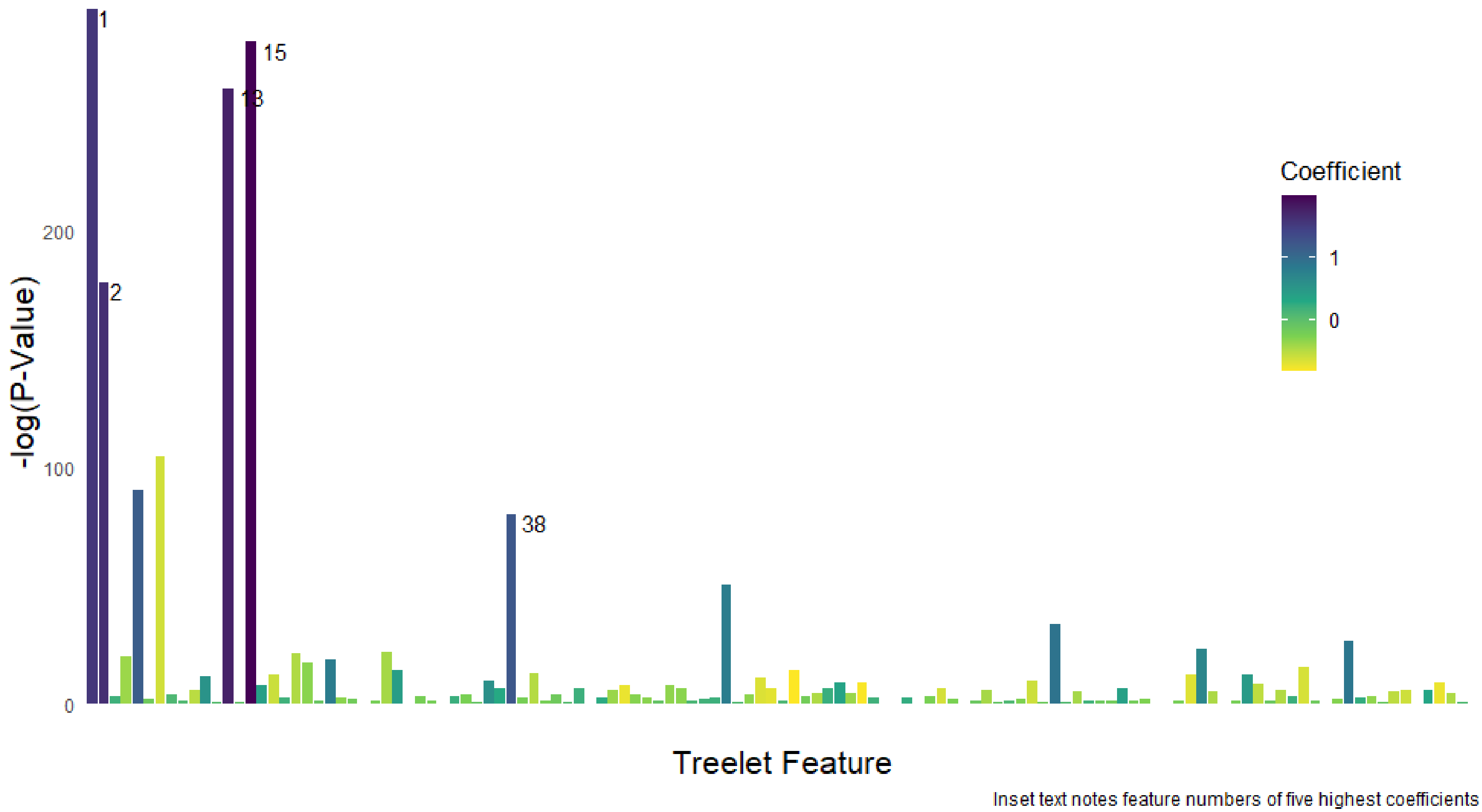
Results

In-Hospital Mortality

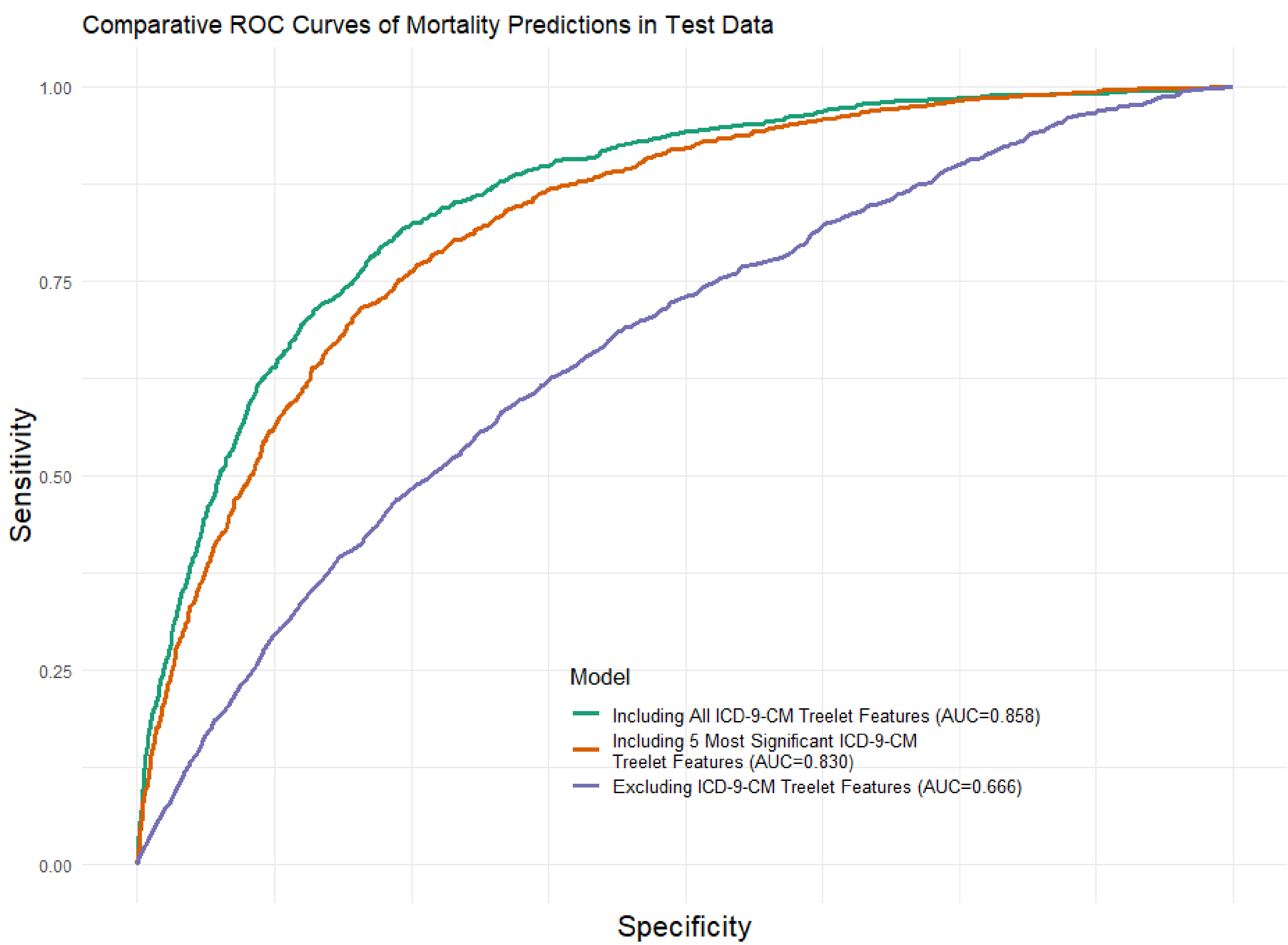
Mortality (Cross-Validation)



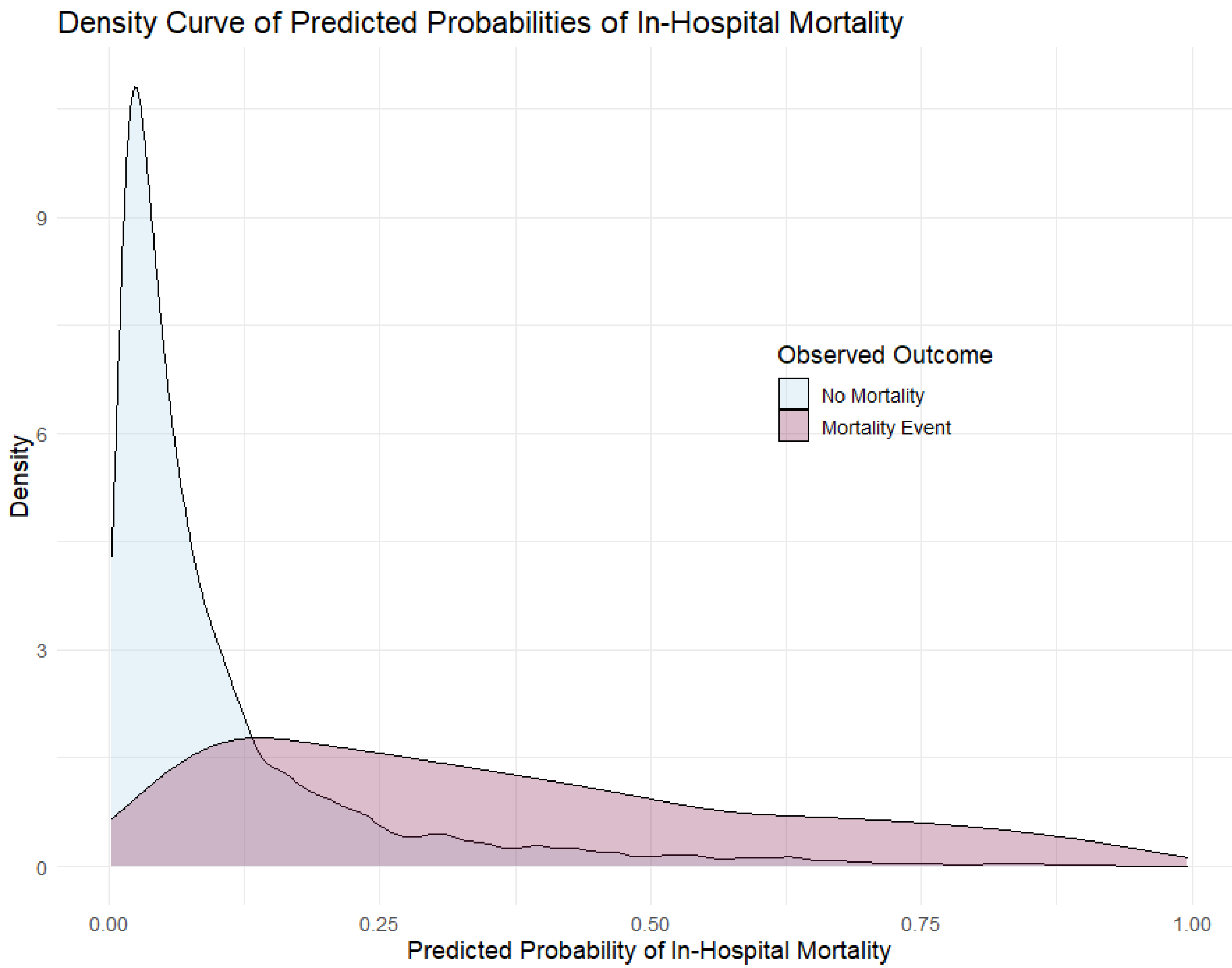
Mortality (Covariate Importance)



Mortality (ROC Curves)



Mortality (Predicted Probabilities)



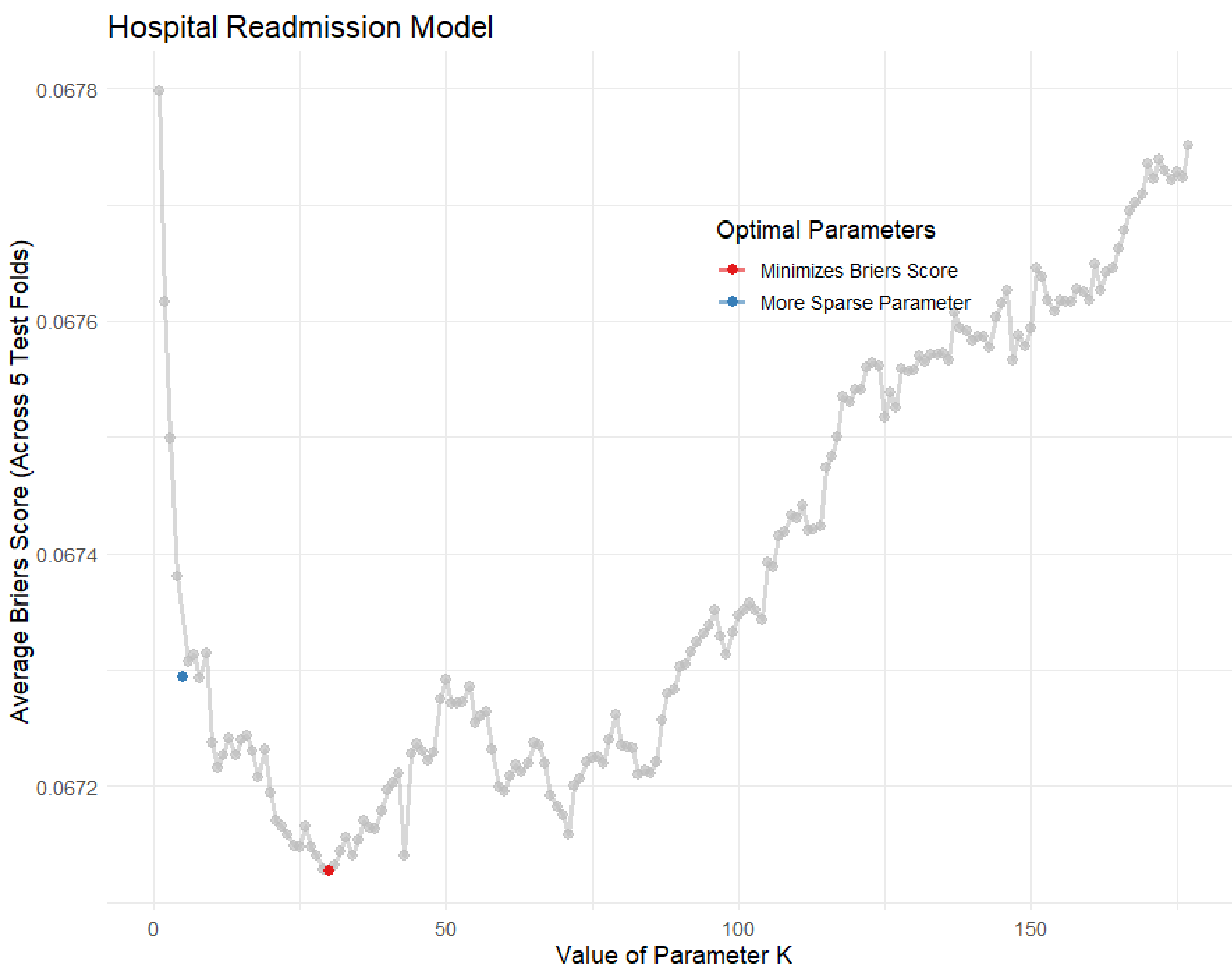
Mortality (Summary)

- Treelet reduction (and cross-validation) did *not* yield a sparse feature space
  - $K = 123$ dimensions retained loadings from all 178 diagnosis codes

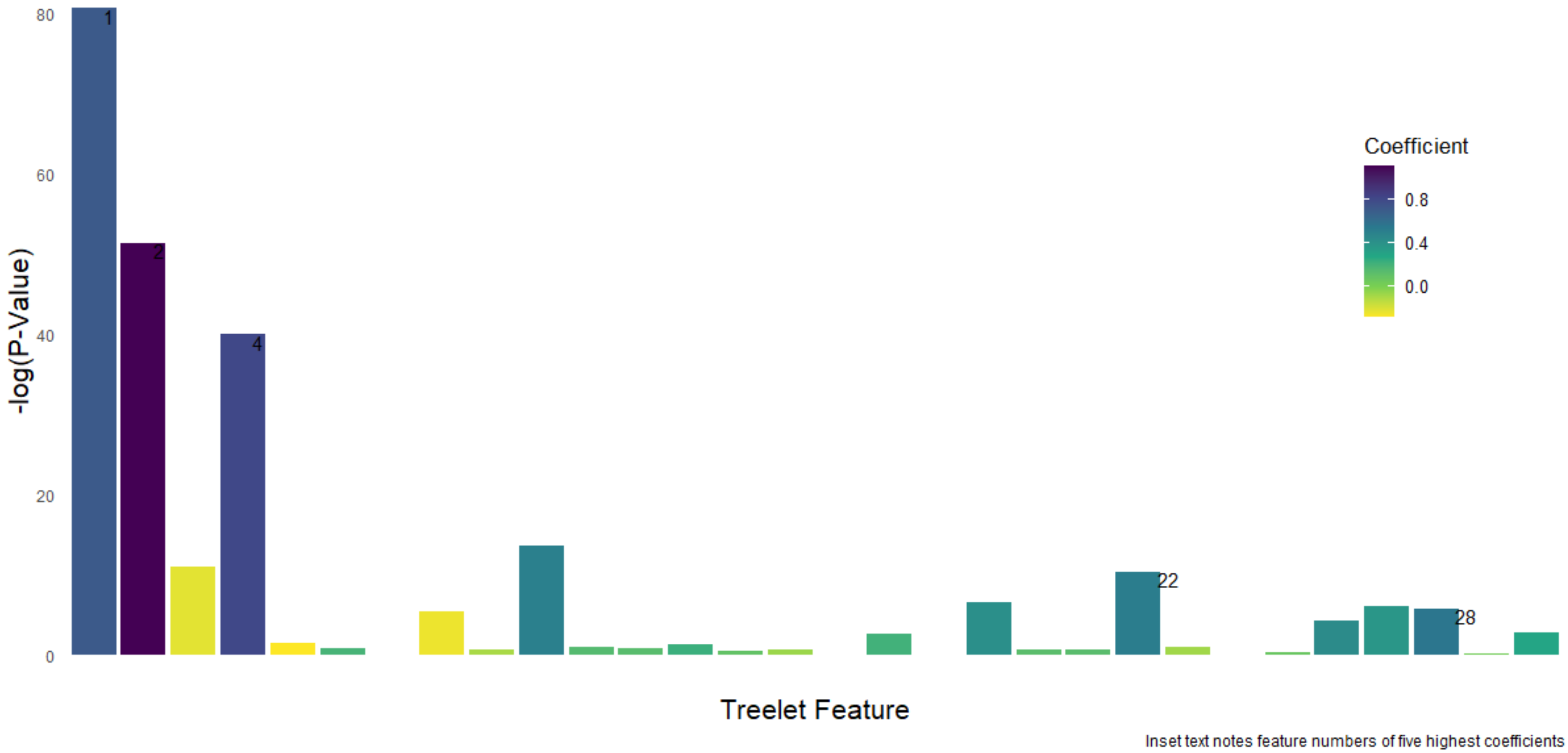
Model	Test AUC
Including All Treelet Features	0.858
Including 5 Most-Significant Treelet Features	0.830
Excluding Treelet Features	0.666

Unplanned Hospital Re-admission

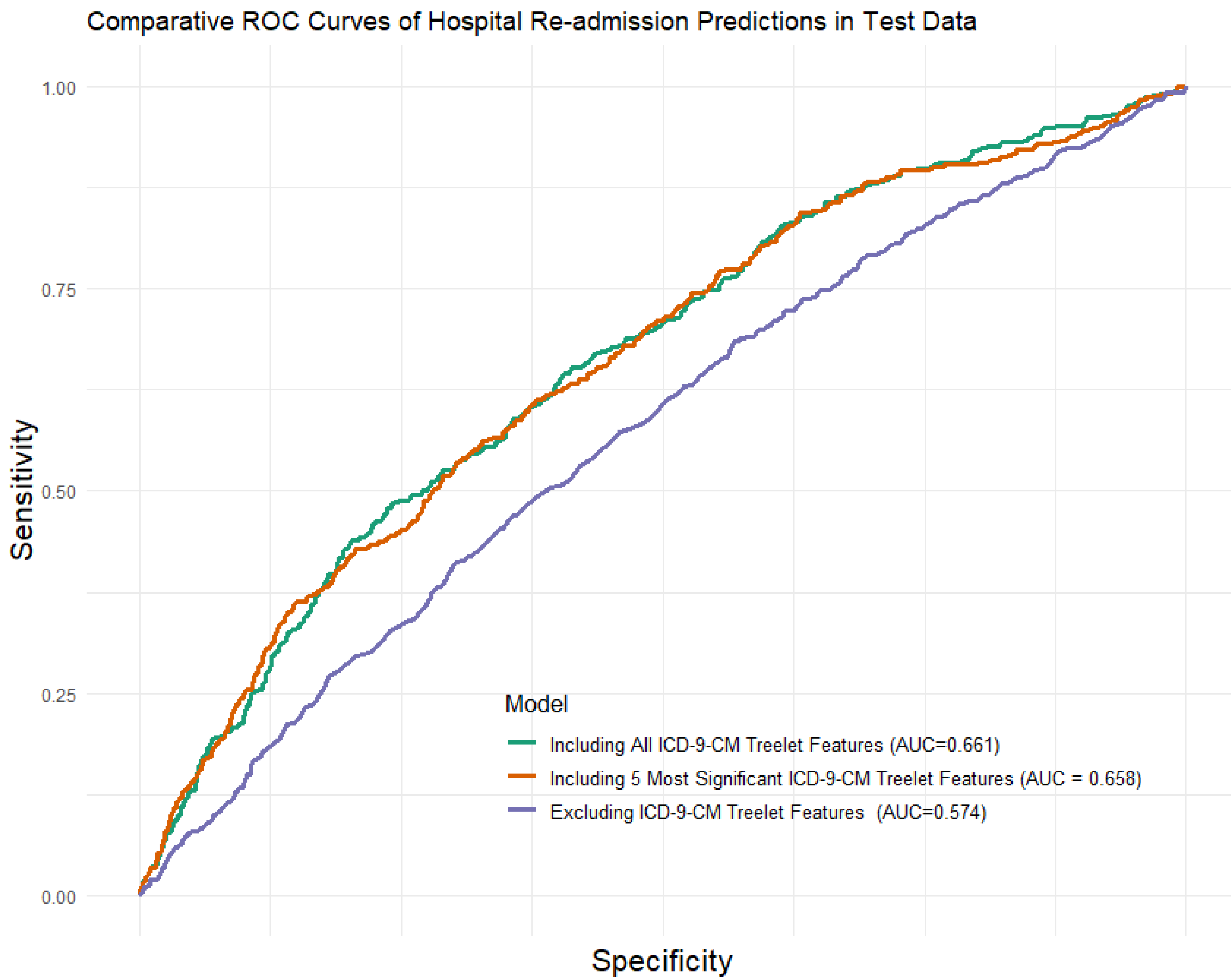
Readmission (Cross-Validation)



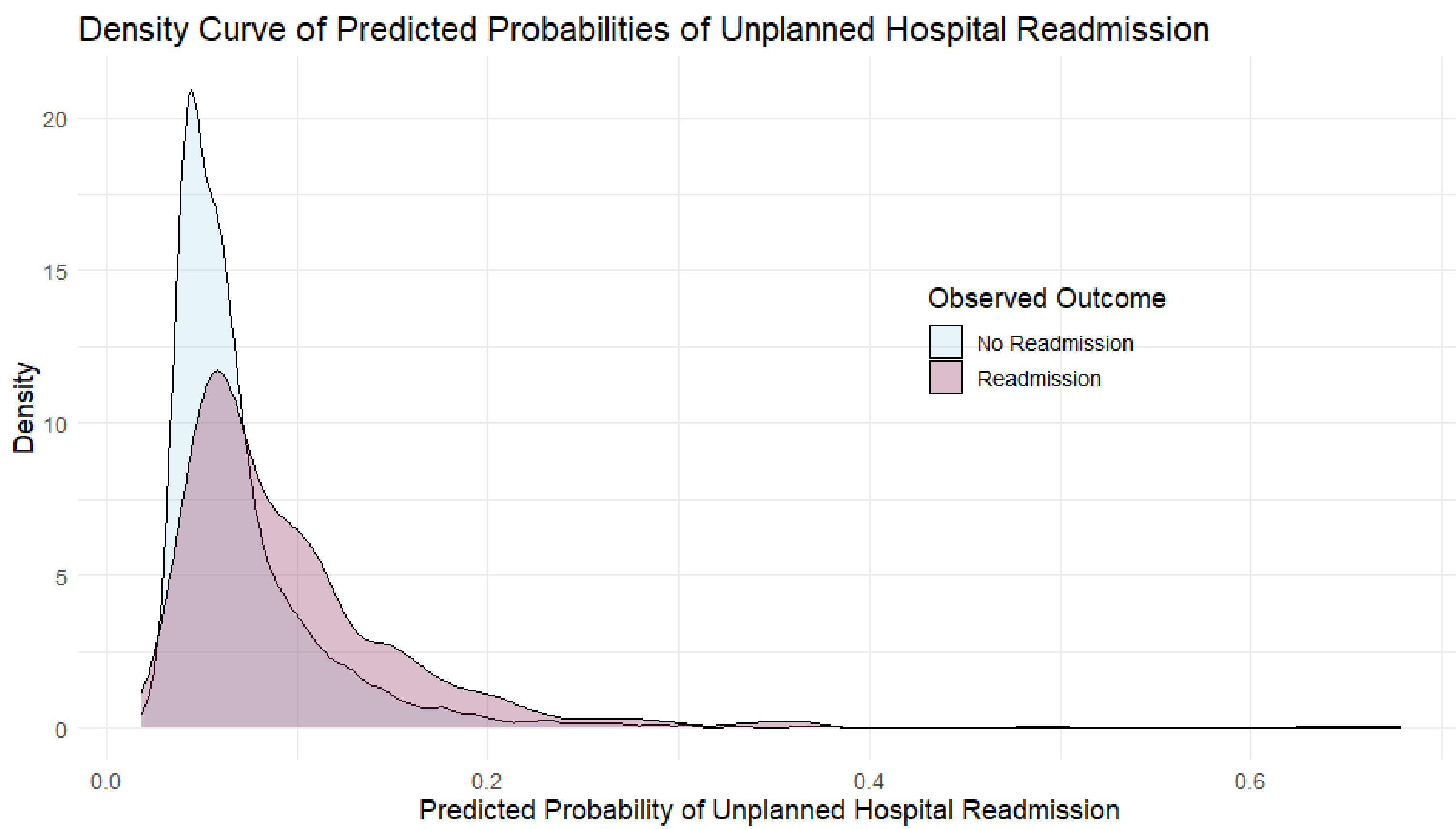
Readmission (Covariate Importance)



Readmission (ROC Curves)



Readmission (Predicted Probabilities)



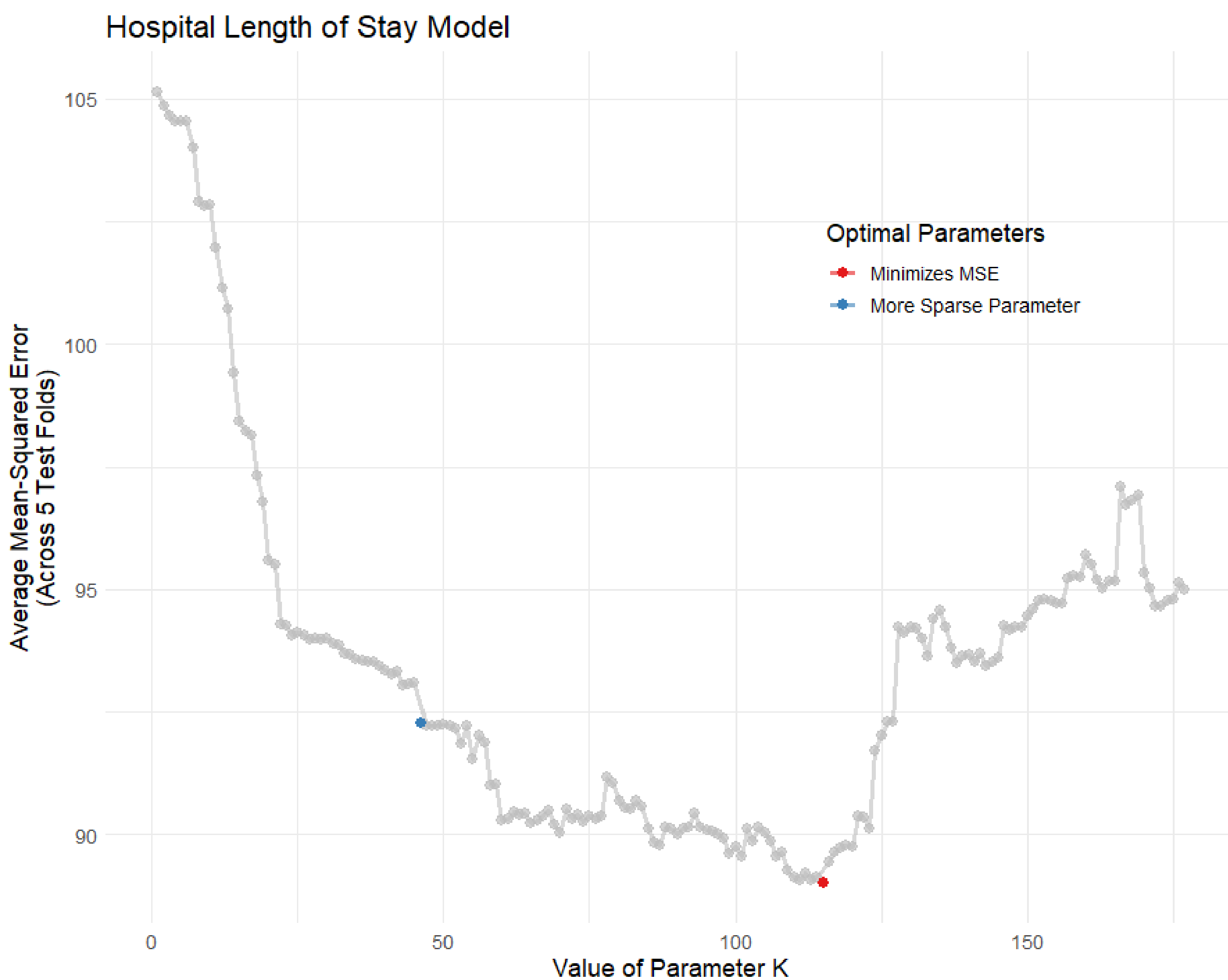
Readmission (Summary)

- Treelet improved performance but still presented only limited discrimination of hospital re-admission
  - While cross-validation reduced our 178 diagnosis codes covariates into $K = 30$ variables, we again retained loadings from all codes

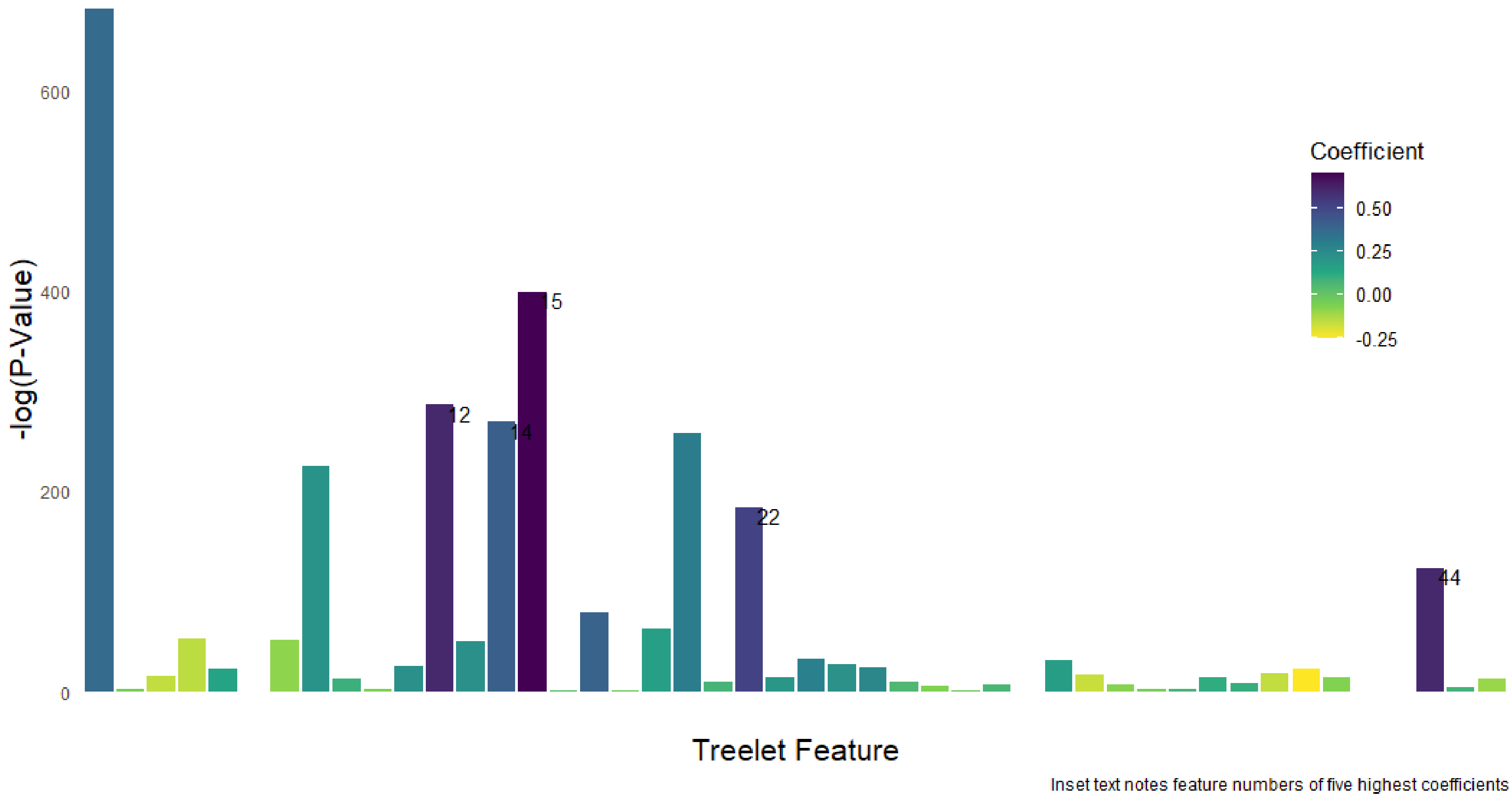
Model	Test AUC
Including All Treelet Features	0.661
Including 5 Most-Significant Treelet Features	0.658
Excluding Treelet Features	0.574

Hospital Length of Stay

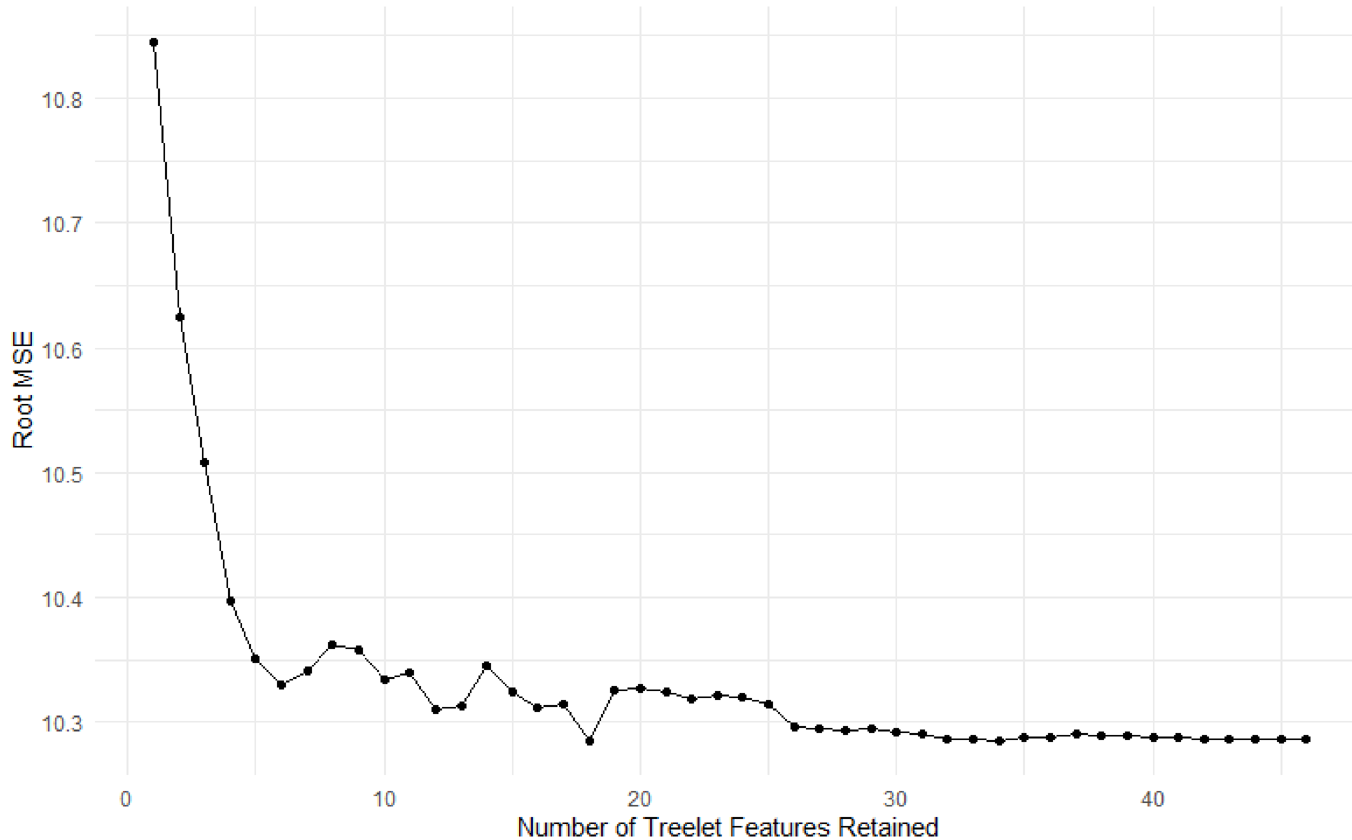
Length of Stay (Cross-Validation)



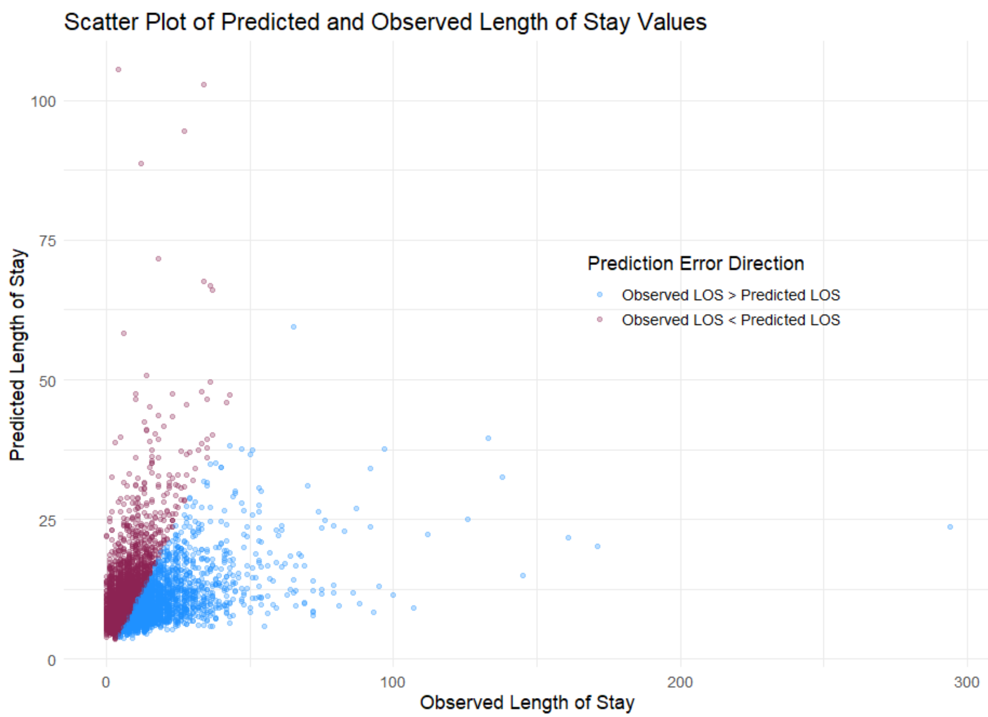
Length of Stay (Covariate Importance)



Readmission (Performance by Included Features)



Readmission (Predicted Probabilities)



Readmission (Summary)

- Treelet identified a reduced dimensionality and a sparse feature set
  - The retained $K = 46$ variables from our treelet model including loadings from 107 of our 178 ICD-9-CM diagnosis codes

Model	Test RMSE
Including All Treelet Features	10.29
Including 5 Most-Significant Treelet Features	10.35
Excluding Treelet Features	11.09

Comparison to LASSO and PCA

Comparative Model Results

Model	Mortality	Re-Admission	Length of Stay
Treelet	0.858	0.661	10.29
Lasso	0.868	0.669	9.94
PCA	0.860	0.666	10.13
Charlson	0.632	0.502	13.48
Elixhauser	0.615	0.513	13.49

Implications & Conclusions

Summary

- ICD-9-CM diagnosis codes improve 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

- Treelet dimension reduction reduces the number of retained covariates in our models but does not outperform PCA, LASSO

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.

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

Supplemental Slides

Retained Diagnoses Codes

Of 178 diagnosis codes included in analyses, each method retained the following number of unique codes:

Outcome	Treelet (Optimal)	Treelet (Sparse/1-St. Dev.)	LASSO
Mortality	178	178	170
Hospital Re-admission	178	178	48
Length of Stay	178	107	178

Mortality (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

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

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