[Final report] Computational prediction of clinical outcome of sepsis from critical care database (**Life Sciences** category)

Yosuke Tanigawa (SUNetID: ytanigaw, ytanigaw@stanford.edu)
Stephen Pfohl (SUNetID: spfohl, spfohl@stanford.edu)
December 16, 2016

Abstract

Sepsis affects over a million patients annually and remains one of the largest contributors to mortality in the ICU. However, prior models for the early detection of sepsis have typically relied on either manual chart review or a small number of hand-selected features. We propose a electronic phenotyping pipeline that utilizes the full numeric and categorical entries in the MIMIC-III database to extract relevant features for sepsis classification and mortality risk prediction. We have demonstrated that these derived feature are informative (1) to retrospectively identify sepsis cases with high performance (Lasso AUC = 0.796, Random Forest AUC = 0.816) and (2) to predict the risk of mortality following sepsis (concordance index = 0.81) without the use of ICD-9 billing codes. This high performance is promising for the future development of real-time risk models that utilize a large EHR-based feature set.

1 Introduction

The use of Electronic Health Records (EHR) over the past several years has generated a large data source that allows for the development of machine learning models for early diagnosis, risk stratification, and clinical decision support. Generating gold-standard labels for the outcome (phenotyping) is critical to the process of developing a training cohort[1], but is often a labor-intensive process requiring manual chart review. Sepsis affects over a million patients annually and remains one of the largest contributors to mortality in the ICU, costing the health care system over 14 million dollars per year. In hopes of facilitating high-throughput development of predictive models, we propose an electronic phenotyping pipeline capable of retrospectively identifying sepsis cases from the EHR that attains high classification performance without the use of ICD-9 billing codes. For this classification task, we derived features obtained from the EHR and we applied Lasso logistic regression and random forest to classify whether sepsis occurred at some point over the admission. Additionally we explored models to predict risk of mortality following sepsis on the basis of the derived EHR features. For this survival modeling task, we used the same inputs as in the classification task and applied Cox survival modeling to predict the probability of survival during the hospital admission. Although these models in their current state cannot be directly used for real-time risk prediction, the high retrospective classification performance is promising for future development of real-time risk models that utilize a similarly large EHR-based feature set.

2 Related Work

Early detection and diagnosis of septic shock is a notoriously difficult task that requires domain knowledge and failure to identify those cases accounts for much of US health care cost and mortality[2][3]. Prior work[4] has shown that data from MIMIC II, an earlier version of MIMIC III with data from 2001 - 2008, is sufficient for developing models that allow for an improvement in real time prediction of sepsis and sepsis-related mortality risk over older scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE II)[5], Simplified Acute Physiology Score (SAPS II)[5], Sequential Organ Failure Assessment (SOFA)[6] scores, Modified Early Warning Score (MEWS)[7][8], and Simple Clinical Score (SCS)[7], [5]. In general, these models attempt to predict the risk of sepsis progression on the basis of some previously hand-selected subset of features. Given that some of their features, however, include ICD-9 billing codes that are used to define clinical outcome, it is unclear to what extent sepsis may be identified if those ICD-9 features are moved. As a result, we see room for potential improvements in unbiased approaches that fully utilize the data in the EHR in that these approaches may identify features that were previously unknown to be predictive.

3 Dataset and Features

3.1 Modeling Overview

For this project, we considered two separate modeling paradigms. First, we handled the task of developing a classification model for retrospectively identifying hospital admissions that result in the development of sepsis at any point prior to discharge. Second, we developed time-to-event models to predict the time-varying risk of death following sepsis on the basis of the same set of clinical features used in the classification models. To accomplish this dual purpose, we processed the EHR to develop a set of summary features for each hospital admission and then use those derived features for classification and survival analysis.

3.2 Database Description

The data of interest is contained within the MIMIC III database[9], an electronic database curated by MIT that houses de-identified demographics, vital signs, lab test results, procedures, medications, notes, imaging reports, and outcomes of 58,000 hospital admissions between 2001 and 2012 for 38,645 adults and 7,875 neonates at the Beth Israel Deaconess Medical Center. The data is represented in a relational database of 26 tables that may be queried with SQL. In this project, we focused on the numeric valued data in 19 **event tables** (tbl. (1), under the dashed line) and the demographic information in the ADMISSIONS and PATIENTS. The ADMISSIONS and PATIENTS tables are additionally used for indexing the patients and admissions. We additionally use the ICD-9 coding data in the DIAGNOSES_ICD table to define the outcome labels on the basis of the criteria laid out by by Angus et. al [2].

3.3 Cohort and Outcome Definition

For the purposes of the classification task, we label hospital admissions as a positive example if sepsis occurs over the course of the admission and negative (normal) if no sepsis occurs over the course of the admission. For this purpose, we have selected the criteria laid out by Angus et. al [2] as the criteria used to construct the labels, as this criteria has been used in similar work to identify sepsis cases[4]. Briefly, cases are labeled as sepsis if the case contains ICD-9 billing codes for a bacterial or fungal process and either a diagnosis of acute organ dysfunction or if the patient required mechanical ventilation.

For survival modeling, we restrict the data to only those

Table 1: Number of features we dropped during feature engineering

Table name	(1) original		(2) wide		(3) sparsity	(4) NZV
	nrow	ncol	nrow	ncol	ncol	ncol
diagnoses_icd	651,047	5	58,976	6,985	N/A	33
admissions	58,976	19	58,976	78	N/A	N/A
labevents	27,854,055	9	58,147	2,880	522	461
inputevents_cv	17,527,935	22	21,879	1,112	166	119
inputevents_mv	3,618,991	31	21,879	1,112	166	119
outputevents	4,349,218	13	51,836	4,556	18	17
procedureevents_mv	258,066	25	21,894	464	52	34
chartevents_1	38,033,561	15	28,687	268	70	61
chartevents_2	13,116,197	15	34,904	36	12	10
chartevents_3	38,657,533	15	29,085	356	108	89
chartevents_4	9,374,587	15	27,210	44	32	28
chartevents_5	18,201,026	15	27,231	168	54	49
chartevents_6	28,014,688	15	34,896	1,644	278	267
chartevents_7	255,967	15	2,030	1,488	6	5
chartevents_8	34,322,082	15	7,990	1,268	184	155
chartevents_9	1,274,692	15	7,452	404	162	156
chartevents_10	9,584,888	15	18,650	528	28	17
chartevents_11	470,141	15	8,672	996	12	10
chartevents_12	265,413	15	1,405	804	4	4
chartevents_13	39,066,570	15	56,716	500	74	53
chartevents_14	100,075,138	15	24,549	3,032	836	535

patients that both have a recorded in-hospital time of death and are labeled as a positive case by the same criteria as before. We take the relative time of death to be the number of days between admission and death.

Since we used ICD-9 billing codes to define the clinical outcomes, we tested our models for three different data sets: (1) features from 19 events tables and ADMISSIONS table where this feature set does not contain ICD-9 codes. (2) Feature sets in (1) and the set of ICD-9 code features that are not used to define sepsis in Angus et. al [2]. (3) Feature sets in (1) and all the ICD-9 features.

3.4 Feature Engineering

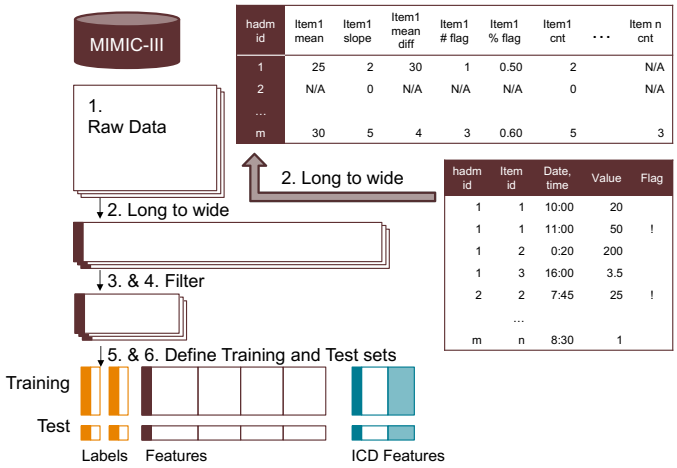


Figure 1: Schematic illustration of our feature engineering

We will now describe the feature engineering process by which we derive a set of time-invariant summary features at the admission level from an inherently temporal electronic health record. The following set of operations derives the training and test set, as illustrated in fig. (1).

1. **Starting point: Raw data (21 SQL tables)**
   Initially, each of the tables in tbl. (1) are processed separately. The number of rows and columns of each of the original tables are shown in column (1) of tbl. (1). Each *event table* is indexed by both the admission ID and an item ID that represents the type of measurement recorded in the value column.
2. **Conversion from long to wide format**
   For 19 events tables, we summarized the time course for numeric-valued measurements by taking the following four statistics for all the unique combinations of hospital admission and items (lab tests, medical procedures, etc.): mean, slope (time-derivative), mean of the difference of consecutive measurements[1], and the total number of measurements. Additionally, we derived features from the *flag* field in the LABEVENTS table, which indicates whether the results of the lab test is within the normal range. We took the number and the fraction[2] of abnormal flags for this table in addition to the four summary statistics. As not all of the patients have measurements for all the possible items, the resulting wide matrix is sparse and contains many missing values.
   For ADMISSIONS table, which contains demographic information of patients, we convert categorical variables such as diagnosis type, insurance, language, religion, marital status, and ethnicity to binary dummy variables and append these to the feature set.
   For DIAGNOSES_ICD table, which contains the ICD-9

[1]We employed the mean of the difference of consecutive measurements instead of the standard deviation to both reduce computational complexity and to ensure numerical stability.

[2]It is possible to calculate the fraction of flags feature based on number of flags and number of measurements. We intentionally included this fractional feature despite potential collinearity because Lasso logistic regression is linear model.

billing codes recorded during the hospital admission, we simply count the number of occurrences of each ICD-9 code during the admission. These features are handled separately and are only conditionally included in the models.
At this stage in the process, the number of rows and columns in tables are as in column (2) in tbl. (1).

3. **Drop columns with 90 %+ missing values**
For 19 events tables, we counted the fraction of missing values for each feature, and dropped features with 90 % or more missing values. The threshold is determined by manual observation of histograms of missing value fraction (data not shown).
This operation is not run for DIAGNOSES_ICD and ADMISSIONS tables.
4. **Drop near zero variance features**
For each table, the near zero variance features are dropped with the caret package[10]. At this stage, the number of rows and columns of each table are given in column (3) of tbl. (1). The resulting features for DIAGNOSES_ICD table may be used in step 11.
5. **Join tables and define the labels**
We sequentially joined each of the processed event table and the demographic information table ADMISSIONS table on combinations of the admission ID and the measurement item ID. The outcome labels as described in section 3.3 are also attached. This operation is not run for DIAGNOSES_ICD table.
6. **Split training and test set**
We divided our cohort to training and test set with $0.9 : 0.1$ ratio[3] while maintaining the ratio of positive and negative classes (tbl. (2)) across the testing/training split.
7. **Drop near zero variance features**
The near-zero-variance filtering step is repeated for the aggregated data set.
8. **Log-transformation and normalization**
We applied a log transformation and normalized the variables so that they have zero-mean and unit-variance.
9. **Median imputation**
Missing values are imputed with the column median.
10. **Drop near zero variance features**
We again performed near zero variance feature selection for the imputed table[4].
11. **(optional) Join ICD-9 billing codes**
We conditionally joined the ICD-9 code features from DIAGNOSES_ICD to produce an expanded feature set containing ICD-9 features.

4 Methods

4.1 Brief Description of Models

4.1.1 Logistic Regression with L_1 Regularization

For the task of binary classification, we first applied logistic regression with L_1 regularization, often called the Lasso[12][13]. The model is fit by minimizing the following loss function:

$$L(\theta) = \sum_i (y^{(i)} - \theta^T x^{(i)})^2 + \lambda \|\theta\|_1 \qquad (1)$$

[3]We considered having 5,000 data points in test set is sufficient to evaluate the performance of models and used $0.9 : 0.1$ split based on guideline discussed in Ng 2016[11].

[4]We have tested Lasso logistic regression with and without this step, and found that the model performs well with this feature selection step (data not shown).

Table 2: # of data points in data sets

	positive	negative	total
training	13,729	39,400	53,129
test	1,525	4,372	5,897
total	15,254	43,772	59,026

The usage of the L_1 regularization over other competing penalty (such as the L_2 penality as in ridge regression) is preferred due to the tendency for the L_1 penalty to find a sparse solution [12][14] such that many elements of the solution are found to be zero.

4.1.2 Random Forest

We additionally considered the usage of the random forest model[15][16][17] for classification. Briefly, the random forest is an ensemble of decision trees where each decision tree is constructed on a random bootstrap sample of the data set where at each split point in the tree a random bootstrap sample of the features are considered as candidate features and the optimal feature and decision rule are determined such that the heterogeneity of the class distribution in the parent node is optimally decreased upon assigning samples to the child nodes. Predictions are made by evaluating the sample on each of the individual trees and then using the aggregate of the predictions across all trees in the ensemble. Each tree is thus a low bias/high variance model and the process of aggregating the predictions across all trees has the effect of lowering the variance of the model overall while maintaining relatively low bias. The fraction of trees that vote for a particular class may be interpreted similarly to the predicted probabilities returned from logistic regression.

4.1.3 Cox Proportional Hazards with L_1 Regularization

For the purposes of the survival modeling, we used the Cox Proportional Hazards model with L_1 Regularization[18][19][13][20]. We considered the model without censoring. Interested readers should refer to Tibshirani 1997[19] for more details.

In this model, it is assumed that the time-dependent hazard for each example may be represented as

$$\lambda^{(i)}(t|x^{(i)}) = \lambda_0(t) \exp(\theta^T x^{(i)}) \qquad (2)$$

where λ_0 is some unspecified baseline hazard. The solution may be found by maximizing the partial likelihood:

$$L(\theta) = \prod_{r \in D} \frac{\exp(\theta^T x^{(j)})}{\sum_{j \in R} \exp(\theta^T x^{(j)})} \qquad \text{where } \|\theta\|_1 \leq s \qquad (3)$$

where D is the set of indices at which deaths occur, R is the set of indices of the samples at risk at time t, and j is the index of the failure at time t.

4.2 Model Fitting and Validation

We implemented our pipeline in R language and used packages[21][22][10][23][24][13][25][26][17][27][28][20][29] to train our models.

The regularization parameter λ in the logistic regression and Cox models is tuned with 10-fold cross validation on the training set such that an appropriate performance metric optimized. For binary classification we maximize the cross-validation area under the receiver operating curve (AUC) and for Cox modeling we minimize the partial likelihood deviance[19][13]. Given that random forests are fairly insensitive to changes in the model parameters[15], we fix the parameters of the algorithm such that each forest contains

250 trees, the number of features considered at each split-point is fixed at $\sqrt{p}$ where p is the total number of features, and we put no restrictions on the depth of each tree.

For evaluation on the testing set, the value of the tuning parameter that was found to give optimal cross-validation performance is used to fit a model on the entire training set and the resulting model is used to make predictions on the testing set. Testing performance is evaluated with the area under the receiver operating curve for classification models and with the concordance index (c-index)[28][30][31] for survival models. The c-index may be interpreted as a measure of the quality of the rank order of the survival predictions and may be interpreted similarly to the AUC of classification.

5 Results and Discussions

5.1 Retrospective classification of sepsis

Table 3: AUC of Lasso and random forest

Model	ICD features included	Total # of features	Training (n = 53,079)	Test (n = 5,897)
Lasso	None	797	0.908	0.791
	Non Sepsis	824	0.910	0.817
	All	830	0.947	0.9
RF	None	797	0.998	0.816
	Non Sepsis	824	0.999	0.855
	All	830	1.000	0.921

For the Lasso logistic regression models, we performed 10-fold cross validation on the training set (fig. (2)-(a)) to determine the weight λ of regularization term on the objective function that maximizes the cross-validation AUC. As mentioned before, the L_1 regularization has the tendency to select a sparse solution and this is evident in the trajectories of the coefficients for various levels of the L_1 norm (fig. (2)-(b)). Given the optimal value of the tuning parameter, we predict the label for both the training set and testing set. The ROC curves for the Lasso logistic models built on the three feature sets are shown in (fig. (2)-(c)) and the corresponding AUC values are shown in tbl. (3).

We fit the random forest models with 250 trees, as previously described. The performance results are summarized in tbl. (3) and the ROC curves are shown in (fig. (3)-(c)). We can see that the random forest model performs better than the Lasso logistic regression models on the test set for each of the feature sets considered. However, this performance boost is perhaps not as large as one would expect given that the random forest is capable of modeling non-linear interactions of the features. We suspect that the variance of the random forest models may be further reduced by the usage of a larger number of trees and thus a large forest may generalize better to the test set.

The usage of the ICD-9 code features tends to improve the performance of the models and there is a step-wise increase in performance when the non-sepsis ICD-9 codes are added to event derived features and when the full set of ICD-9 code features are used (tbl. (3), fig. (2)(c) fig. (3)(c)). This is not surprising, as those ICD-9 code features were used in the definition of the outcome labels.

Analysis of the random forest variable importance measures (in terms of the mean decrease in the Gini Coefficient) in the no-ICD case (figs. (3(a)–3(b))) indicates that a relatively small number of the variables are important in the classification. Interestingly, we find that the most informative features tend to be the number of times that a specific lab test was measured rather than the actual recorded value of the tests. A possible interpretation is these features serve as a proxy for the perceived wellness of the patient, as clinicians will be more likely to order a lab test if they suspect that the patient is unwell. However, given that the number of measurements is highly informative, it is clear that data are not missing at random and thus the results may be improved by more advanced imputation methods.

5.2 Prediction of risk of mortality

Table 4: C-index of Cox proportional hazard model

Model	ICD features included	Total # of features	Training (n = 5,827)	Test (n = 330)
Cox	None	797	0.92	0.81

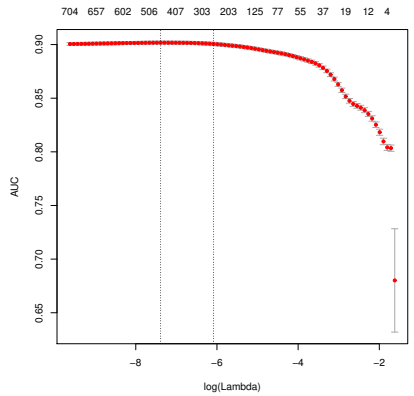
As in Lasso logistic regression models, we performed 10-fold cross validation on the training sets to determine the the value of the L_1 penalty term on the objective (fig. (4)(a)) for the Cox model. Using the λ that minimizes the partial likelihood deviance, we use the resulting Cox proportional hazards model for to predict the in-hospital survival functions for the test set on the basis of the feature set without ICD-9 billing code features (tbl. (4), fig. (4)(c)). With this feature set, we are able to achieve a c-index of 0.81 when predicting the survival functions of the admissions in the test set. This is promising, as it indicates that the same set of features that can be used for electronic phenotyping of sepsis may also be used to explain the variation in the risk of mortality at any particular time point. Also, it is likely that the usage of more powerful survival modeling methods such the random survival forest[32][33] would give considerable gains in model performance.

6 Conclusion & Future prospects

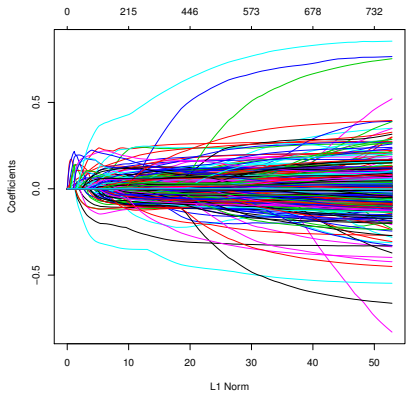
We developed a pipeline that extracts informative features to characterize sepsis patients from EHR in an unbiased manner. With these features, we classified sepsis patients with high AUC (0.816 for random forest), indicating that we could construct an automated high-throughput electronic phenotyping system with modest fidelity. We speculate that random forest performs better compared to Lasso logistic regression in the classification task because it can capture non-linear relationship among features. We also demonstrated that our feature set is informative to predict the risk of mortality during the hospital admission.

In this class project, we focused on numeric and categorical data in MIMIC-III database. In the future, we could expand our feature space so that that we handle images and clinical notes. Additionally, we could further optimize the parameters in our models, such as the number of trees in random forest model, to avoid overfitting to the training data set.

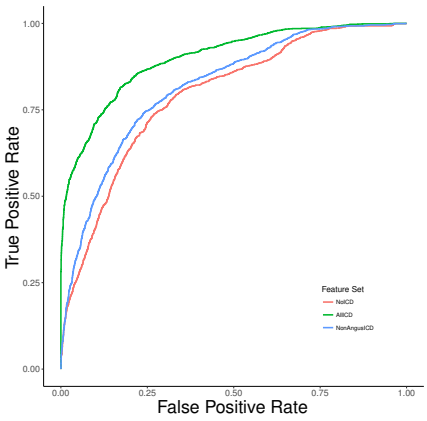
While our classification pipeline is useful for high-throughput phenotyping, one of the limitations of this work is that the classification models are only valid for retrospective classification. However, given our success on the classification problem and mortality prediction, it is likely that similar unbiased methods of deriving features from the EHR may perform well in the real-time risk prediction setting with some prior algorithm for determining the precise time of septic event.



(a) 10-fold Cross Validation. The left dotted line indicates the selected λ.

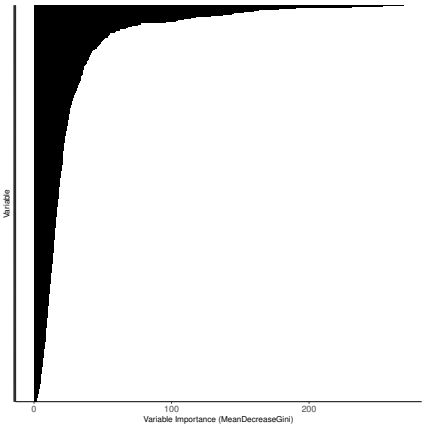


(b) Coefficient plot for no ICD-9 feature sets

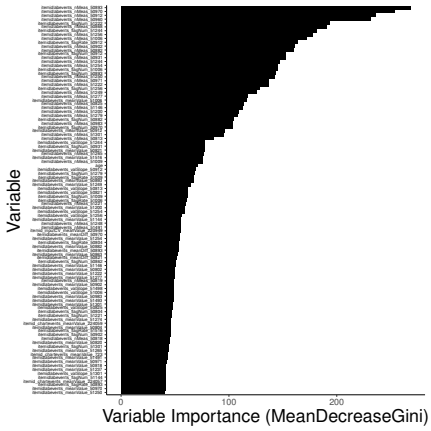


(c) ROC curves

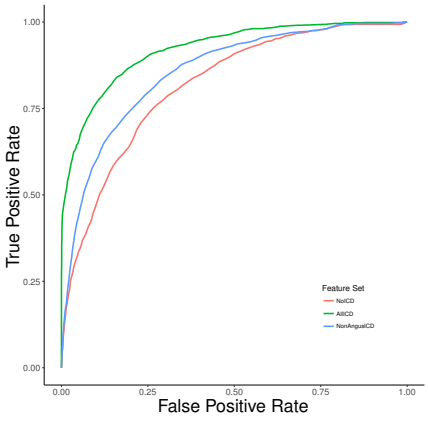
Figure 2: Results of Lasso logistic regression model for the classification problem



(a) Variable importance scores for no ICD-9 feature sets

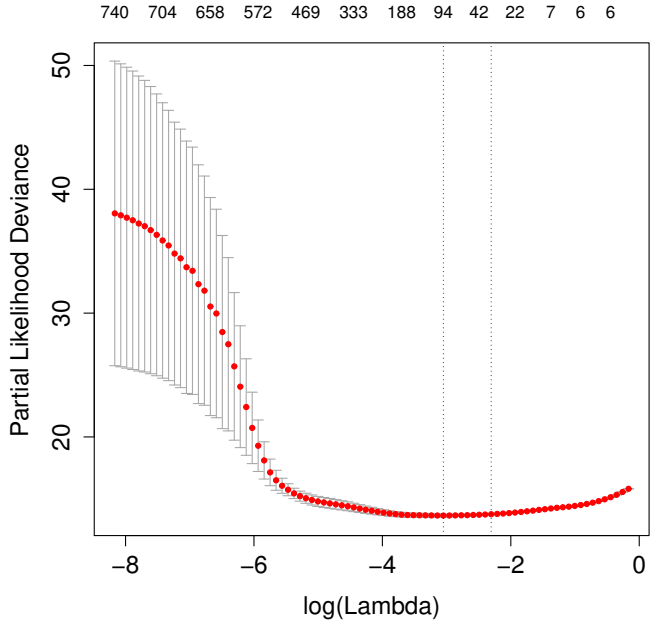


(b) Top 100 variable importance scores

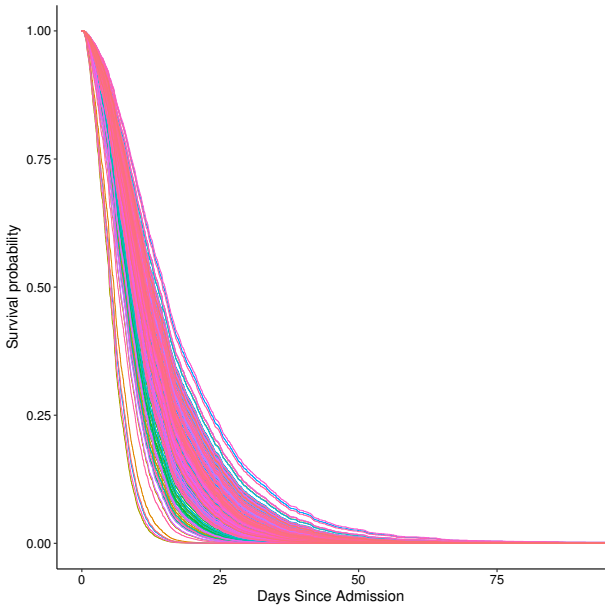


(c) ROC curves

Figure 3: Results of random forest model for the classification problem



(a) 10-fold Cross Validation. The left dotted line indicates the selected λ.



(b) Survival curve of patients in training set.

Figure 4: Results of binary classification with L_1 constrains for data without sepsis ICD-9 codes

References

[1] S. Yu, K. P. Liao, S. Y. Shaw, V. S. Gainer, S. E. Churchill, P. Szolovits, S. N. Murphy, I. S. Kohane, and T. Cai, "Toward high-throughput phenotyping: Unbiased automated feature extraction and selection from knowledge sources," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 993–1000, 2015.

[2] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, "Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care.," *Critical care medicine*, vol. 29, no. 7, pp. 1303–1310, 2001.

[3] G. Kumar, N. Kumar, A. Taneja, T. Kaleekal, S. Tarima, E. McGinley, E. Jimenez, A. Mohan, R. A. Khan, J. Whittle, E. Jacobs, and R. Nanchal, "Nationwide trends of severe sepsis in the 21st century (2000-2007)," *Chest*, vol. 140, no. 5, pp. 1223–1231, 2011.

[4] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (TREWScore) for septic shock.," *Science translational medicine*, vol. 7, no. 299, p. 299ra122, 2015.

[5] S. L. Barriere and S. F. Lowry, "An overview of mortality risk prediction in sepsis.," *Critical care medicine*, vol. 23, pp. 376–393, feb 1995.

[6] A. E. Jones, S. Trzeciak, and J. A. Kline, "The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation.," *Critical care medicine*, vol. 37, pp. 1649–1654, may 2009.

[7] N. O. Ghanem-Zoubi, M. Vardi, A. Laor, G. Weber, and H. Bitterman, "Assessment of disease-severity scoring systems for patients with sepsis in general internal medicine departments," *Critical Care*, vol. 15, no. 2, p. R95, 2011.

[8] C. Vorwerk, B. Loryman, T. J. Coats, J. A. Stephenson, L. D. Gray, G. Reddy, L. Florence, and N. Butler, "Prediction of mortality in adult emergency department patients with sepsis," *Emerg Med J*, vol. 26, no. 4, pp. 254–258, 2009.

[9] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database.," *Scientific data*, vol. 3, p. 160035, 2016.

[10] M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, and T. Hunt., *caret: Classification and Regression Training*, 2016.

[11] A. Ng, *Machine Learning Yearning (draft v0.5)*. 2016.

[12] R. Tibshirani, "Regression Selection and Shrinkage via the Lasso," 1996.

[13] J. Friedman, T. Hastie, N. Simon, and R. Tibshirani, *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, 2016.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, pp. 273–282, jun 2011.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[17] L. Breiman, A. Cutler, A. Liaw, and M. Wiener, *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*, 2015.

[18] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for Cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[19] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[20] T. M. Therneau, *survival: Survival Analysis*, 2016.

[21] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, *ROCR: Visualizing the Performance of Scoring Classifiers*, 2015.

[22] J. Conway, D. Eddelbuettel, T. Nishiyama, S. K. Prayaga, and N. Tiffin, *RPostgreSQL: R interface to the PostgreSQL database system*, 2016.

[23] H. Wickham and R. Francois, *dplyr: A Grammar of Data Manipulation*, 2016.

[24] H. Wickham and W. Chang, *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2016.

[25] G. Grolemund, V. Spinu, and H. Wickham, *lubridate: Make Dealing with Dates a Little Easier*, 2016.

[26] S. M. Bache and H. Wickham, *magrittr: A Forward-Pipe Operator for R*, 2014.

[27] H. Wickham, *stringr: Simple, Consistent Wrappers for Common String Operations*, 2016.

[28] B. Haibe-Kains, M. Schroeder, C. Olsen, C. Sotiriou, G. Bontempi, and J. Quackenbush, *survcomp: Performance Assessment and Comparison for Survival Analysis*, 2015.

[29] H. Wickham, *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2016.

[30] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar, "On Ranking in Survival Analysis: Bounds on the Concordance Index," in *Advances in Neural Information Processing Systems 20* (J. Platt, D. Koller, Y. Singer, and S. Roweis, eds.), pp. 1209–1216, Cambridge, MA: MIT Press, 2007.

[31] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar, "On ranking in survival analysis: Bounds on the concordance index," in *Advances in neural information processing systems*, pp. 1209–1216, 2008.

[32] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, pp. 841–860, sep 2008.

[33] H. Ishwaran, H. Ishwaran, U. B. Kogalur, and U. B. Kogalur, "Random Survival Forests for R," *October*, vol. 7, no. October, pp. 25–31, 2007.