Bates College
SCARAB

Honors Theses

Capstone Projects

5-2018

A Machine Learning Approach for Predicting Inpatient Discharge at Central Maine Medical Center

Mats Terwiesch
Bates College, mterwiesch@gmail.com

Follow this and additional works at: https://scarab.bates.edu/honorstheses

A Machine Learning Approach for Predicting Inpatient Discharge at Central Maine Medical Center

An Honors Thesis

Presented to

The Faculty of the Department of Economics

Bates College

In partial fulfillment of the requirements for the Degree of Bachelor of Arts

By

Mats Terwiesch

Lewiston, Maine

May 2nd 2018

Acknowledgements

This project would not have been possible without Dr. John Dickens of Central Maine Medical Center. I would like to thank Dr. Dickens for showing me both the complexities of health care delivery and sharing his approaches to these challenges.

I am grateful for the guidance of my thesis advisor, Professor Michael Murray. I often entered Professor Murray's office unsure of the progress I had made or of the road forward, but always left inspired by his wisdom and excited to keep learning.

I would like to thank Professors Henry Boateng and Nathan Tefft for their helpful feedback throughout the course of the project and following the thesis defense.

I owe a huge debt to the entire Bates Rowing program, which has defined my time at Bates.

Finally, I would like to thank my parents, friends and family around the world. No words can adequately capture how much they mean to me.

Introduction

Operating with a finite quantity of beds, medical resources, and physicians, hospitals are constantly allocating resources under conditions of scarcity. The American health care system is highly overburdened, facing more demand than it can handle. According to a series of reports by the Institute of Medicine, ambulances are turned away from American hospitals once every minute due to a lack of bed capacity. [1] The same reports find that the growth in demand for medical attention is dramatically outpacing the growth in supply, posing significant public health risks. Moreover, misallocation of resources and operational inefficiencies are a substantial driver of the United States' strikingly high healthcare costs.[2]

A specific instance of such inefficient resource allocation relates to the use of hospital beds. Patients get admitted to the hospital at random points in time for varying medical conditions. Accurately forecasting the duration with which a specific patient will stay in a hospital, also known as a patient's length of stay (LOS), can assist hospital decision makers in optimizing their workflow and allocating their resources efficiently. Short of having a good prediction of when a patient will leave the hospital and make room for new patients being admitted, the hospital will leave the patient's bed underutilized in the time immediately following discharge and delay the admission of a sick patient who needs the bed. Moreover, having a good estimate of future patient censuses in the various parts of the hospital could help facilitate staffing decisions. Finally, in case of bed shortages, methods of determining how much more hospital time a particular patient needs could help the hospital identify those patients best

---

[1] Berger, Eric. "Breaking Point: Report Calls for Congressional Rescue of Hospital Emergency Departments." *Annals of Emergency Medicine*

[2] Bentley, Tanya G.K., et al. "Waste in the US Healthcare System: A Conceptual Framework." *The Milbank Quarterly*

suited for discharge and to schedule nurse workflow accordingly. Thus, accurately forecasting the lengths of stay of patients has the potential to improve quality of care by moving patients into beds faster, guaranteeing sufficient staff and discharging the right patients while simultaneously decreasing costs by avoiding underutilized beds and idle staff.

Anecdotally, a significant bottleneck to patient discharge is the availability of transportation home from the hospital. For a large portion of the patients typically served by Central Maine Medical Center, it cannot be reliably assumed that patients will be able to arrange a ride home on short notice. One potential application of prediction modelling would be to help patients and social workers schedule transportation several days in advance.

How does one approach the task of predicting patient length of stay? The analysis of duration data goes by different names across disciplines. But, whether one is an engineer looking to calculate time until a component fails or a biologist predicting time until the death of an organism, the methods are fundamentally the same. I refer to the estimation of the time until an event of interest as survival analysis. In this context, the event of interest is the patient being discharged from the hospital.

Broadly speaking, approaches towards survival analysis can be classified as either classical econometric approaches or machine learning methods. In this thesis, I demonstrate the superiority of machine learning methods over the classical econometric approach. I go on to compare the performance of machine learning to an approximation of current forecasting practice at Central Maine Medical Center. My evaluation of the strengths and weaknesses of the machine learning model inform a brief discussion of potential real-world applications.

Methods

My study of patient length of stay uses a machine learning strategy called survival random forests. I use the survival random forest because of limitations in the classical econometric approach to survival analysis. I study a set of individuals admitted to the hospital from 27 February 2015 to 24 July 2017 from arrival to discharge. By relying on complete cohorts of admitted patients, I avoid selection biases. The risk of selection bias is elaborated on later in this section. I first describe classical survival modeling, its limitations, and the motivation to turn towards machine learning methods. I then describe the data available and the survival random forest model.

A commonly used classical econometric approach to survival analysis is the Cox proportional hazards model. In the Cox model, a baseline hazard rate is determined solely by the effect of time.[3] The hazard rate is the rate at which spells are completed at time t, given that they have already lasted at least until t. This baseline hazard rate is expressed as $\lambda_0(t)$. Heterogeneity in the sample is then accounted for by a vector of covariates x (e.g. dummy variable for admission on a weekend) and a vector of parameters expressed as β. The hazard function is expressed as:

$$\lambda(t) = \lambda_0(t) \exp(x'\beta)$$

Each individual's characteristics shifts the baseline hazard rate, while the effect of time elapsed remains constant. A restrictive feature that emerges due to the separation of time and covariates is that each individual's hazard function is expected to be parallel to the hazard

---

[3] Kennedy, Peter. "Limited Dependent Variables." *A Guide to Econometrics*, 4th ed., The MIT Press, 1998, pp. 259–261.

function of each of the other individuals. (Hence the name proportional hazards) Such proportionality is a fundamental assumption of the Cox model. As I demonstrate in Appendix A this assumption cannot be supported in my data.

An additional shortcoming of the econometric approach is that it is not clear what to make of the assumed effect of time. Does a patient who has already been in the hospital a week have an increasing probability or decreasing probability of discharge in the next day? In some cases, one might assume that the patient is steadily moving towards discharge, but in other cases one might take the fact that a patient has been in the hospital for a long time as an indicator that the stay will not be over soon.

Machine learning offers more flexible models that can handle the effects of interactions with time and between the covariates. I provide a background of the components of the machine learning model before turning to the final model itself.

Early machine learning estimation algorithms were proposed as alternatives to linear regression models. [4] These early machine learning estimation models were called "regression trees" when the predicted variable was continuous (e.g. length of stay) and "classification trees" when the predicted outcome was binary (e.g., readmitted or not). Later machine learning algorithms relied on multiples trees and were therefore named "random forest" models. [5]

The survival random forest (SRF) is the primary machine learning method used in this study.[6] The SRF is an extension of the random forest specifically designed for survival analysis.

[4] Breiman, Leo, et al. Classification and Regression Trees. 1984
[5] Breiman, Leo. "Random Forests." *Machine Learning*. 2001.
[6] Ishwaran, Hemant, et al. "Random Survival Forests." *The Annals of Applied Statistics*. 2008.

Understanding the survival random forest necessitates an understanding of survival trees, which in turn demands understanding of regression and classification trees.

In machine learning, trees can be used to visually express the relationship between predictors and an outcome variable. For example, the classification tree below considers whether or not the predicted length of stay for a patient is greater than 10 days.

Figure 1: Illustrative Classification Tree
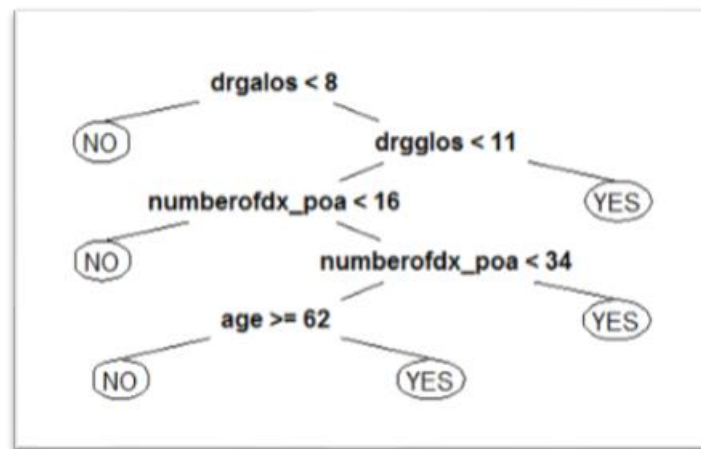
The algorithm begins by considering each predictor, and elects to make a "split" in the data that has the largest effect on the size of the total residual sum of squares (RSS). Each split generates two decision "nodes". At each node, the algorithm considers all of the predictors again and uses the predictor that makes the next largest split. This process is continued at each node until a stopping rule has been reached. [7] Once the tree is generated, it can be used to make predictions for new data. For example, in the classification tree above, if a new patient's diagnosis has a national average length of stay (*drgalos*) of less than 8 days, the algorithm's best

---

[7] James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. 2013.

prediction is that the patient will not have a length of stay greater than 10 days. However, if *drgalos* is greater than 8, the algorithm elects to look for more information by considering *drgglos,* the geometric mean of the national length of stay for the patient's diagnosis. If *drgglos* is less than 11 days, the algorithm considers the number of diagnoses present on admission, and so on. Generally, the performance of the tree will be measured by applying it to testing data that has previously been randomly separated from the sample.

When a tree is modified to handle survival data, it is called a survival tree.[8] The general structure of the tree is the same as in regression and classification trees, but instead of using residual sum of squares as the splitting criterion, the splits seek to maximize the difference in survival times at each split. The default approach towards measuring the difference in survival times is called the log-rank method. While I do not discuss it in detail here, the log-rank test statistic is a way to measure the magnitude of the difference between two survival curves. [9]

One downside of trees is that a particularly important variable may come to dominate the tree. Often, this results when a predictor makes a particularly large split at the top of the tree. This issue is known as "path-dependency", which can lead to a potentially significant portion of the total RSS remaining unreduced.[10] For example, a pair of variables used for the first two steps might in concert lead to a better prediction than a third does alone, but the third might dominate either of those two variables when the algorithm considers each sequentially. In response to the path-dependency problem, Breiman (2003) developed the random forest. The random forest is an aggregation of thousands of generated trees. The key element of the random forest is that, unlike

---

[8] Bou-Hamad, Imad. "A review of survival trees" *Statistics Surveys*. 2011.
[9] Statistics 331."Logrank Test". Class notes. Stanford University School of Medicine. https://web.stanford.edu/~lutian/coursepdf/unitweek3.pdf
[10] James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. 2013.

the tree generation process, only a random subset of the predictors are considered at each node. In this way, variables that lead to particularly large splits are only sometimes considered. Random forests and similar machine learning methods have been used in a variety of attempts to forecast patient length of stay. [11,12] Notably, Barnes (2015) uses a random forest to outperform physician predictions of patient length of stay.

To use a random forest approach in survival analysis, Ishwaran, Kogan, Blackstone and Lauer introduced the survival random forest in 2008. Ishwaran et al. outline the following algorithm.

1. Draw *B* bootstrap samples from the original data. The data not included (on average 37% of the data) in each bootstrap is called out-of-bag data.

2. A survival tree is grown for each bootstrap sample, with a random selection of variables considered at each node. At each node, the algorithm chooses the split that maximizes survival difference between nodes using the log-rank statistic.

3. Grow each tree to a default stopping rule.

4. Calculate the hazard function for each tree, and average the cumulative hazard function of all the trees

5. Calculate the prediction error rate of the combined cumulative hazard function using the out-of-bag data.

The hospital records available for analysis, normally used for insurance reimbursement and quality control purposes, are compiled upon patient discharge. So, patients are only

---

[11] Van Walraven, Carl, and Alan J Forster. "The TEND (Tomorrow's Expected Number of Discharges) Model Accurately Predicted the Number of Patients Who Were Discharged the Next Day." *Journal of Hospital Medicine*. 2017.

[12] Jones, Spencer S., et al. "Forecasting Daily Patient Volumes in the Emergency Department." *Academic Emergency Medicine*. 2008.

available for analysis once they are discharged from the hospital.[13] Left unaddressed, this creates a sample selection issue. Early in the dataset, only those patients who have a length of stay long enough to stay past the starting date of the study make it into the sample. Similarly, the end of the study excludes longer staying patients who do not experience discharge during the study period.

Therefore, near the beginning and end points of all the data available, some patients are observed while other patients who arrived on the same day as those observed remain unobserved. To rectify this issue, an analysis subsample was selected to ensure that only full cohorts (groups of patients arriving on the same day) were considered. Patients discharged within the first 150 days of the study (before 27 February 2015) were dropped from the sample, a decision that was informed by the longest patient length of stay being 146 days. Similarly, at the end of the sample, patients who arrived within the last 150 days of the sample (after 24 July 2017) were not considered.

The primary intended use of the findings of the study are to assist hospital administrators in inpatient management. Therefore, only inpatients were included in the analysis. Elective patients were not considered. In addition, patients in the maternity ward or under the age of 18 were not considered.

Data and estimation

In the analysis period from 27 February 2015 to 24 July 2017 the Central Maine Medical Center admitted 21,230 inpatients that met the criteria outlined in the previous section. This

---

[13] One of the primary advantages of the survival random forest over other approaches is that the algorithm automatically considers right-censored observations. However, since full cohorts of discharge data is being used in the analysis, there are no observations that have not had any outcome. An additional advantage of the algorithm is that competing risks to discharge can also be assessed. However, this is not included in the analysis.

corresponds to 103,247 patient days, with a mean length of stay of 4.86 days. Available

predictors ranged from personal information (e.g. insurance status) to clinical information (e.g.

primary diagnoses). An overview can be seen in Figure 2. To keep the creation of the model

grounded in its potential applications, only information that would be readily available to the

hospital during the patient's stay) was used.

Figure 2: Prediction data

Variable name	Description	
los	Length of stay (outcome)	Mean: 4.86, median: 3, s.d: 5.8, IQR: 4
age	Patient's age on admission	Mean: 62.4, median: 65, s.d: 18.1, IQR: 25
sex	Patient's sex	49.2% female
drg_glos	Geometric mean of length of stay for the patient's diagnosis related group, across a national sample	Mean: 4.1, median: 3.6, s.d: 2.2 IQR: 2.0
drg_alos	Arithmetic mean of length of stay for the patient's diagnosis related group, across a national sample	Mean: 5.11, median: 4.4, s.d: 2.8, IQR: 2.7
drg_type	Diagnosis related group type (surgery vs medical)	74.7% medical, 25.3% surgery
numberofdxs_poa	Number of diagnoses that are present on admission	Mean: 15.7, median: 15, s.d: 7.27, IQR: 10
weekend_admit	Dummy variable, whether the patient was admitted on a weekend	25.2% admitted on weekend
off_hour_admit	Dummy variable, whether the patient was admitted between 11pm and 7am	17.6% admitted off hours
uninsured	Dummy variable, whether the patient has health insurance	5% uninsured
elos	The elapsed length of stay in days (how long the patient has already been in the hospital)	Changes dynamically

census	Number of patients in the hospital	Mean: 123.4, median: 108, s.d: 37.9, IQR: 22
month	Month on patient day (changes dynamically)	January: 6.25% February: 6.51% March: 9.75% April: 9.37% May: 9.52% June: 8.98% July: 11.46% August: 8.99% September: 8.54% October: 8.52% November: 5.97% December: 6.12%

It is worth noting that the sample size of this study is approximately five times larger than the samples in prior machine learning survival studies. [14] Figure 3 depicts the distribution of length of hospital stays across the cohorts of patients in our sample. The mean length of stay was 4.86 days, with an interquartile range of 4 days. Hospitals vary in their distributions of length of stay. For example, in the hospital studied by Barnes (2015) the mean stay was only 2.16 days, with an interquartile range of 1.76.

---

[14] Barnes, Sean, et al. "Real-Time Prediction of Inpatient Length of Stay for Discharge Prioritization." *Journal of the American Medical Informatics Association.* 2015.

Figure 3: Distribution of Patient LOS at Central Maine Medical Center
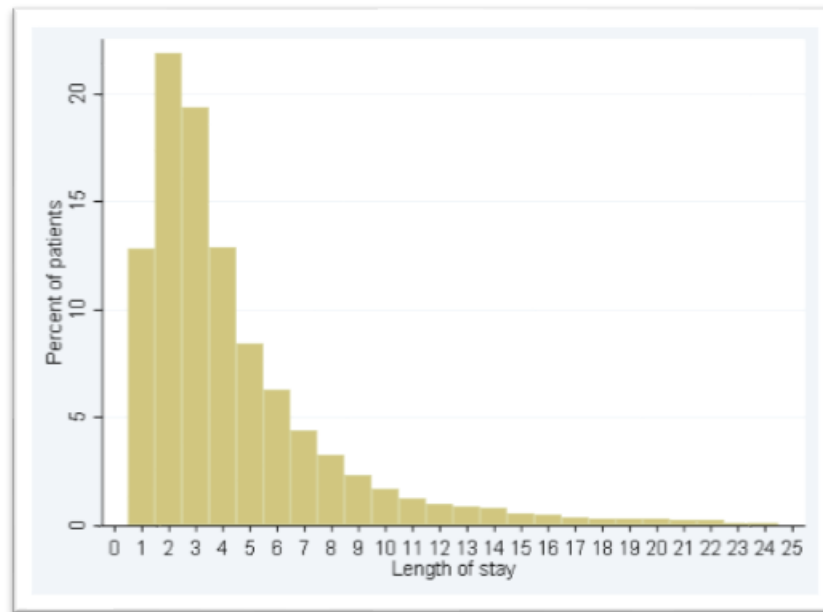
The survival random forest model was generated using the randomForestSRC [15] package in R. In machine learning, a common practice is to separate the data into "training" and "testing" samples. The training sample is used to fit the model, while the testing data is used to assess the model's performance. Patients discharged before 27 February 2017 (approx. 2 years of data) were assigned to the training data, while the remaining patients were assigned to the testing data (approx. 6 months of data). This corresponds to a roughly 80%-20% split of the data, which is common practice in machine learning methods.

When using the Ishwaran et al. algorithm described above, splitting the sample into a training and testing set is not necessarily required, as error can be evaluated on the out of bag samples created by the algorithm. However, splitting the data became necessary in this study to compare the performance of the survival random forest to other prediction methods.

---

[15] Ishwaran, Hemant, et al. "Random Survival Forests." *The Annals of Applied Statistics*. 2008.

Due to the aggregation of thousands of survival trees to generate the survival random forest, it cannot be visualized as simply as a survival tree. Instead, the relative impact of each variable on predictions can be displayed in a variable importance plot. The Breiman-Cutler importance measure is the most frequently applied importance measure for random forests, and is displayed in Figure 4.[16] To calculate the importance of a variable in a tree, the given variable is given random values and dropped down the tree. Then, the true values are dropped down the tree. The variable importance is the difference in error of the tree with random values and the tree with the true values for the variable. A variable's importance in a random forest is the average of the variable's importance across all of the trees.

Figure 4 shows that in the model estimated here, deleting drg_glos alone, the national geometric mean length of stay among the observed patient's diagnosis group, or numberofdx_poa alone, the number of diagnoses present on admission, would decrease the model's Breiman-Cutle accuracy by .05 and .02 respectively. Survival prediction error is evaluated by measuring the squared difference at each time t between the predicted event probability and the true event probability (zero or one). This measure is called the Brier score. Brier scores range from 0 to 1, with scores closer to 0 being indicative of a more accurate model. A Brier score of .25 corresponds to predicting a 50% probability of an event occurring at each time point.

---

[16] Strobl, Carolin, et al. "Conditional Variable Importance for Random Forests." *BMC Bioinformatics*. 2008.

Figure 4: Variable Importance Plot for SRF
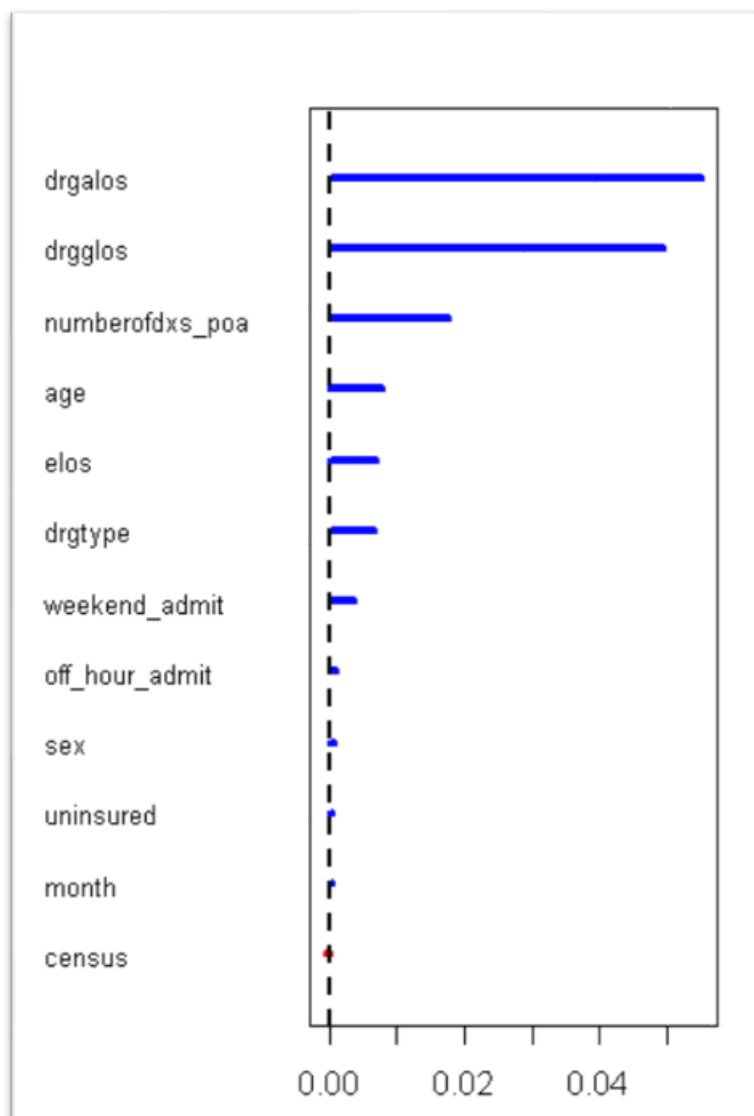
In all, this machine learning algorithm uses twelve explanatory variables for its predictions. Upon generating the survival random forest, test observations can be inputted into the model to generate predicted probabilities of discharge for each patient on each day. In addition, hazard functions can be generated for each patient. The quality of these predictions will be discussed in the following section.

Analysis

In this section, I take three approaches to evaluating the survival random forest's prediction accuracy. First, I replicate the prediction error curves used in Mogensen et al. [17] This approach allows for visual comparison of the prediction error between the survival random forest, the Cox proportional hazards model and the forecasting method currently employed by CMMC. The second approach is to consider the model's performance in terms of true positive and true negative rates. Finally, I evaluate the model in an applied context.

In survival analysis, the most important benchmark is the Brier score of a prediction model which ignores all predictor variables.[18] This null model is generally estimated using the Kaplan-Meier statistic. The Kaplan-Meier is a commonly used nonparametric method of survival analysis.[19] In addition, to provide a comparison of the performance of the models to forecasts currently available to CMMC, I generated predicted dates of discharge for each patient using the geometric mean length of stay nationally for the patient's diagnosis related group (DRG). Under

[17] Mogensen, Ulla B., et al. "Evaluating Random Forests for Survival Analysis Using Prediction Error Curves." *Journal of Statistical Software.* 2012.

[18] Goel, Khanna and Kishore. "Understanding survival analysis: Kaplan-Meier estimate". *International Journal of Ayurveda Research.* 2010.

[19] Kaplan, E L, and Paul Meier. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association.* 2008.

this model, the patient's estimated probability of discharge is zero until their DRG estimated discharge date, on which the estimated probability of discharge is one.

Approximately 80% of patients were randomly assigned to training data. All models were trained on this data. The models generated probability of discharge predictions for the remaining patients. The performance of the models relative to true outcomes is visualized in Figure 5.

Figure 5: Prediction Error Curves
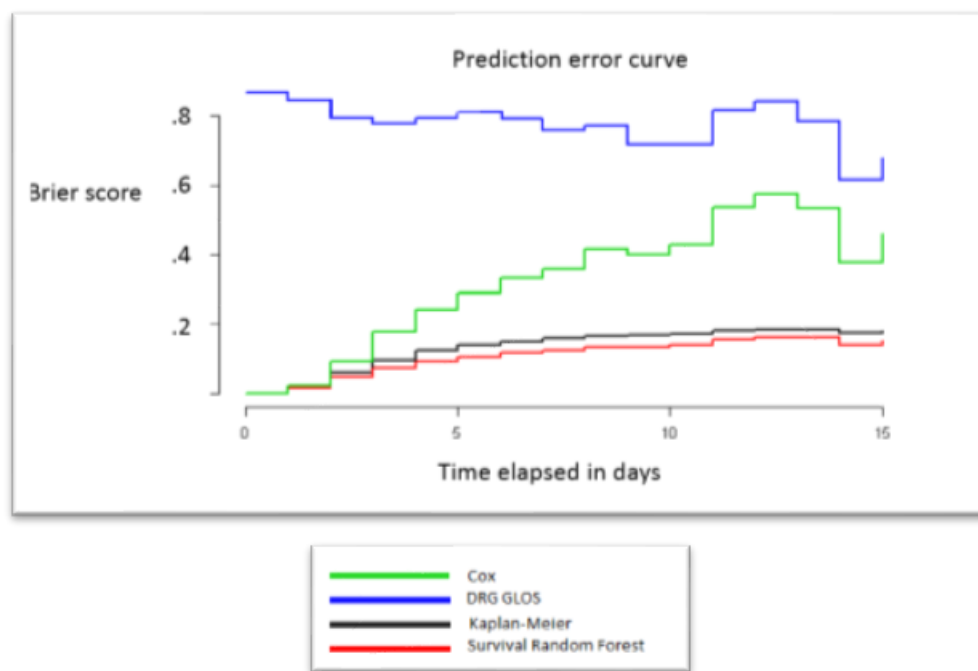
As can be seen in the prediction error curves, the survival random forest had the lowest prediction error of all models, including the Kaplan-Meier null model reference line. The superiority over the Cox model is not surprising, given its previously discussed limitations. The survival random forest also clearly outperforms current CMMC forecasting practice in terms of overall accuracy. However, it is worth noting that the model exploits the fact that relying solely

on the geometric mean of length of stays does not take into account the distribution of length of stay.

The survival random forest model approaches the task of estimating a patient's length of stay through a far more complex approach than simply taking the national mean length of stay for a patient's diagnosis. Given the significant costs associated with forecasting and changing workflows, hospital administrators should consider for which specific applications a machine learning approach is worth the increased complexity.

To inform such a decision, I calculate the true positive and true negative rates of the model and of the drg_glos-only approach. In prediction modelling, as with medical diagnostics, errors are generally classified as one of the following two types. True positives [20] occur when the model correctly predicts that a patient will be discharged on a certain day. True negatives [21] occur when the model correctly identifies a patient who does not end up being discharged on that day.

All survival analysis models face a tradeoff between their true positive and true negative rates. For example, the surest way to a true positive rate of 1 (100%) is to predict that on each day, every single patient will be discharged. In practice, a model forecasts the likelihood of discharge on each day, and then a user must make a decision as to where the probability cutoff should lie for the prediction to be "discharge" or "stay". This tradeoff can be represented through the receiver operating characteristic ("ROC") curve, which graphs the false positive rate on the x-axis and the true positive rate on the y-axis as a function of the selected decision cutoff. If equal weight is assigned to the false positive rate and the true positive rate, the optimal cutoff is

---

[20] The true positive rate is commonly referred to as "sensitivity"
[21] The true negative rate is commonly referred to as "specificity"

at the maximum of (true positive rate) + (true negative rate) - 1. The accuracy of the model can

be distilled into one measure by calculating the area under the curve. ("AUC")

Figure 6: True vs False Positive Rate of SRF

	Survival Random Forest	DRG GLOS prediction
True positive rate	.70	.18
True negative rate	.62	.88
Area Under Curve (AUC)	.72	.52

At first glance, the AUC scores suggest that the survival random forest is the superior model. However, the true positive and negative rates inform which sort of applications would benefit from the use of machine learning forecast. Interestingly, while the survival random forest is more apt at making individual discharge predictions, when aggregated together to estimate daily total discharges, the DRG GLOS approach comes closer to the true number of daily discharges.

A more meaningful comparison for hospital administrators would be an analysis of how the model performs in an applied setting. To that end, I have selected an additional, one-month testing subset to compare the abilities of the models to predict the daily discharge count relative to the true number of discharges. A caveat here is that the selected testing subset could be subject to seasonal effects. However, the month variable is intended to pick up the effect of changing seasons.

The results of aggregating the predictions into daily discharge predictions are shown in Figure 7. While the survival random forest generates more reliably accurate for specific individuals, it does not seem well suited to making predictions of total discharge. The survival random forest had a lower mean squared error than the DRG GLOS model, but the mean DRG GLOS prediction came much closer to the mean number of true discharges.

Figure 7: Aggregated Discharge Predictions

	SRF predicted discharges	DRG predicted discharges	True discharges
Mean	27.05	22.47	22.72
Standard deviation	3.03	4.00	5.38
Mean error	-4.33	.25	-
Mean squared error (relative to truth)	49.3	55.81	-
R-squared	.04	.09	-

In the 32 days between 15 March 2017 and 15 April 2017, the mean difference between

the true number and the survival random forest predicted number of discharges was 4.3 days. In

contrast, using the DRG GLOS prediction method was only -.25. This likely reflects how the

DRG GLOS model is stronger in true negative rates versus true positive rates. The survival

random forest overestimates the total number of discharges. Specifically, SRF estimated total

discharge was greater than true discharge 68.7% of days. The DRG GLOS model only

overestimated true discharge 46.8% of the time.

Physicians and hospital administrators often need to make predictions more specific than

a total number of discharges. Individuals and their families are likely more concerned with their

own length of stay forecast than an aggregation of all current patients. Moreover, hospital

administrators are challenged to balance the needs of specific, heterogeneous units within the

hospital by allocating the right workers with the right skills to each unit. For example, specialist

nurses, social workers or physical therapists. The survival random forest provides a more useful prediction in this regard.

Conclusion

To summarize the findings of this study, the survival random forest is more accurate than the Cox proportional hazards model and an approximation of current CMMC practice when estimating discharge probabilities for specific patients. For a small subsample of the data, this advantage is lost when attempting to predict total number of discharges. This information is worth knowing for hospital decision makers considering the implementation of forecasting analytics. Specifically, this model is best suited to inform workflows centered on specific patients. Hospital administrators seeking forecasts of their total bed occupancy should instead use models developed for such a purpose. An avenue for further research could be evaluating the accuracy of the model or models like it in more specific contexts, e.g. in specific sub-units of the hospital.

Quantifying the economic impact of improved information technology in hospitals could also be the subject of interesting further research. While improvements in operations management may not remedy all of the challenges facing the US healthcare system, improved utilization could foreseeably help bring healthcare costs down.

Works Cited

Barnes, Sean, et al. "Real-Time Prediction of Inpatient Length of Stay for Discharge

    Prioritization." *Journal of the American Medical Informatics Association*, vol. 23, no. e1,

    July 2015, doi:10.1093/jamia/ocv106.

Bentley, Tanya G.K., et al. "Waste in the U.S. Health Care System: A Conceptual Framework."

    The Milbank Quarterly, vol. 86, no. 4, 2008, pp. 629–659.

Berger, Eric. "Breaking Point: Report Calls for Congressional Rescue of Hospital Emergency

    Departments." *Annals of Emergency Medicine*, vol. 48, no. 2, 2006, pp. 140–142.,

    doi:10.1016/j.annemergmed.2006.06.022.

Bou-Hamad, Imad, et al. "A Review of Survival Trees." *Statistics Surveys*, vol. 5, 2011, pp. 44–

    71., doi:10.1214/09-SS047.

Breiman, Leo, et al. *Classification and Regression Trees*. Vol. 1, Brooks/Cole Publishing, 1984.

Breiman, Leo. "Machine Learning." *Machine Learning*, vol. 45, no. 1, 2001, pp. 5–32.,

    doi:10.1023/a:1010933404324.

Goel, Manish Kumar, et al. "Understanding Survival Analysis: Kaplan-Meier Estimate."

    International Journal of Ayurveda Research, vol. 1, no. 4, 2010, pp. 274–278.,

    doi:10.4103/0974-7788.76794.

Goel, Manish Kumar, et al. "Understanding Survival Analysis: Kaplan-Meier Estimate."

    International Journal of Ayurveda Research, vol. 1, no. 4, 2010, pp. 274–278.,

    doi:10.4103/0974-7788.76794.

Ishwaran, Hemant, et al. "Random Survival Forests." *The Annals of Applied Statistics*, vol. 2, no.

    3, 2008, pp. 841–860.

James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer,

    2013.

Jones, Spencer S., et al. "Forecasting Daily Patient Volumes in the Emergency Department."

    Academic Emergency Medicine, vol. 15, no. 2, 7 Feb. 2008, pp. 159–170.,

    doi:10.1111/j.1553-2712.2007.00032.x.

Kaplan, E L, and Paul Meier. "Nonparametric Estimation from Incomplete Observations."

    Journal of the American Statistical Association, vol. 53, no. 282, June 1958, pp. 457–

    481., www.jstor.org/stable/2281868.

Kennedy, Peter. "Limited Dependent Variables." *A Guide to Econometrics*, 4th ed., The MIT

    Press, 1998, pp. 259–261.

Mak, Gregory, et al. "Physicians' Ability to Predict Hospital Length of Stay for Patients

    Admitted to the Hospital from the Emergency Department." *Emergency Medical

    International*, no. 824674, 2012, doi:10.1155/2012/824674.

Mogensen, Ulla B., et al. "Evaluating Random Forests for Survival Analysis Using Prediction

    Error Curves." *Journal of Statistical Software*, vol. 50, 18 Sept. 2012,

    www.jstatsoft.org/article/view/v050i11.

Strobl, Carolin, et al. "Conditional Variable Importance for Random Forests." *BMC

    Bioinformatics*, vol. 9, 2008, p. 307., doi:10.1186/1471-2105-9-307.

Van Walraven, Carl, and Alan J Forster. "The TEND (Tomorrow's Expected Number of

Discharges) Model Accurately Predicted the Number of Patients Who Were Discharged the Next

Day." *Journal of Hospital Medicine*, 23 Aug. 2017, doi:10.12788/jhm.2802.

Appendix 1: Testing proportional hazards assumption

A common way to test the proportional hazards assumption is to create an additional variable for each explanatory variable that measures the interaction of time and the explanatory variable. [22] For the proportional hazards assumption to be substantiated, the new variables should have estimated coefficients insignificantly different from zero. Figure A demonstrates that age, female, drg_glos, drg_alos and off_hour_admit all are significantly different from zero to the .05 level. Therefore, the proportional hazards assumption cannot be justified.

<u>Figure A: Interaction with time and explanators</u>

Variable name	Coefficient on time interaction	Chi squared	P-value
age	.02626	10.71	.0011
female	.01969	6.19	.0128
drg_glos	.10292	156.80	< .00001
drg_alos	-.04843	35.09	< .00001
numberofdx_poa	.00745	.86	.3538
weekend_admit	-.00775	.96	.3280
off_hour_admit	-.02102	6.99	.0082
uninsured	-.00395	.25	.6189

---

[22] Kennedy, Peter. "Limited Dependent Variables." *A Guide to Econometrics*, 4th ed., The MIT Press, 1998, pp. 259–261.