Death versus Data Science: Predicting End of Life

Muhammad A Ahmad, Carly Eckert, Greg McKelvey,
Kiyana Zolfagar, Anam Zahid, Ankur Teredesai
KenSci Inc., Seattle, WA USA
{muhammad, carly, greg, kiyana, anam, ankur}@kensci.com

Abstract

Death is an inevitable part of life and while it cannot be delayed indefinitely it is possible to predict with some certainty when the health of a person is going to deteriorate. In this paper, we predict risk of mortality for patients from two large hospital systems in the Pacific Northwest. Using medical claims and electronic medical records (EMR) data we greatly improve prediction for risk of mortality and explore machine learning models with explanations for end of life predictions. The insights that are derived from the predictions can then be used to improve the quality of patient care towards the end of life.

Introduction

In the United States, 22.2% of Medicare decedents die in acute care hospitals (Bekelman et al. 2016). For many, the last six months of life are full of physician visits, medical procedures, and hospital stays (Marik 2015) (Setoguchi, Stevenson, and Schneeweiss 2007) (Halpern 2015). Such aggressive medical care comes with a high price. In 2011, the U.S. health system spent $205 billion on the care of individuals in their last year of life (Aldridge and Kelley 2014). Yet, this incessant and expensive care is without commensurate improvement in outcomes or quality of life (Zhang et al. 2009). Additionally, 70% of Americans wish to die at home (Barnato et al. 2009). Services such as palliative care and hospice are options designed to provide an alternative to hospital-based medicine for patients to spend their last months before death. Referral to these services can also save on healthcare costs, if done in a timely manner (Hogan et al. 2001). Recent work by Morrison and colleagues suggest cost savings associated with palliative care patients of between nearly $1700- $5000 per admission (Morrison et al. 2008). Other research suggests that Advance Care Planning, the professionally facilitated approach to discussing end of life care needs, is also associated with net cost savings in end of life populations (Klingler, in der Schmitten, and Marckmann 2016). The need to improve patient experience and quality outcomes while reducing health care costs has fueled a growing interest in identifying opportunities for improving end of life care (Hogan et al. 2001).

The number of Americans using palliative care services continues to grow and was estimated at 1.7 million, or about 46% of those who die (NHPCO 2016). Yet these services are being utilized too late: the median length of stay in hospice care in 2016 was only 23 days. Additionally, 28% of hospice patients were discharged or died within 7 days of hospice enrollment (NHPCO 2016). In work by Christakis and colleagues, they suggest that hospice clinicians consider 80-90 days of hospice care as optimal for the needs of patients and their families (Christakis 1997). Surveys of family members of decedents indicate that satisfaction with end of life care is correlated with their perception of timeliness of hospice referral (Teno et al. 2007). Finally, providers that commonly encounter in-hospital patient death, like intensivists and critical care nurses, have high rates of professional burnout (Embriaco et al. 2007). It follows to conclude, therefore, that timely and appropriate end of life care impacts all aspects of the Quadruple Aim in healthcare (quality, satisfaction, cost savings, and provider satisfaction).

Patients require physician referral for hospice services and there is evidence that physician discomfort with end of life conversations may be partly responsible for the delay in these referrals (McGorty and Bornstein 2003). Additionally, physicians have difficulty with predicting mortality, particularly among patients that they know well (Christakis 1998). For patients to qualify for hospice programs, clinicians must estimate that the patient has less than six months to live (Medicare 2017). The prognostic error associated with these estimations is widespread and quite large. Studies show that physicians are inaccurate when predicting time to patient death and tend to be overly optimistic about patient survival. In one study of hospice patients, physicians overestimated patient survival five-fold (Christakis et al. 2000). This deficit of prognosis can be detrimental for patients, their families, and the health care system. This prognostic error can result in late referrals to hospice, can set false expectations of survival for families and patients, and can lead to missed opportunities to focus on quality of life. Such expectations can delay important discussions between families thus leading to more difficult decisions made during times of crisis. Studies show that delays in discussing end of life choices ultimately lead to more aggressive in-hospital treatments, more in-hospital deaths, and higher costs (Wright et al. 2008). The ability to provide an earlier option of palliative care and hos-

pice services decreases emergency department visits, stays in the intensive care unit (Teno et al. 2007), hospital days, and rates of thirty-day hospital readmission and in-hospital death (Kelley et al. 2013). Utilizing palliative care and hospice can significantly improve care value for patients and their families near the end of life (Kelley et al. 2013). Although less often described, the situation may occur when a patient is discharged from hospice alive. This may be related to prognostic error, unexpected changes in disease course, or patient preferences to resume curative care (Teno et al. 2014). Nearly 1 in 5 hospice patients is discharged alive, a situation that proves expensive and potentially detrimental to further care. In a study by Teno et al., 7% of patients discharged alive from hospice were hospitalized immediately after and then reentereed hospice, costing Medicare an additional $95 million in hospital expenditures (Teno et al. 2014). To improve quality and reduce unnecessary hospitalizations near the end of life, there is an urgent need to help healthcare providers identify the most vulnerable and at-risk patients by providing them with beneficial care coordination and supportive care services (Donzé, Lipsitz, and Schnipper 2014).

Machine learning has much to offer in the area of end of life care. However, little has been published on machine learning techniques in this space and none, to our knowledge, are widely used clinically. By analyzing an array of patient specific features we present a framework for predicting patient mortality 6-12 months from the date of prediction as shown in Figure . We suggest our method as clinically important as it gives providers crucial insight into the likelihood of patient death. Combined with clinical insight, this prediction model can better inform patient and family discussions, lead to earlier hospice referrals, and better end of life experiences for the sick and dying. We consider data from two health systems which correspond to two different populations. In one health system the data comes from the general Medicare population while in the other the data is from a cohort of patients diagnosed with heart failure.

Although machine learning techniques for predicting risk of mortality and the probability of death have been studied in a variety of problems as a risk scoring problem, to the best of our knowledge the machine learning literature on predicting mortality to inform end of life care is quite sparse. In this paper, we seek to address this deficiency. The main contributions of this paper are as follows: (i) Present a machine learning model and framework to predict end of life or risk of mortality for at risk patients. (ii) Describe a scalable cloud based system that delivers insights to multidisciplinary care teams. (iii) Create model explanations for end of life predictions for individual patients which may foster trust in the predictions from the perspective of clinicians using the system.

Related Work

Survival analysis has a long history in the medical domain where life tables and statistical inference have been used to predict life expectancy for patients (Chiang 1984). Most models that are used in the medical domain are either based on actuarial tables (Cox 1992) or scoring based models

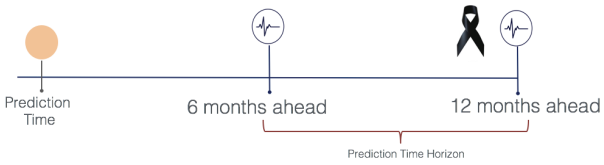

Figure 1: Time window of Prediction

which provide a score as a linear combination of factors identified by domain experts (Pollack, Ruttimann, and Getson 1988). Alternatively, work by Moss et al. demonstrated the predictive performance of a simple ”surprise question” in predicting time to death among dialysis patients (Moss et al. 2008). In this study, clinicians are prompted to ask themselves: ”Would I be surprised if this patient died within the next 12 months?” Initially touted as a powerful and useful predictive tool, a recent systematic review of the tool's use in predicting 6-18 month mortality in other patient cohorts revealed poor performance (Downar et al. 2017). Another recent systematic review evaluated 16 validated non-disease specific prognostic indices for mortality (Yourman et al. 2012). One of the better performing indices evaluated (AUC 0.79), by Gagne et al, utilized administrative data including age, sex, and indicator variables corresponding to 20 comorbidities used in the Elixhauser and Charlson indices predicted 1 year mortality among low-income elderly (Gagne et al. 2011). Similarly, disease specific cohorts have also been used in machine-learning based predictive algorithms. The Seattle Heart Failure Model, for example, is a disease specific predictive tool that is in widespread clinical use. Derived from a multivariate Cox model using easily obtained clinical features, the Seattle Heart Failure Model has been prospectively validated and has excellent performance predicting mortality at 1 year (AUC 0.75 to 0.81) (Levy et al. 2006). Setoguchi et al used machine learning models to predict risk of worsening conditions for patients with cardiovascular diseases (Setoguchi, Stevenson, and Schneeweiss 2007). Finally, there are machine learning based models that predict mortality across disease cohorts. Makar et al. used administrative data to predict mortality within 6 months among specific disease cohorts within the Medicare population. Their findings concluded that administrative models required augmented features informed by clinical status to boost performance over baseline methods with an AUC of 0.826 (Makar et al. 2015).

Model Interpretability

Model interpretability is an important aspect of prediction in the medical domain where the people involved in decision making may ask for explanations for the predictions (Lakkaraju, Bach, and Leskovec 2016). Opacity of the machine learning model can lead to mistrust, possibility of error, and even abuse (Domingos 2016).

Model interpretability in machine learning has multiple definitions (Lipton 2016): (i) It may refer to interpretability as providing causal links which can then be used to generate hypotheses to test experimentally (Lou, Caruana, and

Gehrke 2012). (ii) Simulatability refers to the notion of comprehending the entire model all at once by a human (Lipton 2016), (Ribeiro, Singh, and Guestrin 2016). Thus a decision trees with 8 nodes and depth 2 may be cosidered more interpretable than a decision tree with 800 nodes and depth 20. (iii) Interpretability may refer to all the components of the model being amenable to intuiative explanations (Lou, Caruana, and Gehrke 2012) (Turner 2016) (Lakkaraju, Bach, and Leskovec 2016) e.g., the relative strength of parameters in linear models correspond to strengths of associatoins between the features and the label to be predicted. The other implication of this notion of interpretability is that certain complex engineered features may render a model non-interpretable (Lipton 2016). (iv) It can refer to as means to engender trust in the model (Kim 2015) which in turn may refer to trust in a model's performance or robustness. In this paper, the notion of model interpretability as engendering trust and giving intuitive explanations is employed.

A Scalable Framework for Mortality Prediction

Creating predictive models around the possibility of death can greatly enhance the welfare of individual patients and their families. Predicting end of life care could play a role in improving time to hospice referral, reducing readmissions, and increasing patient and family satisfaction. Given the complexity of medical data and problems inherent in integrating predictive models in patient care cycle, we propose a framework for predicting risk of mortality over timescales relevant to optimizing end of life care.

Nearly all end of life prediction problems are set up to address cases where the prediction window is from the immediate present to a point in time in the future e.g., predict end of life probability for the next six months, predict end of life probability for the next five years etc. The primary deficiency of current approaches to anticipating impending mortality and initiating end of life care planning is timeliness. In order to maximize the benefit of end of life care, adequate timeliness of prediction is needed, both to enable the optimal duration of intervention and to allow for the multiple conversations often necessary to arrive at a plan of care (Balaban 2000). Thus, the models that we consider in this paper are for predicting end of life for six to twelve months from the time of prediction.

The system described in this paper is uses machine learning models to issue mortality scores as well integrate explanations associated with these predictions. The machine learning model is deployed as a single layer binary classification layer that can be accessed by a cloud based app from any browser. This system uses data from the hospital systems' claims live feed or Electronic Health Record (EHR) data feed. New data can be continuously pulled into the cloud which is then transformed into a standardized schema. The standardized schema allows the capability of adding new data sources as well as enabling the transfer of data sources in the system with relative ease. Lack of standardization and difficulty in making many medical data formats interchangeable is an important factor in the slow growth of

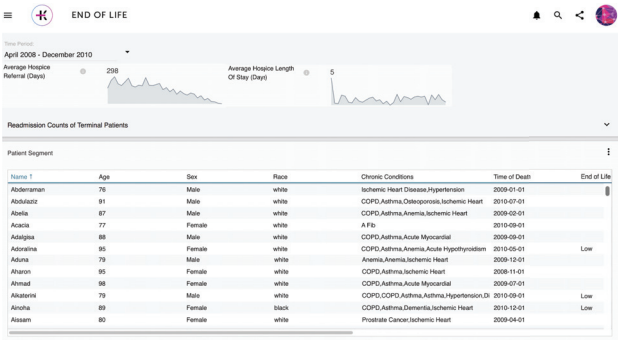

Figure 2: End of Life Prediction App

adoption of advanced machine learning in the medical informatics domain (Luxton, Kayl, and Mishkind 2012).

Figure 2 shows a screen capture of the cloud based app. The cloud based app allows analysis of end of life prediction at the cohort level and the patient level with drill down capabilities. Additionally, the healthcare provider can create custom cohorts of patients to compare e.g., patient outcomes by chronic disease diagnosis, demographic, and geographic patterns. The cloud based app also allows the care provider to explore a detailed patient risk profile and link it to other aspects of patient care e.g., predicting patient's risk of readmission. A care provider can make a decision regarding patient care based on both readmission risk and mortality risk; for example, a care provider may decide to consult palliative care, patient outcomes by chronic disease cohort, demographic, and geographic patterns or have a discussion regarding end of life wishes with a patients family.

While business intelligence tools within health care have varying levels of success and integration with existing health care systems (Raghupathi and Raghupathi 2014), it is their lack of integration with existing systems and difficulty in accessability that has hindered widespread use. The current system can work with a variety of existing medical data formats and can be accessed through any web browser. It should be emphasized that at no point does the data leave the premise of any of the Hospital Systems' cloud to ensure security and compliance with HIPAA. Also, the data that is used for model building has been stripped of Protected Health Information (PHI) in compliance with HIPAA Safe Harbor de-identification (USDHHS 2015).

Data

We consider data from two major hospital systems based in the Pacific Northwest in the United States. The names of these health systems are omitted because of privacy reasons and are henceforth referred to Health System A and Health System B. The population from Health System A comprises patients with a history of heart failure (HF), the population from Health System B consists of patients with any type of illness. We predict readmission for both cohorts regardless of the reason for readmission, this formulation of the problem of predicting readmission is referred to as 'All Cause Readmission Prediction'. Different feature sets were avail-

able for the two data sets which resulted in different models for prediction. Therefore, the results from the two models are not directly comparable and are addressed separately. The data set for Hospital System A consists of 4,888 patients and that for Hospital System B consists of 48,365 patients. In both cases the data spans over the course of less than two years.

Experiments

Feature Set

For Hospital System A the data set is derived from the Electronic Medical Records (EMR) data for and Medical Claims Data for Hospital System B. The claims data contains information regarding billing i.e., how much a medical system is charged for with respect to a procedure that was performed on a patient or any other type of medical claims related information.

Health System A The feature sets that are available for Health System A come from the Electronic Medical Records (EMR) data. This includes demographic features, patient length of stay, overall cost related features, specific cost related features (in-patient, out-patient, home health, hospice, skilled nurse facility), readmissions related features, and counts of procedures performed for different procedure types. The Healthcare Common Procedure Coding System (HCPCS) codes are used after feature transformation (for Medicare & Medicaid Services and others 2003). HCPCS codes include indicators for services like ambulance services, durable medical equipment, prosthetics, orthotics and supplies (DMEPOS). This coding system is also used as an official code set for outpatient hospital care, chemotherapy drugs, Medicaid, and other medical services. There are more than a thousand HCPCS codes in the data set and using the count of each individually would lead to a very sparse representation. We first map the set of HCPCS code to a TF-IDF space (Sparck Jones 1972) and then use features from the transformed space in our models.

Health System B Only the claims data set is available from Health System B. The feature set used for Health System B consisted of demographic features, patient length of stay, cost related features, and counts of comorbidities. We note that all information is deidentified so that it is not possible to link a record back to a patient.

Results

We pose the problem of end of life prediction as a binary classification problem. The training period includes data from six months prior to the patient's death for deceased patients and six months prior to the last available encounter for non deceased patients. The test period is six month to one year. The baseline can be described as follows: We compare prediction metrics derived from the implicit mortality prediction model used by healthcare organizations. A prediction by a healthcare organization is defined as happening when a patient is enrolled in hospice. It should be noted that hospice enrollment represents only a subset of actions that a healthcare system may make to indicate that death is predicted. The reason to use hospice for baseline is that it is visible in the data, and it includes an explicit prediction window of 6 months (CMS 2016). A confusion matrix for healthcare organization prediction can then be constructed by the following ruleset:

- True positive - The patient was enrolled in hospice within 6 months of death
- False positive - The patient was enrolled in hospice and did not die within 6 months
- True negative - The patient has never been enrolled in hospice and has not died
- False negative - The patient died and was not enrolled in hospice. However it may be the case that the patient was referred but not enrolled. The implication being that we do not know if the false negative was due to patient or family refusal.

For both data sets, the following models were tried for the binary classification problem: Adaboost, Random Forests, Support Vector Machines, Naive Bayes, Bayes Net, Extreme Gradient Boosting, CART and GLM (Generalized Linear Models). The best results were obtained from Extreme Gradient Boosting and we report the results from this model in all the cases described here. The output from the model is a scaled risk score between zero and one. We use a threshold function such that if the score is above the threshold then it is flagged as prediction for end of life otherwise it is flagged as surviving.

Table 1: Performance Metrics for EOL Models for Hospital System B

Metric	Accuracy	Recall	Precision	AUC
Baseline	0.88	0.37	0.07	0.71
ML Approach	0.89	0.41	0.09	0.79

The results for predicting end of life for Hospital System B are given in Table 1. The precision of both the baseline and the proposed approach is low. In the EOL prediction domain practitioners usually focus on AUC and recall as the metric for assessing performance (Arabi et al. 2003). The proposed approach performs much better than the baseline on all of these metrics. Hanley et al descibe the problem and solution for comparing AUC coming from different models on the same source data (Hanley and McNeil 1982). Using their approach we get a p-value of 0.0957 for the significance of the difference between the two ROC Curves, which implies a statistically significant difference between the two. Since decisions regarding how to handle care for sick patients, allocating hospital resources and sending patients to hospice are based on these models, any improvement over the baseline is considered to have major repercussions for patient care, quality of life, and cost savings to health systems.

In hospital System B the average number of days spent in hospice by a patient is 37 days. We also compare how our models perform as compared to the baselines for the 37 days time period for the patients who are admitted to hospice. The feature set that we use is the same as the feature set used for predicting end of life for the 6 to 12 month period.

Table 2: Prediction for Hospice for Hospital System B

Metric	Accuracy	Recall
ML Approach	0.944	0.517

Table 3: Performance Metrics for EOL Models for Hospital System A

	Accuracy	Recall	Precision	AUC
Month 1	0.853	0.371	0.054	0.706
Month 2	0.894	0.312	0.119	0.705
Month 3	0.871	0.459	0.096	0.766
Month 4	0.914	0.616	0.162	0.879

The results are given in Table 2. These results give better performance as compared to the results for predicting for a longer period of time i.e., six to twelve months.

A baseline for predicting end of life for months in advance is not a current practice for Hospital System A. However, their accuracy for predicting end of life for shorter periods of time, days to weeks, is 0.89. Among these, the median days to death for patients who were referred to hospice was 6 days. The performance for the proposed model is given in Table 3 which shows that the model does perform quite well. Since we have access to data that spans a longer time period, we also use a moving window for the training and the test period to make predictions for four different months. A noticeable improvement in performance is observed for months 3 and 4 as compared to the first two months. In all cases, however, the results are as good or better than what has been reported for performance metrics in the literature. Also, the data set from Hospital System A is large enough to cross validate the models which showed similar performance.

Interpretable Prediction Models

One problem with many machine learning risk prediction models is that it is difficult to interpret the results of prediction i.e., figuring out why a certain prediction was made. There are exceptions to black box models in the form of models like Decision Trees (Quinlan 1986), Bayesian Rule Lists (Yang, Rudin, and Seltzer 2016), Regression Trees (De'ath and Fabricius 2000) etc. While these techniques give

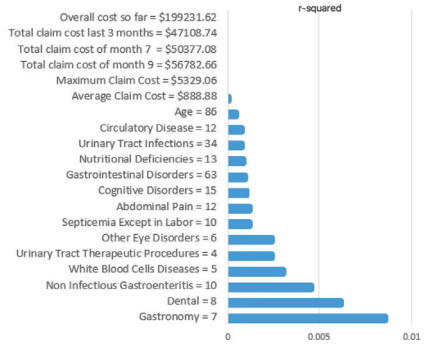

Figure 3: EOL: 0.580, R-Square: 0.57

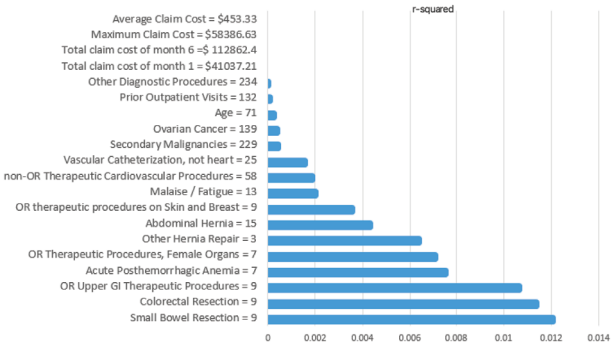

Figure 4: EOL: 0.994, R-Square: 1.00

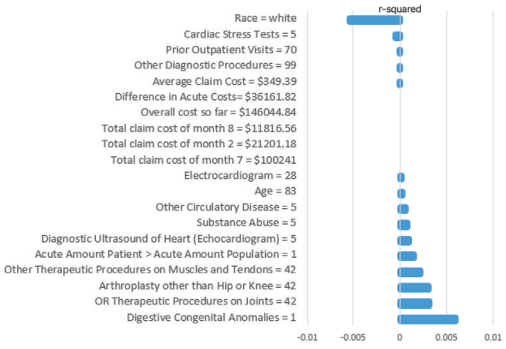

Figure 5: EOL: 0.993, R-Square: 0.94

us a global model of how predictions are made, it is not always clear how predictions for individual instances were made. A recent model proposed by (Ribeiro, Singh, and Guestrin 2016) Local Interpretable Model-agnostic Explanations (LIME) builds approximate models on top of more complex prediction models to give reasons for prediction for individual instances. Given a class of interpretable models $g \in G$ and let $\Omega(g)$ be the complexity of the model. Let $\pi_x(z)$ be a proximity measure between x and z i.e., how similar the two instances are to one another. Let $\Lambda(f, g, x)$ measure of how the explanation g is faithful to the original model $f : \mathbb{R}^d \rightarrow \mathbb{R}$, where for a classification task f. The task of producing explanations is given by :

$$\xi(x) = \underset{g \in G}{argmin}\ \Lambda(f, g, x) + \Omega(g) \qquad (1)$$

In linear models, the complexity of the model $\Omega(g)$ which is directly related to interpretability is given by the number of variables with non-zero weights. Given that our feature sets consist of 500 variables, using LIME to come up with an explanation for predictions, the resulting explanations would not be interpretable by humans. Thus even before running LIME, we reduce the feature space by running regression on the feature space and choosing only top k features before there is a sharp changes in the weights associated with the features. In the current end of life prediction case $k = 20$.

LIME uses a linear model to approximate a global model in the local space. To quantify how well the local model, in this case a regression model, is approximating the global

model we use the R^2 of the linear model as a measure of fidelity between the two models. Figures 3, 4, 5 show examples of output from the LIME model for three different instances. Here we use the convention used by (Ribeiro, Singh, and Guestrin 2016) where the width and intensity of the bar represent how much the variable is contributing towards the prediction.

Given the space constraints we consider three examples of models explanations that correspond to three different scenarios of fidelity between the global and the local model. Consider example in Figure 3 which shows a case where there is a somewhat medium level probability for EOL as well as medium level value for R-square for the linear model. The top factors that were associated with the correct prediction are the counts of certain medical procedure medical procedures performed on the patient. These counts are proxies for how many times a patient has been admitted and the higher count is an indicator for deteriorating condition of a patient. Thus one gets a lower risk score prediction for mortality risk for the non-serious conditions. We note that it there may be instances where the R^2 for the underlying model is quite low which implies that the model fidelity for the local model is low with respect to the underlying model. Alternatively it may be the case that the explanations given by the local model are correct but the confidence in the explanations is quite low in this case.

Consider the case in Figure 4 which corresponds a true positive example from the top factors for prediction are related to some serious conditions and the local models align almost exactly with the global model with an R-square value equal to one. Lastly, consider in Figure 5 where both the probability from the underlying model and the local model is again high. In this case if we look at the top factors that went into prediction for the local model then it becomes clear that all these factors correspond to procedures associated with serious ailments. The main idea behind showing the overall prediction and the confidence in the local model is that the care giver can still make a decision regarding care even if the fidelity of the local model is low based on their past experience in patient care.

One way to informally evaluate the explanation-based models is to show the explanations to clinicians and ask if they make sense. While this is not completely rigorous, it can give us some confidence about the face value of the predictions. The evaluation criteria in this case is not the correctness of the prediction i.e., if a person died or not but rather if the explanation for the prediction makes intuitive sense. Consider the following, if a prediction score is high due to many counts of cardiac procedures in a heart failure patient, this explanation may prove rational to providers. We discovered that for cases where the R^2 associated with the local model is high (≥ 0.65) the explanations provided make medically makes sense. The opposite is also true. It should, however, be noted that a low value for R^2 does not imply that the global model is incorrect. It is possible that the decision boundary is non-linear at the neighborhood space where the local model is constructed leading to incorrect explanations (Ribeiro, Singh, and Guestrin 2016). The main idea behind use explanatory models for prediction is that it can

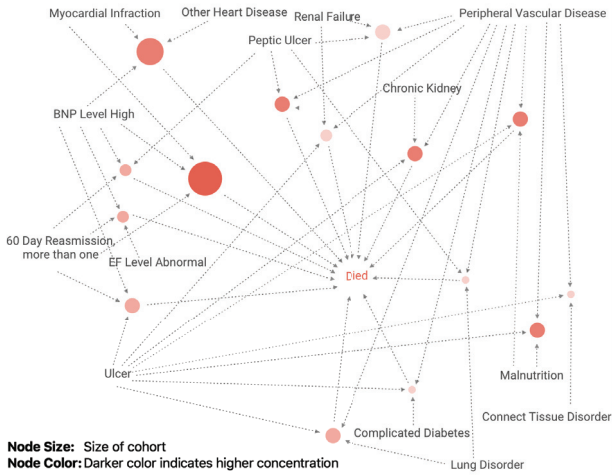

Figure 6: Graphical Factor Analysis of EOL Model

give the clinicians insights about why the prediction is being made. It is possible that there may be cases that for any given machine learning model, the correct prediction is because of an artifact of some underlying distribution of a variable. However if the clinicians can see the factors associated with a prediction then they are able to interrogate and build trust in the model. In cases where the model make a correct prediction based on "unexpected" factors then it could either be because of some artifact of the data, in which case the physician can choose to ignore the prediction. In such a case this information can be used to improve the model. The other possibility is that the model may have picked up a factor which the clinician may not have considered. This observation can be used to do follow up studies that can potentially result in improvement in the state of the art in the field. These insights are integrated into the prediction application where the clinicians can use them to make decisions regarding patient care. We discuss follow up studies to insights models in the follow up section.

Association Analysis Insights Models

We also explore a global insights model of the factors that lead to the death of a patient. We start with Association Rule Mining of the CHF data set for Hospital System A. All rules below a threshold s for support and c for confidence are pruned. We then filter the rules where right had side has the target variable with the interesting outcome. After automatic pruning, all remaining association rules are displayed to the clinician to examines these rules and removes those that do not make clinical sense. For each remaining rule, the clinician compiles zero or more interventions directed at the reason made evident by the rule. A visualization of the top 20 rules with simplified factors is shown in Figure 6. Most real world rule sets, however, are much more complicated. But even with simplified rules, as shown in the Figure 6, it is still possible to gain interesting insights e.g., having multiple readmissions in the last 60 days coupled with high level

of BNP (Brain Natriuretic Peptide) and ulcers greatly increases the risk of mortality. We note that the association rules are not causal implications but are to be used as additional signals by physicians to further explore factors that may be associated with the patient's condition.

Future Work

This paper descibes a system which can provide predictions as well as explanations for a patient's end of life. A rigorous assessment of the system requires that it should be in operation long enough to collect statistically significant data about the clinical use of the system in operation by a sufficient number of clinical staff. However, for both hospital systems more than a hundred thousand patients are covered by each hospital system. Thus, once the system has been in operation for at least six months, we plan to follow up on this analysis with usage information, and how the model is being used in an operational setting. We plan to use insights gained from such an analysis to improve the model itself and use feedback from clinicians to improve the model.

Conclusion

The problem of mortality prediction has been considered in a number of contexts. In this paper we considered the problem of mortality prediction in the context of predicting all-cause mortality for two groups of patients: all-comers and those with heart failure. Data from two different large hospital systems in the Pacific Northwest of the United States was used. Using feature sets derived from claims data and EMR we built binary classification models to predict the end of life for a period of six to twelve months from the time of prediction. The results that we obtained were better than the baseline that is currently used in the healthcare industry. We also created explanation based prediction models based on the LIME (Locally Interpretable Model-agnostic Explanations) to surface model insights. The crux of the contribution is in creating a system that integrates various data sources from health care and providing on demand insights to primary care givers who can then use this knowledge to make better informed decision regarding the lives of their patients.

Acknowledgments

The authors would like to acknowledge our research parters in the two hospital systems for providing us with the data and valuable feedback. We would also like to acknowledge the following people for their valuable feedback: Jamie Chung, Boby George, James Marquardt, Rohan D'Souza and Samir Manjure.

References

Aldridge, M., and Kelley, A. 2014. Epidemiology of serious illness and high utilization of health care. *Dying in America: Improving Quality and Honoring Individual Preferences Near the End of Life*.

Arabi, Y.; Al Shirawi, N.; Memish, Z.; Venkatesh, S.; and Al-Shimemeri, A. 2003. Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: a prospective cohort study. *Critical care* 7(5):R116.

Balaban, R. B. 2000. A physician's guide to talking about end-of-life care. *Journal of general internal medicine* 15(3):195–200.

Barnato, A. E.; Anthony, D. L.; Skinner, J.; Gallagher, P. M.; and Fisher, E. S. 2009. Racial and ethnic differences in preferences for end-of-life treatment. *Journal of General Internal Medicine* 24(6):695–701.

Bekelman, J. E.; Halpern, S. D.; Blankart, C. R.; Bynum, J. P.; et al. 2016. Comparison of site of death, health care utilization, and hospital expenditures for patients dying with cancer in 7 developed countries. *Jama* 315(3):272–283.

Chiang, C. L. 1984. The life table and its applications.

Christakis, N. A.; Smith, J. L.; Parkes, C. M.; and Lamont, E. B. 2000. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *Bmj* 320(7233):469–473.

Christakis, N. 1997. Survival of medicare patients enrolled in hospice programs. In *Present. at the annual American Academy of Hospice and Palliative medicine meeting, Chicago*.

Christakis, N. A. 1998. Predicting patient survival before and after hospice enrollment. *Hospice Journal* 13:71–88.

CMS. 2016. Cms manual system: Pub 100-04 medicare claims processing: Transmittal 2303.

Cox, D. R. 1992. Regression models and life-tables. In *Breakthroughs in statistics*. Springer. 527–541.

De'ath, G., and Fabricius, K. E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81(11):3178–3192.

Domingos, P. 2016. A mystery in the machine. *Organisation for Economic Cooperation and Development. The OECD Observer* (308):1H.

Donzé, J.; Lipsitz, S.; and Schnipper, J. L. 2014. Risk factors for potentially avoidable readmissions due to end-of-life care issues. *Journal of hospital medicine* 9(5):310–314.

Downar, J.; Goldman, R.; Pinto, R.; Englesakis, M.; and Adhikari, N. K. 2017. The surprise question for predicting death in seriously ill patients: a systematic review and meta-analysis. *Canadian Medical Association Journal* 189(13):E484–E493.

Embriaco, N.; Papazian, L.; Kentish-Barnes, N.; Pochard, F.; and Azoulay, E. 2007. Burnout syndrome among critical care healthcare workers. *Current opinion in critical care* 13(5):482–488.

for Medicare & Medicaid Services, C., et al. 2003. *Healthcare Common Procedure Coding System (HCPCS)*. Centers for Medicare & Medicaid Services.

Gagne, J. J.; Glynn, R. J.; Avorn, J.; Levin, R.; and Schneeweiss, S. 2011. A combined comorbidity score predicted mortality in elderly patients better than existing scores. *Journal of clinical epidemiology* 64(7):749–759.

Halpern, S. D. 2015. Toward evidence-based end-of-life care. *New England J. of Medicine* 373(21):2001–2003.

Hanley, J. A., and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143(1):29–36.

Hogan, C.; Lunney, J.; Gabel, J.; and Lynn, J. 2001. Medicare beneficiaries costs of care in the last year of life. *Health affairs* 20(4):188–195.

Kelley, A. S.; Deb, P.; Du, Q.; Carlson, M. D. A.; and Morrison, R. S. 2013. Hospice enrollment saves money for medicare and improves care quality across a number of different lengths-of-stay. *Health Affairs* 32(3):552–561.

Kim, B. 2015. *Interactive and interpretable machine learning models for human machine collaboration*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Klingler, C.; in der Schmitten, J.; and Marckmann, G. 2016. Does facilitated advance care planning reduce the costs of care near the end of life? systematic review and ethical considerations. *Palliative medicine* 30(5):423–433.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1675–1684. ACM.

Levy, W. C.; Mozaffarian, D.; Linker, D. T.; Sutradhar, S. C.; Anker, S. D.; Cropp, A. B.; Anand, I.; Maggioni, A.; et al. 2006. The seattle heart failure model. *Circulation* 113(11):1424–1433.

Lipton, Z. C. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Lou, Y.; Caruana, R.; and Gehrke, J. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 150–158. ACM.

Luxton, D. D.; Kayl, R. A.; and Mishkind, M. C. 2012. mhealth data security: The need for hipaa-compliant standardization. *Telemedicine and e-Health* 18(4):284–288.

Makar, M.; Ghassemi, M.; Cutler, D. M.; and Obermeyer, Z. 2015. Short-term mortality prediction for elderly patients using medicare claims data. *International journal of machine learning and computing* 5(3):192.

Marik, P. E. 2015. The cost of inappropriate care at the end of life: implications for an aging population. *American Journal of Hospice and Palliative Medicine* 32(7):703–708.

McGorty, E. K., and Bornstein, B. H. 2003. Barriers to physicians decisions to discuss hospice: insights gained from the united states hospice model. *Journal of evaluation in clinical practice* 9(3):363–372.

Medicare. 2017. Medicare coverage: Hospice & respite care. https://www.medicare.gov/coverage/hospice-and-respite-care.html. Accessed: 2017-11-20.

Morrison, R. S.; Penrod, J. D.; Cassel, J. B.; Caust-Ellenbogen, M.; Litke, A.; Spragens, L.; and Meier, D. E. 2008. Cost savings associated with us hospital palliative care consultation programs. *Archives of internal medicine* 168(16):1783–1790.

Moss, A. H.; Ganjoo, J.; Sharma, S.; Gansor, J.; Senft, S.; Weaner, B.; Dalton, C.; et al. 2008. Utility of the surprise question to identify dialysis patients with high mortality. *Clinical Journal of the American Society of Nephrology* 3(5):1379–1384.

NHPCO. 2016. Nhpco: Facts and figures: Hospice care in america. *Alexandria, VA: National Hospice and Palliative Care Organization*.

Pollack, M. M.; Ruttimann, U. E.; and Getson, P. R. 1988. Pediatric risk of mortality (prism) score. *Critical care medicine* 16(11):1110–1116.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1(1):81–106.

Raghupathi, W., and Raghupathi, V. 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems* 2(1):3.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Setoguchi, S.; Stevenson, L. W.; and Schneeweiss, S. 2007. Repeated hospitalizations predict mortality in the community population with heart failure. *American heart journal* 154(2):260–266.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28(1):11–21.

Teno, J. M.; Shu, J. E.; Casarett, D.; Spence, C.; Rhodes, R.; and Connor, S. 2007. Timing of referral to hospice and quality of care: length of stay and bereaved family members' perceptions of the timing of hospice referral. *Journal of pain and symptom management* 34(2):120–125.

Teno, J. M.; Plotzke, M.; Gozalo, P.; and Mor, V. 2014. A national study of live discharges from hospice. *Journal of palliative medicine* 17(10):1121–1127.

Turner, R. 2016. A model explanation system. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on*, 1–6. IEEE.

USDHHS. 2015. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule.

Wright, A. A.; Zhang, B.; Ray, A.; Mack, J. W.; Trice, E.; Balboni, T.; Mitchell, S. L.; et al. 2008. Associations between end-of-life discussions, patient mental health, medical care near death, and caregiver bereavement adjustment. *Jama* 300(14):1665–1673.

Yang, H.; Rudin, C.; and Seltzer, M. 2016. Scalable bayesian rule lists. *arXiv preprint arXiv:1602.08610*.

Yourman, L. C.; Lee, S. J.; Schonberg, M. A.; Widera, E. W.; and Smith, A. K. 2012. Prognostic indices for older adults: a systematic review. *Jama* 307(2):182–192.

Zhang, B.; Wright, A. A.; Huskamp, H. A.; Nilsson, M. E.; Maciejewski, M. L.; Earle, C. C.; Block, S. D.; Maciejewski, P. K.; and Prigerson, H. G. 2009. Health care costs in the last week of life: associations with end-of-life conversations. *Archives of internal medicine* 169(5):480–488.