A hybrid model using LSTM and decision tree for mortality prediction and its application in provider performance evaluation

Peichang Shi
Customer Value Partners
Columiba, MD, USA
pshi1@umbc.edu

Aryya Gangopadhyay
Dept. Informaiton Systems
University of Maryland, Baltimore County
gangopad@umbc.edu

Carolyn Owens
Customer Value Partners
Towson, MD, USA
carolynowens@cvpcorp.com

Brenda Blunt
Customer Value Partners
Towson, MD, USA
brendablunt@cvpcorp.com

Christine Grogan
Customer Value Partners
Towson, MD, USA
christinegrogan@cvpcorp.com

***Abstract*—The risk adjusted mortality rate, which is also called standardized mortality ratio (SMR), is one widely used quality measure to evaluate healthcare provider performance. Logistic regression and decision tree are two traditional risk models for mortality rate calculation. Though some machine learning based approaches could achieve higher accuracy, they are hard to interpret and may have poor calibration scores. In this paper, we evaluated multiple machine learning approaches with different formats of longitudinal data, and proposed a hybrid approach based on long short-term memory (LSTM) model and decision tree. The new hybrid method provides a comparable area under the receiver operating characteristic curve (AUC) performance as LSTM with a better calibration score. Using a set of 3,473 patients with 10 months of data from historical, large scale ESRD patient data, the LSTM with long format data approach achieved AUC for prediction of mortality of 0.772 compared to 0.758 for logistic regression and 0.726 for a decision tree model. The hybrid approach could reach 0.783, a little higher than both LSTM and decision tree model. The hybrid approach has the best calibration performance based on the Hosmer Lemeshow test.**

***Index Terms*—Mortality prediction, LSTM, decision tree, ESRD**

I. Introduction

Healthcare performance evaluation is one critical step for the United States to monitor provider performance and improve value of care through reducing the number of treatments and using lower cost interventions where appropriate. Standardized mortality ratio (SMR) is probably one of the most widely adopted quality measures, which is a ratio between the observed numbers of deaths in the study population and the number of deaths which would be expected [18], [19]. For example, the Hospital Value-Based Purchasing (HVBP) program was introduced by the Centers for Medicare and Medicaid Services (CMS) in 2011 to reward or penalize hospital based on their performance [1]. One important measure for the HVBP program is 30-day mortality for acute myocardial infarction, pneumonia, and heart failure. Mise reported that 90-day postoperative mortality is a legitimate measure of hepatopancreatobiliary surgical quality [2].Risk-adjusted mortality has also been proposed as a quality of care indicator to gauge cardiovascular intensive care Unit (CICU) performance [2], [3]. There is no universal definition regarding short term or long term mortality rate measure. 30-day, 90-day and one-year mortality are the most common mortality periods reported in the literature. During the mortality measure development, the key part is the risk adjustment model development since the predicted mortality is the base for the expected rate in the SMR calculation. Invalid estimation of the expected rate could have serious consequences related to financial and reputations of the healthcare providers due to the association with performance comparison.

II. Related studies

Mortality prediction is not a new topic in healthcare. For cross sectional mortality data, Kuo etc. tested several machine learning methods to predict the mortality of hospitalized motorcycle riders [7]. They found logistic regression (LR) and support vector machine(SVM) models had a significantly higher area under the receiver operating characteristic curve (AUC) than decision tree models. No significant difference was observed in the AUC of LR and SVM. Karhade developed an algorithm to predict 90-day and one-year mortality in spinal metastaic disease [4]. Pedersen also did analysis on 30-day , 90-day and one-year mortality after emergency colonic surgery [5]. Karhade compared five machine learning algorithms for prediction of mortality in spinal epidural(stochastic gradient boosting model, elastic-net penalized logistic regression, random forest, neural network and SVM), and found the stochastic gradient boosting model achieved the best performance across discrimination. Tayler's group compared the accuracy of different techniques (random forest, gradient boosted machine, bagged trees, SVM, neural network,

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

logistic regression, k-nearest neighbors, decision tree) for the composite of ward cardiac arrest, ward to ICU transfer, or death on the wards without attempted resuscitation [8]. They identified that random forest had the best performance regarding AUC.

However those models focus on cross sectional data and lack options for models with longitudinal data. For longitudinal data, one primary traditional approach is cox regression model [20]–[22].However in cox regression, the c statistics or AUC is different from the traditional AUC in other data mining methods. In cox regression, the AUC is applied to time to event data [14], and is non comparable to the conventional c statistics, so this method is not covered in this paper. Though hierarchical logistic regression and mixed effect regression tree could be used for longitudinal data, Hannan reported that standard logistic regression performed similarly to hierarchical models [32], where hierarchical logistic regression is mostly used to account for correlation between patients within the same hospitals.No literature was found for mortality analysis using mixed effect regression tree.

Recently,there has been an increasing interest in deep learning models for mortality prediction with longitudinal healthcare data. Interest in deep learning for healthcare has grown for two reasons. First, for healthcare researchers, deep learning models yield better performance in many tasks than traditional machine learning methods and require less manual feature engineering. Second, large and complex datasets are available in healthcare and enable training of complex deep learning models.

Several papers applied recurrent neural networks with LSTM to electric health records [9], [11]. When modeling longitudinal EHR data, LSTM was used to establish relationships between mixed effect observations and future events.Rajkomar used 114,003 patient records from University of California, San Francisco (UCSF), from 2012 to 2016, and the University of Chicago Medicine (UCM) from 2009 to 2016 for prediction tasks [9]. They tried three deep learning models: one based on recurrent neural networks, one on an attention-based time-aware neural network model, and one on a neural network with boosted time-based decision stumps. They discovered that deep learning methods were capable of accurately predicting multiple medical events (e.g., the prediction of in-hospital mortality, readmission, length of stay, and discharge diagnoses) from multiple centers. Jo combined LSTM and latent topic modeling for mortality prediction using MIMIC-iii data set and showed their model significantly outperformed prior models [10].

Different from the above approaches, our contribution is:

1) We explored an alternative way to use traditional models (logistic regression and decision tree) for longitudinal data by converting the longitudinal data into wide format. This simple transformation proved to have similar performance compared to hierarchical logistic regression and mixed effect regression tree, which models have options to handle longitudinal data

2) We compared models with one time point and multiple time points. Through this, we showed that prior clinical results could affect future outcome and should be included in the model.

3) We proposed a hybrid approach by combing the LSTM and decision tree model, which not only provide interpretability , but also improved the calibration analysis score.

4) We provided a case study for provider performance evaluation using the hybrid approach. This hybrid approach could provide more accurate predictions than traditional hierarchical logistic regression and information about whether the provider had better performance in each subgroup.

III. ESRD Data

The End Stage Renal Disease (ESRD) Population Public Use File (PUF) contains the current and historical ESRD patient population and associated clinical data for both Medicare and non-Medicare patients. This deidentified patient-level dataset is the first large scale public data from the CMS ESRD Program.

The ESRD Population PUF includes ESRD patient data from 1973 to present and ESRD patient clinical data from 2012 to present. Patient and clinical data associated with transient admissions and acute discharges are excluded from this PUF. The ESRD patient population is selected and extracted from the CMS Renal Management Information System (REMIS), which contains all ESRD dialysis and transplant patients. The CMS Consolidated Renal Operations in Web-Enabled Network (CROWNWeb) system is the source of clinical data/measures extracts and supporting patient data. Together, REMIS and CROWNWeb comprise the mandated System of Record (SOR) for all ESRD Patients in the U.S. including both Medicare-entitled and non-Medicare. The current dataset has about 3.2 million patients, with data on admissions, hospitalizations, chronic conditions and 13 monthly clinical measures.

The challenges of this dataset are:

- A mixture of static and dynamic variable information: The static variables include: gender, 20 chronic conditions, and age. The dynamic variables include: clinical testing data, such as clinical hemodialysis adequacy and clinical hemodialysis infection.
- Missing values: Patients may have irregular clinical data, which means one patient may have 10 monthly clinical data, another patient could have 12 monthly data. For one patient, he/she may have five monthly serum albumin data, but have 8 monthly calcium data.

IV. Data selection

Since this is a deidentified dataset, there are no patient visiting dates, the date the patient entered into the ESRD Program is deemed as the reference date and all clinical collection and reporting dates are removed and reported as study Days. The study date is calculated as the number of days elapsed from the patient's deference date so the passage of time between clinical events and the sequenced order of clinical events is preserved.

For a pilot study, we randomly chose 3,473 patients from 3.2 million patients. We first chose a baseline date (for example,

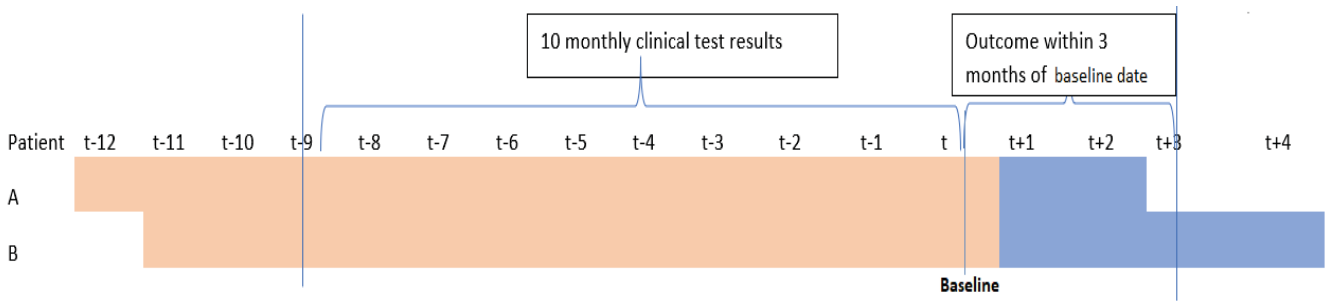

Fig. 1. Patient selection processing

TABLE I
SURVIVAL TIME OF SELECTED PATIENTS

Survival months after baseline date	number of patients	Percentage
1	51	1.47
2	14	0.40
3	676	19.46
4	1499	43.16
5	261	7.52
6	60	1.73
7	16	0.46
8	8	0.23
9	3	0.09
10	1	0.03
12	1	0.03
15	1	0.03
16	7	0.20
>16	875	25.19

400 days after entering the programs for all patients), then we took all 9 monthly data points before the baseline date, and the 12 monthly data points after the baseline date to check the mortality status(see Fig. 1). The outcome variable is whether the patient was still alive within the 3 months following the baseline date. For example, in Fig. 1, let t equal the baseline date. We took 9 monthly clinical data before t, and checked the outcome after t. If patient A died at t+2, and patient B was still alive at t+3, patient A would be coded as death(1),and patient B coded alive (0).The survival months of those selected patients are listed in Table I.

The basic demographic information and clinical data information at baseline date are described in Tables II and III.

V. METHODS

A. Variable selection

Over fitting is one common issue in regression analysis, which is normally caused by extraneous predictors in the model.When this occurs, the coefficients may have inflated magnitude, and then the R square will be large too. To reduce this effect, feature selection becomes very important. We applied Lasso selection [23]. The Lasso selection will impose a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero. Variables with coefficients equal to zero were excluded from the model.

B. Hierarchical logistic regression

Logistic regression is one simple but powerful method, especially for binary outcome. One key component is the logistic function, which could convert the multi variate input into the probability of the outcome between 0 and 1. Among all the machine learning algorithms, logistic regression has multiple advantages. Firstly, no assumption is made regarding the dependent variables following a normal distribution. There is also no assumption about a linear relationship between outcome and covariates. Most importantly however, it is easy to understand and interpret the results [16].

The hierarchical logistic regression is designed for studying data with group structure and a binary response [12] ,which believes that subjects within the group are correlated more than the other subjects within other groups. The groups may be different hospitals, or even different time points for the same subject.The correlation matrix within the group could be exchangeable, independent or auto correlation depending on the data nature. The hierarchical logistic regression is implemented in R geepack in this study.

C. Mixed effect regression tree

This is a methodology that combines the structure of mixed effects models for longitudinal and clustered data with the flexibility of tree-based estimation methods [15], [17]. This approach accounts for the possibilities of systematic differences or auto correlation within objects across time periods. This method is less sensitive to parametric assumptions and was implemented using R REEMtree package.

D. Long Short-Term Memory(LSTM)

LSTM is one special type of recurrent neural network, which works well for long term time series data. LSTM can capture long range dependencies and nonlinear dynamics. Mortality prediction based on electronic health record (EHR data) has been widely reported [24]–[28].We used keras in Python 3.7 for our LSTM model analysis.

VI. EXPERIMENT DESIGN

A. Data with wide format vs data with long format

We explored the option to reshape the data from long format to wide format. One advantage of wide format data is that it can be analyzed using the traditional approach. In the wide

TABLE II
DEMOGRAPHIC DESCRIPTION OF THE BASELINE DATA

		Frequency	%
Age			
	18 - 44	5	0.14
	45 - 54	84	2.42
	55 - 64	694	19.98
	65 - 74	1512	43.54
	75 - 84	975	28.07
	85+	203	5.85
Gender			
	Female	1469	42.3
	Male	2004	57.7
Primary cause of renal failure			
	AIDS nephropathy	8	0.23
	Acquired obstructive uropathy	21	0.62
	Cholesterol emboli, renal emboli	4	0.12
	Chronic interstitial nephritis	6	0.18
	Complications of transplanted kidney	91	2.67
	Diabetes with renal manifestations Type 1	79	2.32
	Diabetes with renal manifestations Type 2	1404	41.19
	Etiology uncertain	56	1.64
	Focal Glomerulonephritis, focal sclerosing GN	27	0.79
	Glomerulonephritis (GN) (histologically not examined)	33	0.97
	Hepatorenal syndrome	12	0.35
	Hypertension: Unspecified with renal failure	632	18.54
	Hypertensive chronic kidney disease with stage 1	188	5.51
	IgA nephropathy, Berger's disease	8	0.23
	Lupus erythematosus, (SLE nephritis)	16	0.47
	Membranous nephropathy	10	0.29
	Multiple myeloma	14	0.41
	Nephropathy caused by other agents	5	0.15
	Other Primary Diag	416	12.17
	Other renal disorders	25	0.73
	Polycystic kidneys, adult type (dominant)	33	0.97
	Renal artery stenosis	12	0.35
	Tubular necrosis (no recovery)	35	1.03
	Type 1 diabetes mellitus with diabetic chronic kidney disease	23	0.67
Primary cause of death			
	Atherosclerotic heart disease	29	0.87
	Cachexia/failure to thrive	42	1.26
	Cardiac arrest, cause unknown	1115	33.42
	Cardiac arrhythmia	55	1.65
	Cardiomyopathy	25	0.75
	Cerebrovascular accident including intracranial hemorrhage	44	1.32
	Congestive Heart Failure	54	1.62
	Gastro-intestinal hemorrhage	22	0.66
	Hyperkalemia	8	0.24
	Ischemic brain damage/Anoxic encephalopathy	12	0.36
	Malignant disease (not 82)	49	1.47
	Malignant disease, patient ever on Immunosuppressive therapy	12	0.36
	Myocardial infarction, acute	97	2.91
	Other Primary Causes	172	9.1
	Other cause of death	292	4.65
	Pulmonary edema due to exogenous fluid	11	0.33
	Pulmonary infection (pneumonia, influenza)	43	1.29
	Septicemia due to peripheral vascular disease, gangrene	20	0.6
	Septicemia, other	143	4.29
	Unknown	652	19.54
	Withdrawal from dialysis/uremia	576	17.27
Comorbid condition			
	congestive heart failure	318	7.83
	ischemic heart disease, CAD	211	4.7
	myocardial infarction	217	4.88
	cardiac arrest	118	1.99
	cardiac dysrhythmia	127	2.25
	pericarditis	1294	36.34
	cerebrovascular disease, CVA, TIA	73	0.67
	peripheral vascular disease	555	14.75
	history of hypertension	230	5.26
	diabetes (primary or contributing)	140	2.63
	diabetes, currently on insulin	116	1.93
	chronic obstructive pulmonary disease	129	2.31
	tobacco use (current smoker)	108	1.69
	malignant neoplasm, Cancer	124	2.16
	alcohol dependence	53	0.09
	drug dependence	63	0.38
	HIV positive status	54	0.12
	AIDS	93	1.26
	inability to ambulate	71	0.61
	inability to transfer	164	3.33

2776

TABLE III
CLINICAL DATA OF BASELINE DATA

Clinical variables	Mean	Std.
Kt/V Hemodialysis	1.286	0.668
Hemodialysis BUN pre-dialysis run	42.376	26.325
Hemodialysis BUN post-dialysis run	11.310	7.891
Weight before Hemodialysis	63.210	35.946
Weight after Hemodialysis	61.430	34.881
Minutes taken to complete hemodialysis	173.790	89.653
Normalized Protein Catabolic (nPCR) rate	0.590	0.460
Iron Saturation (TSAT) Percentage	14.990	17.105
Serum Phosphorus	4.200	2.287
Corrected Serum Calcium	7.690	3.588
Serum Albumin	2.900	1.432
ESA Monthly dose	20.110	36.165

pat_id	Month	age	hd_ktv
10928	1	65	1.1
10928	2	65	1.22
10928	3	65	1.38
10928	..	..	..
20928	1	60	1.21
20928	2	60	1.31

Fig. 2. Data with long format

format, the patient's repeated responses will be in a single row (see Fig. 3), while in the long format, each row is one time point per subject, which is the default structure of ESRD data (see Fig. 2). The wide format for repeated measures could be seen from the repeated ANOVA analysis.However in repeated ANONA, it assumes the covariance structure among the time points for the same subject is compound symmetric.

To evaluate the feasibility of wide format data for repeated measures, we compared 1) logistic regression with wide format data vs hierarchical logistic regression with long format data; 2) decision tree with wide format data vs mixed effect regression tree with long format data. The hierarchical logistic regression and mixed effect regression tree models have the option to handle longitudinal data. We utilized the repeated subsampling approach ten times, and each time used 80% of the data as training data, and 20% of the data as testing data.

B. LSTM with wide format vs LSTM with long format

We tried two types of LSTM. One LSTM model is based on wide format data. For this LSTM, we treat each variable value as one word, each additional time point values will be added to the end. To do that, we need to transform the raw data structure to wide format so that one patient has only one row (see Fig. 5). In the model we will treat each value as a

pat_id	age	hd_ktv_1	hd_ktv_2	hd_ktv_3
10928	65	1.1	1.22	1.38
20928	60	0	1.21	1.31

Fig. 3. Data with wide format

LSTM structure with long format data

Layer (type)	Output Shape	Param #
lstm_73 (LSTM)	(None, 100)	72000
dense_73 (Dense)	(None, 1)	101

Total params: 72,101
Trainable params: 72,101
Non-trainable params: 0

None

Fig. 4. Data structure with long format

LSTM structure with wide format data

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 123, 8)	40000000
flatten_2 (Flatten)	(None, 984)	0
dense_2 (Dense)	(None, 1)	985

Total params: 40,000,985
Trainable params: 40,000,985
Non-trainable params: 0

None

Fig. 5. Data structure with wide format

different word. To differentiate different variables so that value 40 in age is different from value 40 in bun variable, we add one fixed value to each variable, thus all variables will have different values. For example, adding 100 to all age variable, and 1000 to sex variable etc.

For LSTM with long format (Fig. 4), we first needed to encode all categorical variables into dummy variables. Then we simply use the formula y=(x-min)/(max-min) to normalize all the variables.

The two LSTM models include 100 LSTM units in the hidden layer, however he structure for LSTM with wide format is apparently more complicated than LSTM with long format. The former has 40 million parameters, while the latter only has 72 thousand.

C. Single time point vs multiple time points

We tried to identify whether previous clinical time points have effect on model prediction performance. To do that, we set up 10 group trials. We started from one time point data(t), then added successive time points until we had 10 time points. For each time point, we ran 4 models We evaluated the model performance based on AUC.

D. AutoML approach

AutoML is one automatic machine learning which includes automatic training and tuning of many models. It is automatically trained and tuned based on collections of models. One of the frameworks is H2O, which supports the most widely used statistical and machine learning algorithms, including gradient boosted machines, generalized linear models, deep learning models, and more. During this procedure, very few parameters are needed from the user.

In our case, we only provided the number of models and time limit for the models, then the AutoML chose the best

TABLE IV
EXAMPLE OUTPUT FROM AUTOML

model_id	AUC
StackedEnsemble_BestOfFamily_0	0.760
StackedEnsemble_AllModels_0	0.760
GLM_grid_0_model_0	0.756
GBM_grid_0_model_0	0.746
GBM_grid_0_model_1	0.737
DRF_0	0.728
XRT_0	0.717

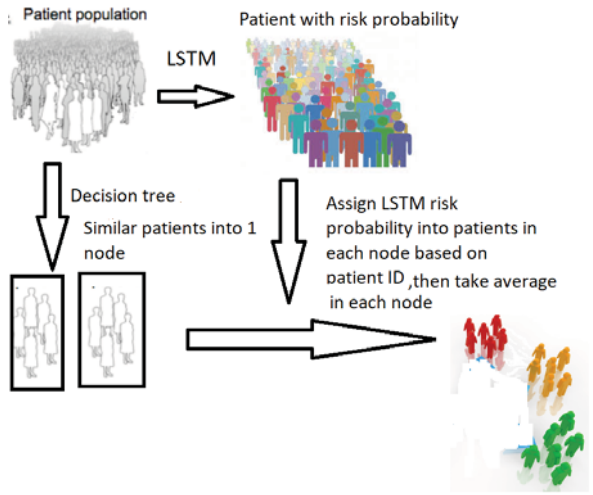

Fig. 6. Hybrid approach flowchart

performers from the models. Our analysis is based on H2O AutoML package in Python. We limited the number of models to five.

In Table IV, the H2O will list the model names and also the performance of the model, then AutoML will choose the top model performer as the final model.

E. Hybrid approach

Compared to decision tree model, LSTM has better prediction performance, but it is hard to explain and acts like a black box. Also, the calibration analysis of LSTM showed poor performance. The decision tree is easy to interpret but has poor performance. We explored the possibility of improving the calibration score through combining decision tree and LSTM models. we first built a decision tree framework, and each patient was then assigned to different nodes. We replaced the patient predicted probability with the probability from LSTM model, then by taking the average of that node for all patients within that node (See Fig. 6 for the flowchart).

VII. EVALUATION

A. Discrimination analysis

Discrimination refers to how well the model differentiates those at higher risk of having an event from those at lower risk.Mean Brier score and Receiver Operating Characteristic (ROC) or AUC curves are used to assess the model performance. The Brier score measures the accuracy of probabilistic predictions as the mean squared difference between the predicted probability assigned to the possible outcome and the actual outcome. The closer the score is to 0, the better the sharpness and calibration of the prediction.

AUC values between 0.7–0.8 were considered to represent moderate predictive validity and $>$0.8 to represent high predictive validity.However, discrimination alone is insufficient to assess a model's prediction capability, since this analysis may fail to discriminate patients in a more homogenous population. For each dataset, we took 10 subsamples for cross validation and took the average of the AUC for each model [29].

B. Calibration analysis

Calibration or goodness of fit is often considered the most important property of a model and reflects the extent to which a model correctly estimates the absolute risk (ie, if the values predicted by the model agree with the observed values). Poorly calibrated models will underestimate or overestimate the outcome of interest.

Hosmer-Lemeshow test is one good measure to evaluate calibration between predicted and observed event rates [31].These predicted probabilities are ordered from lowest to highest and subdivided into number of equally sized groups representing subgroups of risk (In our case, the number is 50). The expected event rate for each group is the average of probabilities of those patients in the group; observed event rate is number of patients in the subgroups.

Discrimination and calibration are both important characteristics in the evaluation of model performance; however, they remain underreported in the published medical literature [30]. A systematic review addressing prediction models of cardiovascular outcomes noted that only 63% reported on discrimination and only 36% reported on calibration [29].

C. Model validation

To avoid overfitting, we used repeated random subsampling for model validation. Each time we used 80% of the data as training dataset, 20% of the data for testing. The performance is compared among different models by taking the average of c statistics or AUC for all testing data tests.

VIII. RESULT AND DISCUSSION

A. Wide format data vs long format data

After comparing the results between traditional logistic regression , decision tree with wide format longitudinal data and hierarchical logistic regression model , mixed effect regression tree with long format longitudinal data, we found that these two format gave very close results (see table V). These results may suggest that we could use wide format data for longitudinal data analysis, which could give us more flexibility to apply more traditional approaches to longitudinal dataset.

B. LSTM with wide format vs LSTM with long format

The results of these two LSTM models in Table VI showed that LSTM with wide format had comparable performance with LSTM with long format. However, due to the excessive

TABLE V
RESULTS OF WIDE FORMAT DATA VS LONG FORMAT DATA

	Mixed effect regression tree	Hierarchical logistic regression	Decision tree	Logistic regression
Subsample 1	0.744	0.769	0.745	0.729
Subsample 2	0.758	0.752	0.754	0.786
Subsample 3	0.751	0.759	0.729	0.771
Subsample 4	0.724	0.757	0.772	0.748
Subsample 5	0.771	0.761	0.753	0.752
Subsample 6	0.736	0.758	0.704	0.737
Subsample 7	0.710	0.758	0.743	0.778
Subsample 8	0.735	0.762	0.739	0.758
Subsample 9	0.748	0.764	0.740	0.742
Subsample 10	0.726	0.759	0.770	0.778
Mean AUC	0.740	0.760	0.745	0.758

TABLE VI
RESULTS OF SINGLE TIME POINT VS MULTIPLE TIME POINTS

# Time points in the model	LSTM (Wide)	LSTM (long)	Decision tree	Logistic regression
10 time points	0.770	0.754	0.711	0.751
9 time points	0.755	0.753	0.710	0.755
8 time points	0.747	0.758	0.708	0.751
7 time points	0.744	0.734	0.737	0.745
6 time points	0.780	0.779	0.741	0.745
5 time points	0.764	0.767	0.705	0.748
4 time points	0.774	0.773	0.697	0.745
3 time points	0.757	0.747	0.710	0.743
2 time points	0.747	0.745	0.699	0.741
1 time points	0.746	0.737	0.700	0.737

number of parameters in the model , LSTM with wide format took much longer time than LSTM model with long format(20 minutes vs 2 minutes). We decided to use LSTM with long format for further analysis.

C. Single time point vs multiple time points

From the table VI , we can see: 1) all models increased the AUC from the single time point to 10 time points; 2) LSTM models were more sensitive to number of time points, and decision tree was the least.
The above findings may suggest that the previous clinical results did have effect on the future outcome, and we should try to include previous data in the model for better performance.

D. AutoML vs LSTM with long format data

The AutoML approach was an easy way to explore more advanced machine learning methods. In this approach we tried five additional algorithms, which provided good results, however no better than LSTM model (see results in Table VII).

E. Hybrid approach vs LSTM

Given the outputs from the above experiments, we realized that LSTM with long format data may provide the best model accuracy, so in the Table VIII, we evaluated our hybrid approach with LSTM with long format data based on both model AUC and Hosmer Lemeshow score.
With combining information from decision tree and LSTM,

TABLE VII
COMPARISON OF MODEL PERFORMANCE BETWEEN LSTM AND AUTOML

	LSTM (long)	AutoML
Subsampel 1	0.740	0.750
Subsampel 2	0.795	0.792
Subsampel 3	0.789	0.779
Subsampel 4	0.773	0.767
Subsampel 5	0.781	0.769
Subsampel 6	0.758	0.757
Subsampel 7	0.759	0.776
Subsampel 8	0.776	0.776
Subsampel 9	0.745	0.749
Subsampel 10	0.799	0.777
Mean	0.772	0.769

TABLE VIII
COMPARISON OF MODEL PERFORMANCE AMONG LSTM ,DECISION TREE AND HYBRID APPROACH

	LSTM (long)	Decision tree	Hybrid approach
Subsampe 1	0.747	0.705	0.756
Subsampe 2	0.799	0.728	0.817
Subsampe 3	0.785	0.718	0.787
Subsampe 4	0.774	0.750	0.774
Subsampe 5	0.781	0.732	0.782
Mean AUC	0.777	0.726	0.783

the hybrid approach achieved better performance than both LSTM and decision tree 0.783 vs 0.777 vs 0.726(see Table VIII).When we evaluated 50 groups for Hosmer test using five subsampling groups, the hybrid approach had much better results. This may be due to the fact that the average of the LSTM probabilities in each node removed some outliers , and smoothed the subgroup rate (see Table IX).

IX. CASE STUDY

The simplest and most useful form of indirect adjustment is the standardized mortality ratio, which is also called indirect method. This procedure is to use the population rate of each risk category to multiply the number of hospital patient numbers in that category to calculate the expected number of hospital patients. The standard ratio is then calculated by the sum of observed hospital event numbers in each category divided by the sum of expected event numbers. Since each hospital may have a different number of patients, the ratio for each hospital is not comparable to other hospital ratios. Therefore,it can only show whether a hospital is better or worse than the reference population.

TABLE IX
HOSMER LEMESHOW TEST AMONG LSTM,DECISION TREE AND HYBRID APPROACH

	LSTM (long)	Hybrid approach
Subsampe 1	0.003	0.118
Subsampe 2	0.001	0.194
Subsampe 3	0.008	0.171
Subsampe 4	0.000	0.257
Subsampe 5	0.012	0.091

TABLE X
EXAMPLE OF INDIRECT STANDARDIZATION MORTALITY RATIO CALCULATION

Age	Age specific mortality rate for US population (Standard population)	Hospital A population	observed number of death for hospital A	Expected number of death for hospital A =Hospital A population * national age specific rate
<20	0.02	1000	25	20
20-44	0.03	2000	55	60
45-64	0.05	3000	145	150
65+	0.3	4000	1000	1200

Suppose we have hypothetical data in table X. The SMR formula is:

$$SMR = \frac{observednumberofdeaths}{expectednumberofdeaths} = \frac{25+55+145+1000}{20+60+150+1200} = \frac{1225}{1430} = 0.86 \quad (1)$$

When we applied the risk adjustment model to the SMR calculation, the expected number of deaths would be the sum of predicted probability for each patient.Currently logistic regression is the gold standard for SMR calculation which appears in both the Agency for Healthcare Research and Quality (AHRQ) and CMS programs.

Compared to the traditional approach for SMR, the hybrid approach has the following advantages:

A. Higher AUC and calibration score

From Table IX , apparently the hybrid approach has a high accurate prediction and calibration score. Though LSTM alone could provide comparable model accuracy, the calibration test suggests this may not be a good model for SMR calculation.

B. Additional valuable information

The hybrid approach could also provide extra valuable information regarding the performance of the provider in each subgroup, which thus possibly enables the provider to identify their weakness and help them improve healthcare quality. The traditional logistic regression only provided an overall SMR, which means whether the provider has a better performance (if SMR <1) or has worse performance if (SMR > 1). With the new hybrid approach, each node is a subgroup, hence we can check whether the provider has similar performance in each sub group.

Suppose Fig 7 is one example node in our hybrid approach. From the figure, the two nodes are based on whether the patient has a recoded diagnosis number 16508, which means diabetes. If yes, then the patient will go to leftmost node, otherwise, the patient will go to rightmost node. We can see that in leftmost node, that subgroup has total 1,594 patients

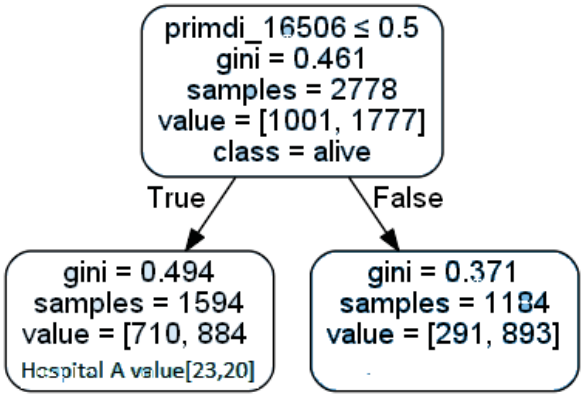

Fig. 7. Tree node

TABLE XI
CASES TO SHOW PERFORMANCE OF ONE PROVIDER IN EACH NOE

Subgroup	Categories	Hospital A vs Subgroup p value
1	Patients with diabetes, and hd KTV <0.8	0.03
2	Patients with more than 1000 hd delivered minutes	N/A
3	Patients with chronic kidney and tubular necrosis, and albumin <2	0.25

with 714 deaths. For hospital A, it has total 43 patients with 23 deaths. With the Fisher exact test, the p value 0.28 suggest that hospital A is not significantly different from the population in this subgroup. Similar to this example,we can check all the subgroups to identify which area hospital A is good at. Table XI shows how it worked. For hospital A, we could see that it does not have patients in category 2, and has better performance in category 1, but no different in category 3. Given this fact, the hospital A may focus on how to improve the healthcare quality in subgroup 1.

X. CONCLUSION

While our hybrid LSTM produces promising results, this is our initial exploration based on first public large scale ESRD data file. As a pilot study, we only randomly chose a small amount of samples from millions of records. On the methodological side, we explored different approaches for the mortality analysis including the deep learning techniques. We also proposed a hybrid algorithm for a better calibration score,and tried to apply this to provider performance evaluation. In future work, we plan to utilize the whole data set. Additionally,we would like to include developing LSTM architectures to handle missing values, which are common issues in the current data set. Beside mortality, we would also like to apply the reinforcement learning for an ESRD treatment policy study to identify risk factors and higher value intervention pathways.

REFERENCES

[1] Figueroa, Jose F., et al. "Association between the Value-Based Purchasing pay for performance program and patient mortality in US hospitals: observational study." bmj 353 (2016): i2214.

[2] Mise, Yoshihiro, et al. "90-day postoperative mortality is a legitimate measure of hepatopancreatobiliary surgical quality." Annals of surgery 262.6 (2015): 1071.

[3] Goldfarb, Michael. "Risk-adjusted overall mortality as a quality measure in the cardiovascular intensive care unit." Cardiology in review 26.6 (2018): 302-306..

[4] Karhade, Aditya V., et al. "Predicting 90-day and 1-year mortality in spinal metastatic disease: development and internal validation." Neurosurgery (2019).

[5] Pedersen, T., et al. "30-Day, 90-day and 1-year mortality after emergency colonic surgery." European Journal of Trauma and Emergency Surgery 43.3 (2017): 299-305.

[6] Taylor, R. Andrew, et al. "Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach." Academic emergency medicine 23.3 (2016): 269-278.

[7] Kuo, Pao-Jen, et al. "Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: a cross-sectional retrospective study in southern Taiwan." BMJ open 8.1 (2018): e018252.

[8] Taylor, R. Andrew, et al. "Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach." Academic emergency medicine 23.3 (2016): 269-278.

[9] Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. npj Digital Med. 2018;1(18):1-10

[10] Jo, Yohan, Lisa Lee, and Shruti Palaskar. "Combining LSTM and latent topic modeling for mortality prediction." arXiv preprint arXiv:1709.02842 (2017).

[11] Makar, Maggie, et al. "Short-term mortality prediction for elderly patients using Medicare claims data." International journal of machine learning and computing 5.3 (2015): 192

[12] Hannan, Edward L., et al. "Predicting risk-adjusted mortality for CABG surgery: logistic versus hierarchical logistic models." Medical care (2005): 726-735.

[13] .Akins, Cary W., et al. "Guidelines for reporting mortality and morbidity after cardiac valve interventions." Eurobpean Journabl of Cardio-Thoracic Surgery 33.4 (2008): 523-528.

[14] Guo, Changbin, S. Yo, and Woosung Jang. "Evaluating predictive accuracy of survival models with PROC PHREG." SAS, 2018.

[15] Sanagou, Masoumeh, et al. "Hospital-level associations with 30-day patient mortality after cardiac surgery: a tutorial on the application and interpretation of marginal and multilevel logistic regression." BMC medical research methodology 12.1 (2012): 28.

[16] Austi, Peter C., and David A. Alte. "Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: should we be analyzing cardiovascular outcomes data differently?." American heart journal 145.1 (2003): 27-35.

[17] Sela, Rebecca J., and Jeffrey S. Simonoff. "RE-EM trees: a data mining approach for longitudinal and clustered data." Machine learning 86.2 (2012): 169-207.

[18] van Gestel, Yvette RBM, et al. "The hospital standardized mortality ratio fallacy: a narrative review." Medical care (2012): 662-667.

[19] Pouw, Maurice E., et al. "Hospital standardized mortality ratio: consequences of adjusting hospital mortality with indirect standardization." PloS one 8.4 (2013): e59160.

[20] Chertow, G. M., et al. "Mortality after acute renal failure: models for prognostic stratification and risk adjustment." Kidney international 70.6 (2006): 1120-1126.

[21] Fenton, Stanley SA, et al. "Hemodialysis versus peritoneal dialysis: a comparison of adjusted mortality rates." American Journal of Kidney Diseases 30.3 (1997): 334-342.

[22] Kulkarni, Girish S., et al. "Longer wait times increase overall mortality in patients with bladder cancer." The Journal of urology 182.4 (2009): 1318-1324.

[23] Tibshirani, Robert. "Regression shrinkage and selection via the lasso: a retrospective." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.3 (2011): 273-282.

[24] Maragatham, G., and Shobana Devi. "LSTM Model for Prediction of Heart Failure in Big Data." Journal of medical systems 43.5 (2019): 111.

[25] Esteva, Andre, et al. "A guide to deep learning in healthcare." Nature medicine 25.1 (2019): 24.

[26] Miotto, Riccardo, et al. "Deep learning for healthcare: review, opportunities and challenges." Briefings in bioinformatics 19.6 (2017): 1236-1246.

[27] Choi, Edward, et al. "Doctor ai: Predicting clinical events via recurrent neural networks." Machine Learning for Healthcare Conference. 2016.

[28] Min, Xu, Bin Yu, and Fei Wang. "Predictive Modeling of the Hospital Readmission Risk from Patients' Claims Data Using Machine Learning: A Case Study on COPD." Scientific reports 9.1 (2019): 2362.

[29] Alba, Ana Carolina, et al. "Discrimination and calibration of clinical prediction models: users' guides to the medical literature." Jama 318.14 (2017): 1377-1384.

[30] Cook, Nancy R. "Use and misuse of the receiver operating characteristic curve in risk prediction." Circulation 115.7 (2007): 928-935.

[31] Schmid, Christopher H., and John L. Griffith. "Multivariate classification rules: calibration and discrimination." Wiley StatsRef: Statistics Reference Online (2014).

[32] Hannan, Edward L., et al. "Predicting risk-adjusted mortality for CABG surgery: logistic versus hierarchical logistic models." Medical care (2005): 726-735.