Deep Learning vs Traditional Models for Predicting Hospital Readmission among Patients with Diabetes

Ameen A. Hai, MSc[1] Mark G. Weiner, MD[2], Anuradha Paranjape, MD, MPH[3], Alice Livshits, BS, MFA[3], Jeremiah R. Brown, PhD, MS[4], Zoran Obradovic, PhD[1], and Daniel J. Rubin, MD, MSc [3]

[1]Center for Data Analytics and Biomedical Informatics, Philadelphia, PA; [2]Weill Cornell Medicine, New York, NY; [3]Lewis Katz School of Medicine at Temple University, Philadelphia, PA; [4]Departments of Epidemiology and Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, NH.

Abstract

A hospital readmission risk prediction tool for patients with diabetes based on electronic health record (EHR) data is needed. The optimal modeling approach, however, is unclear. In 2,836,569 encounters of 36,641 diabetes patients, deep learning (DL) long short-term memory (LSTM) models predicting unplanned, all-cause, 30-day readmission were developed and compared to several traditional models. Models used EHR data defined by a Common Data Model. The LSTM model Area Under the Receiver Operating Characteristic Curve (AUROC) was significantly greater than that of the next best traditional model [LSTM 0.79 vs Random Forest (RF) 0.72, p<0.0001]. Experiments showed that performance of the LSTM models increased as prior encounter number increased up to 30 encounters. An LSTM model with 16 selected laboratory tests yielded equivalent performance to a model with all 981 laboratory tests. This new DL model may provide the basis for a more useful readmission risk prediction tool for diabetes patients.

Introduction

Hospital readmission is an undesirable, costly outcome for both patients and hospitals.[1] Patients with diabetes are at higher risk of readmission within 30 days of hospital discharge (30-day readmission) than patients without diabetes.[2-4] Of the nearly 9 million discharges of diabetes patients annually in the US,[5] almost 2 million are 30-day readmissions, corresponding to at least $20 billion in hospital costs.[6, 7] Identifying higher risk patients with diabetes would enable the targeting of interventions to those at greatest need, optimizing the cost-benefit ratio.

We previously published on the development and validation of the Diabetes Early Readmission Risk Indicator (DERRI™), a logistic regression (LR) model that predicts the risk of all-cause 30-day readmission among patients with diabetes.[8] The DERRI™ was designed for use at the point of care based on user input of 10 factors. In split-sample internal validation, performance was modest (Area Under the Receiver Operating Characteristic Curve, AUROC 0.69). In external validation studies, the DERRI™ AUROC was 0.63 and 0.80.[9, 10] In addition to variable predictive performance, application of the DERRI™ requires manual data collection and entry, which are major barriers to its use in clinical practice.

In other published work, we showed that adding variables to the DERRI™ substantially improves predictive accuracy to an AUROC of 0.82.[11] This expanded model (DERRIplus), however, is not feasible for use at the point of care and included employment status, which is not routinely documented in Electronic Health Records (EHRs). Therefore, this model cannot be directly translated to an automated, EHR-integrated tool. There is an unmet need for a readmission risk prediction tool for patients with diabetes that is both accurate and easy to use.

Over the past few years, multiple machine learning (ML) models for predicting 30-day readmission risk of diabetes patients have been published. Several traditional ML modeling approaches have been explored, including random forest (RF), k-nearest nearest neighbor, naïve Bayes, support vector machine (SVM), AdaBoost, and multilayer perceptron (MLP), with a wide range of performance (AUROC 0.53-0.99, accuracy 0.54-0.99).[12-22] Deep learning (DL) models have also been developed for predicting readmission risk of diabetes patients, also with variable performance (AUROC 0.61-0.97, accuracy 0.69-0.95), none of which exceeded that of the best traditional ML models.[23-27] Two of these studies demonstrated a clear advantage of DL approaches over traditional ML models,[23, 24] and two studies found marginal benefit with DL approaches.[25, 27] Comparisons of model performance across all these studies, however, is limited by the lack of standardized reporting of performance characteristics and variable approaches to testing. Therefore, it remains unclear if DL models outperform traditional ML models at predicting readmission risk for patients with diabetes.

Interestingly, all these prior models were developed on the same dataset,[28] except for the DERRI™ and DERRIplus. This publicly available dataset contains hospital encounters with a diagnosis of diabetes and a length of stay between 1 and 14 days at one of 130 US hospitals between 1999 and 2008. Only 3 International Classification of Diseases, Ninth Revision (ICD-9) diagnostic codes per encounter, and only 2 laboratory values (blood glucose and HbA1c) were recorded. Lastly, there is no distinction made between planned and unplanned readmissions. Thus, even the best of these models may not perform as well in patients today. More current, generalizable models are needed.

Therefore, to address these gaps, the aims of the current study were as follows: 1) To develop DL models for the prediction of unplanned, all-cause 30-day readmission, 2) To compare performance of the DL models to traditional ML models, 3) To explore model performance across a range of prior EHR encounters from 1 to 100 being included in model development, and 4) To compare a DL model developed using a subset of laboratory tests selected by domain knowledge with a DL model developed using all available laboratory tests. All models were developed and tested in a dataset of 2,836,569 encounters of 36,641 patients with diabetes using demographics, vital signs, diagnostic and procedure codes, medications, laboratory tests, and administrative data as defined by the National Patient-Centered Clinical Research Network (PCORnet) Common Data Model (CDM).[29]

Materials and Methods

Definition of patient cohort

Inclusion criteria were patients with at least one discharge from any of the three Temple University Health System hospitals in Philadelphia, PA, between July 1st, 2010, and December 31st, 2020, and diabetes defined by at least one of the following: a diagnosis of diabetes (ICD-9: 249.xx or 250.xx or ICD-10: E08.xxx through E13.xxx); a Hemoglobin A1c (HbA1c) level ≥6.5%, or an order for a diabetes specific medication. Encounters were excluded for patient age <18 years, discharge by transfer to another hospital, inpatient death, a diagnosis of gestational diabetes (ICD-9: 648.0x or ICD-10: O24.4x), a diagnosis of prediabetes (ICD-9: 790.29 or ICD-10: R73.03), or pregnancy (positive beta human chorionic gonadotropin laboratory test within 90 days before or after the encounter).

Patients were sorted into one of 2 groups by readmission status: those who had at least one 30-day readmission and those who did not. Among the patients who had a readmission, one admission-readmission pair was randomly selected for analysis. Among the patients who did not have a readmission, one admission was selected randomly for analysis.

Definition of variables and data preprocessing

Tables were extracted from the CDM for each of the following domains: encounters, demographics, diagnoses, laboratory tests, medication orders, procedures, and vital signs. Because features of a given encounter existed in multiple tables, tables were merged by a unique identifier. Merging extracted tables resulted in a sample containing all records for a given encounter. This resulted in substantial missingness. Thus, missingness was used as a separate feature. For continuous features, missing data were replaced with 0, while categorical features were replaced with a unique category.

A total of 23 features were used as input to the models: 14 were extracted from the CDM and 9 were aggregated. Extracted features were: 1) Encounter type (Inpatient, Emergency Department, Observation Stay, Ambulatory Visit, Other Ambulatory Visit, Telehealth and Other; 2) Discharge Status (Assisted Living Facility, Against Medical Advice, Expired, Home Health, Home/Self Care, Hospice, Nursing Home, Rehabilitation Facility, Skilled Nursing Facility; 3) Sex; 4) Hispanic; 5) Race (American Indian/Alaska Native, Asian, Black, Pacific Islander, White, other/no information); 6) Tobacco (Current user, never user, former user, passive exposure, other/no information); 7) age; 8) Diagnosis Clinical Classification System (CCS) codes;[29] 9) Procedure CCS codes;[29] 10) Laboratory results; 11) Medication orders within 1 year before each encounter; 12) Diastolic blood pressure; 13) Systolic blood pressure; and 14) Body mass index (BMI). Aggregated features were: 1) Elixhauser conditions: a binary feature indicating the presence or absence of each condition;[30] 2) Duration of admission (length of stay in days); 3) number of procedure codes before conversion to CCS code; 4) number of diagnosis codes before conversion to CCS code; 5) number of days since the prior encounter regardless of encounter type; 6) number of days since the prior inpatient, observation or emergency department encounter; 7) number of days since the prior encounter of other (non-hospital) encounter types; 8) number of inpatient, observation and emergency department encounters before the current encounter; and 9) number of other (non-hospital) encounters before the current encounter. ICD-9 codes were converted to ICD-10 codes to unify the code format. ICD-10 codes and procedure codes were converted to CCS codes. Based on domain

knowledge, medications relevant to diabetes were categorized as follows: diabetes medications by class, cholesterol, corticosteroids, renin-angiotensin system (RAAS) blood pressure agents, and non-RAS blood pressure agents. Other medications were ignored. Features found not to be reliable, mostly missing, or correlated were removed. Outliers in features such as dates, results, height, weight, BMI, and blood pressure (systolic and diastolic) were removed by observing the data distributions, percentiles, and domain knowledge. Missing values were treated as another category that indicates that a parameter was not collected in relation to the encounter. The primary outcome for model prediction ($\boldsymbol{y}$) was unplanned, all-cause inpatient readmission within 30 days of an inpatient encounter discharge as defined by the Centers for Medicare & Medicaid Services (CMS).[31] Based on the CMS definition, only the first readmission within 30 days was analyzed.

To prepare the data for machine learning models the following data preprocessing techniques were performed. Categorical features were one hot encoded; continuous and discrete features were normalized using min-max normalization techniques,[32] defined as:

$$\mathrm{x}' = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

There were different numbers of recordings in each encounter for each of the following features. Thus, the following statistical values were computed instead. For diastolic and systolic blood pressures, we calculated minimum, maximum, and mean values. For BMI, minimum, maximum, mean, and coefficient of variance were used. These statistical values where normalized and used as features. Moreover, the number of features differed at encounters due to the different number of laboratory tests, diagnoses, and procedures because an encounter could have multiple diagnosis and/or procedure codes, or none. To remedy this and unify the dimensionality of feature vectors, the following data representation techniques were used to enhance the learning of the models. For diagnosis and procedure codes, we used the representation of one-hot encodings, where each value was set to 0 or 1, indicating whether a diagnosis/procedure code existed or not for each encounter. We modified this data representation technique slightly for laboratory tests because each test had an associated result. Hence, we replaced 1, which indicated a code exists, with the laboratory result. Laboratory results were normalized using Equation 1. Because results were of different units and measures, when normalizing laboratory results, we considered the minimum and maximum for each laboratory code separately. This technique created a high dimensional sparse array due to the many unique codes. Then, we utilized Singular Value Decomposition (SVD) algorithm to learn an embedding and reduced dimensionality. SVD was used since it does not assume a square matrix as an input and better for sparse data.[33] Laboratory tests were reduced to 50 components, procedure codes were reduced to 45 components, and diagnosis codes were reduced to 25 components. Different numbers of components were explored and the sum of variance ratio was observed to determine the optimal number of components to reduce dimensionality. All features were concatenated in a feature vector for each encounter. SVD was applied on each encounter separately to reduce and unify dimensions; dimension of encounters was reduced to 50 features per encounter. Then, we concatenated all encounters for a given patient in a

feature vector ordered sequentially by admission date. The class distribution was 27,511 patients without readmission (negative class) and 9,130 patients who were readmitted (positive class).

Experimental Approaches

We conducted extensive experiments using the EHR data to address the following objectives:

- Predict whether patients with diabetes will be readmitted within 30 days
- Compare the performance of the utilized DL methods with several traditional models
- Analyze how many prior encounters (i.e., historical data) within 2 years is optimal to predict readmission
- Evaluate the effects of incorporating all laboratory tests in the data versus learning from a subset of tests chosen by a domain expert

In this study, DL models take as an input a 3-dimensional tensor p x e x f to represent f features for each of e encounters for p patients. In contrast, in traditional models, data is typically represented as a 2-dimensional matrix, with all features of all encounters corresponding to a single patient concatenated in a long feature vector. The dimensionality of each encounter was reduced and unified to 50 features, hence, in a deep model f is of size 50. In a traditional model feature vector consists of all encounters and therefore is of size e x 50. Patients have different numbers of encounters resulting in a nonuniform dimensions; hence, feature vectors were padded with 0s to achieve a unified form. Data representation used as input for DL and traditional models is illustrated at the left and right panels of Figure 1, respectively.

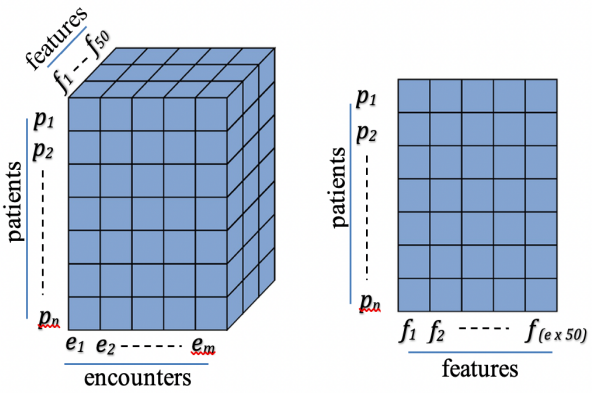

Figure 1. Representation of data input to the deep learning models (left), and traditional models (right).

To model heterogenous sequential data, we developed 2 variants of DL models and compared both versus several traditional models used as baselines. DL models used in our study were: 1) 1-way Long Short-Term Memory (LSTM) networks, which are a variant of Recurrent Neural Network (RNN) that is capable of learning order dependence in sequential data[32]; and 2) Bidirectional Gated Recurrent Unit (GRU), which is another variant of RNN. Traditional models used as baselines were: 1) Random Forest (RF), an ensemble method for classification and regression; during training, it constructs multiple decision trees;[30] RF frequently achieves the state-of-the-art performance in existing literature on predictions using medical data. 2) Multi-layer Perceptron (MLP), a simple neural network model that does not account for temporal information. MLP consists of multiple layers of perceptron and performs backpropagation learning and utilizes a non-linear activation function.[31] 3) Logistic Regression (LR), an interpretable model used frequently in existing literature of readmission predictions and applied on medical data; and 4) AdaBoost, which is less prone to overfitting as its input parameters are not jointly optimized. The DL models were implemented using "Keras" Python libraries, a high-level API of "TensorFlow". "Scikit-learn" library was utilized to implement traditional models in Python.

The architecture of the proposed model, LSTM, comprises 128 neurons, a sequential layer, a reshape layer that was used to reshape the input to 3-dimensional tensor, and a masking layer with a mask value of 0 used to skip the timesteps for which the data were missing. Since padding with 0s was performed to unify dimensions, the masking layer was utilized to avoid any computation with the missing values in all layers following the masking layer, hence, missing values were not accounted for during learning. Additionally, a dropout was added between hidden and output layers. Utilizing this technique to randomly select a given percentage to drop, which is a common regularization technique that assist the model learn general pattern in data.

RNN is a variant of neural networks, which consist of hidden neurons that are capable of analyzing temporal EHR data.[32] RNN comprises of the same structure as the basic neural network, but neurons in the same layer are connected, allowing a neuron to learn from the same neighboring layers, in addition to learning from outputs of the previous

layers and the input data. Thus, RNN neurons include two sources of inputs, the present and the recent past. The process of learning is defined as:

$$b^t = \text{ReLU}\,(b + Wh^{t-1} + Ux^t) \tag{2}$$

$$\bar{y} = \text{sigmoid}\left(b + \sum_t Vb^t\right). \tag{3}$$

To compute value b^t of a hidden neuron, t, a non-linear transformation function, ReLU, is applied to weighted W value of its left hidden neuron b^{t-1} and the weighted U value of its input x^t. Predictions are computed using a sigmoid function of weighted V sum of all hidden neurons with added bias b. The drawback of RNN is that it suffers from the vanishing gradient problem, meaning that weights remain unchanged making it difficult for the model to convergence, hence, the model struggles to learn. To solve this, an LSTM layer was introduced in which sigmoid neurons of RNN are replaced with more complex short-term memory structure. LSTM shares the same weights across layers, which reduces the numbers of parameters that the network compute. The GRU is an alternate solution for a vanishing gradient problem. It substitutes the simple neuron with a gated unit, which has fewer parameters than the LSTM neurons, because it lacks an output gate.[33]

In this study, extensive experiments were conducted to determine how many prior encounters is optimal to predict readmission. We conducted experiments by considering x encounters within the prior 2 years, where $x \in \{1, 2, 4, 8, 15, 30, 60, 80, 100\}$. The average number of encounters per patient in this period was 21, and the 90th percentile was 56. The variation in encounter number resulted in a non-unified length of feature vectors. Thus, in an experiment that considers up to 60 encounters, feature vectors lacking data were padded with 0s to ensure that feature vectors for all patients represent 60 encounters. The hypothesis of this study was that DL models outperform traditional models on a large benchmark, hence, a comparative analysis with a variety of evaluation metrics was performed to evaluate and compare the DL algorithms to the baseline traditional models. Moreover, to examine the importance of domain knowledge, we trained and tested the models on data with all laboratory studies included in the EHR dataset and compared with models trained and tested with a subset of laboratory studies based on prior papers reporting association with readmission (serum albumin, anion gap, arterial pH, bilirubin, blood urea nitrogen, carbon dioxide, serum creatinine, blood glucose, hematocrit, lactate, PaCO2, PaO2, serum sodium, troponin-I, venous pH, and white blood cell count).[11, 34] Using only a subset of laboratory studies may be beneficial by reducing dimensionality.

Patients were sorted randomly into 3 nonoverlapping subsets, where 70% were used for training, 10% for validation, and 20% for testing. We employed cross-validation techniques to find the hyperparameters that yield the best performances. For LSTM and GRU, we varied the number of neurons, dropout, batch size, and the number of epochs using a grid search. Following the literature, in conducted experiments dropout percentage varied from 0 to 50, and the number of neurons varied from 32, to 512. We selected a dropout of 0.1, 128 neurons, a batch size of 512, and 16 epochs for LSTM, and 12 epochs for GRU, since bidirectional GRU converges faster than 1-way LSTM. Sigmoid activation function and Adam optimizer were used. Traditional models were fine-tuned as well and the hyperparameters that yielded best performances were chosen.

Performance metrics and analysis

The performance of the methods used in our study was evaluated by five common metrics: Area Under the Receiver Operating Characteristic Curve (AUROC), Recall (also known as Sensitivity), Specificity, F1-score, and Accuracy. The formal definitions of these evaluation metrics are common and can be easily found.[34]

Statistical significance analysis was performed to evaluate the stability and significance of the proposed model's performance. We randomly selected different patients for training and testing and repeated the random selection 10 times to generate mean performance measures and 95% confidence intervals. LSTM was compared to the best

performing traditional model (RF) by t-test. A p-value <0.05 was considered statistically significant. The Temple University Institutional Review Board approved the protocol.

Results

A total of 36,641 patients with 2,836,569 encounters were analyzed. There were 9,130 patients with at least one readmission and 27,511 without a readmission. Influence of the number of encounters within the prior 2 years was evaluated for five prediction models where x encounters were considered for each model, and experiments were repeated for $x \in \{1, 2, 4, 8, 15, 30, 60, 80, 100\}$. Figure 2 presents the AUROC of the proposed model, LSTM, versus traditional models across various numbers of encounters. Bidirectional GRU was also performed but omitted because it achieved an identical AUROC to LSTM. LSTM outperformed traditional models on a large benchmark across all experiments with different number of encounters. On average, the LSTM models yielded an increase in AUROC of 0.06 when compared to the best performing traditional models, RF. Experiments show that predicting readmission based on a single prior encounter is not sufficient and yielded much lower performance (0.7 using the DL models and 0.68 using the best performing traditional model). DL models reached a plateau when trained using data from 30 encounters with minimal improvement thereafter. The DL algorithm yielded 0.07 increase in AUROC versus best performing traditional model RF when using the optimal number of encounters, 80.

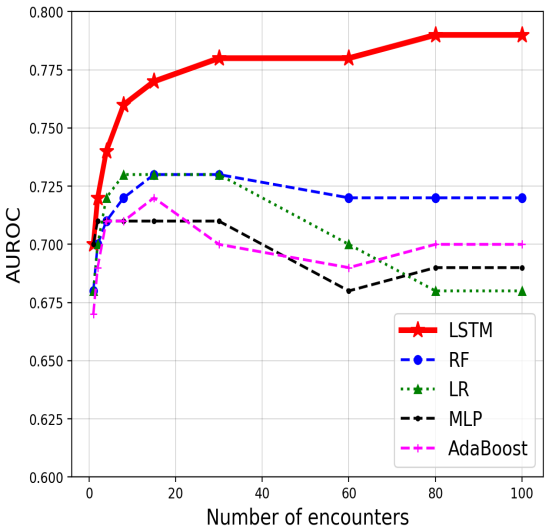

Figure 2. The proposed deep learning (DL) method's performance compared to baselines evaluated using the Area Under Receiver Operating Characteristic (AUROC) metric across varying numbers of prior encounters. Results show that DL methods outperform traditional methods on a large benchmark, and improvement in AUROC reaches a plateau at 80 encounters. Results show that using historical data of up to 30 encounters might be sufficient, 80 is optimal, and ≤15 yields poor performance.
LSTM - Long Short-Term Memory; RF - Random Forest; LR - Logistic Regression; MLP - Multi-Layer Perceptron.

Table 1 shows the performance of LSTM and traditional models using all laboratory tests from up to 80 of the most recent encounters in the prior 2 years. Overall, the confidence intervals were very small (<0.02), indicating a high degree of precision around the means. The proposed method, LSTM, obtained an average AUROC of 0.79 with a 95% CI of 0.001. The p-value obtained by comparing the LSTM AUROC to the second best-performing model (RF) was <0.0001, hence, LSTM performance was significantly greater than the traditional models.

Table 1. Performance of LSTM and traditional models using all laboratory tests from up to 80 of the most recent encounters in testing cohort of 7,329 patients with diabetes. Mean +/- 95% confidence interval (CI) are based on 10 runs.

Model N=7,329	AUROC +/-CI	Recall/Sensitivity	Specificity	F1-score	Accuracy
LSTM	**0.79 ±0.001**	**0.81 ±0.002**	**0.94 ±0.010**	**0.80 ±0.003**	**0.81 ±0.002**
RF	0.72 ±0.004	0.76 ±0.001	0.97 ±0.019	0.71 ±0.002	0.76 ±0.001
AdaBoost	0.70 ±0.000	0.76 ±0.000	0.94 ±0.000	0.73 ±0.000	0.77 ±0.000
LR	0.69 ±0.000	0.77 ±0.000	0.91 ±0.000	0.75 ±0.000	0.77 ±0.000
MLP	0.69 ±0.006	0.75 ±0.009	0.87 ±0.018	0.74 ±0.006	0.75 ±0.009

LSTM - Long Short-Term Memory; RF - Random Forest; LR - Logistic Regression; MLP - Multi-Layer Perceptron.

LSTM models achieved a Recall/Sensitivity of 0.81, indicating that performance was fairly strong at predicting true positives, (i.e., correctly classifying patients with readmissions). All models used in our study achieved a very good specificity, (i.e., the true negative rate). Thus, the trained models performed well at predicting patients who are not likely to be readmitted. LSTM achieved an F1-score of 0.80, indicating very good ability to distinguish between patients who will be readmitted or not.

To determine whether domain knowledge about laboratory studies is helpful, we conducted two different experiments where we trained and tested the model based on a subset of 16 unique laboratory studies selected by domain knowledge versus using all 981 unique laboratory studies included in the data. One Hot encoding techniques were utilized and modified to associate the laboratory result with each laboratory code. A long unique array of laboratory codes $unique\_lab\_codes$ was created. For each encounter, an array of zeros l of the same length as $unique\_lab\_codes$ was created. l consisted of the result at the same index of each laboratory test in $unique\_lab\_results$, to associate the result to a given laboratory code. An encounter without laboratory results would have an l of zeros, indicating that no laboratory test was conducted for a given encounter. Since most encounters contained <3 laboratory codes, this resulted in a sparse array. SVD was therefore utilized to learn an embedding of a sparse feature vector and reduce dimensionality. The Receiver Operating Characteristic (ROC) Curves of the LSTM models based on all laboratory studies or selected laboratory studies were identical (0.79, Figure 3).

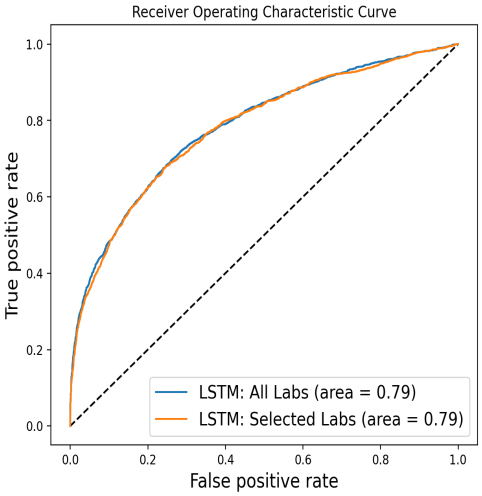

Figure 3. Receiver Operating Characteristic (ROC) Curves of the LSTM models using all laboratory studies or 16 selected laboratory studies.

Discussion

In this retrospective cohort of 36,563 patients with diabetes, DL models outperformed RF, MLP, AdaBoost, and LR models at predicting unplanned, all-cause 30-day readmission. The optimal LSTM model yielded an AUROC of 0.79 and accuracy of 0.81, indicating very good performance. Experiments designed to reveal the relationship between the number of prior encounters and model performance show that AUROC of the LSTM models increased as encounter number increased and plateaued at 30 encounters. Performance of the traditional models increased to a lesser extent up to prior encounter numbers of 15 or 30, then either plateaued (RF) or declined (MLP, AdaBoost, LR) as encounter number increased. Finally, an LSTM model that included a set of 16 laboratory tests selected by domain knowledge yielded equivalent performance to an LSTM model that included all available laboratory tests.

In our study, the DL models performed better than the traditional models. We are aware of 4 studies that compared DL models to traditional models for predicting readmission risk of patients with diabetes. Two of these studies demonstrated a clear advantage of DL approaches over traditional ML models,[23,24] while two studies found marginal benefit with DL approaches.[25,27] Performance of these DL models was variable with AUROC 0.61-0.97 and accuracy 0.69-0.95, none of which exceeded that of the best traditional ML models, which reported AUROC as high as 0.99 and accuracy of 0.99.[23,27] Comparisons of model performance across all these studies, however, is limited by the lack of standardized reporting of performance characteristics and variable approaches to testing. Our study considered with the prior studies that directly compared DL to traditional ML models, suggests that DL approaches usually yield better performance in this population.

We are unaware of other papers that have explored the relationship between the number of prior encounters and readmission risk model performance in patients with diabetes. In related work, however, one paper examined how the performance of models for predicting readmission risk in morbidly obese patients varied as the number of hospitalizations increased from a minimum of 2 up to 5.[35] AUROC increased from 2 to 3 hospitalizations then plateaued. In another broadly related study, we found that the performance of LR models for predicting readmission risk of patients with diabetes tended to increase as sample size increased from 2,000 up to 6,000, then plateaued.[36] This body of research suggests that experimentation across a range of encounter number and sample size may reveal thresholds that could optimize data analysis, balancing information quantity with dimensionality.

We are also unaware of other studies that have compared readmission risk models using laboratory data selected by domain knowledge with all laboratory data available in patients with diabetes. There is a tradeoff between including all laboratory data, which results in higher dimensionality and more computationally expensive models and involving a domain expert to select a subset of laboratory data, which can be costly and less feasible. Like the number of prior

encounters beyond which model performance did not improve, the finding that performance of the model with the laboratory data subset was equivalent to the model with all laboratory data suggests that there is a similar plateau for this domain. Whether or not this phenomenon generalizes to other patient populations should be investigated.

The presented LSTM models, which we are calling *e*DERRI™, are an extension of our prior models, the DERRI™ and DERRIplus.[8,11] In terms of AUROC, the *e*DERRI™ model performed better than the DERRI™ but worse than the DERRIplus. Unfortunately, the performance of 3 models cannot be directly compared in the current study because the dataset does not include zip code, employment status, or payer information. Unlike the DERRI™ and DERRIplus, the *e*DERRI™ models are developed with generally available EHR data such as demographics, vital signs, diagnostic and procedure codes, medications, laboratory tests, and administrative data as defined by the PCORnet CDM.[29] The CDM standardizes the abstraction of EHR data, enhancing the generalizability and scalability of models utilizing it. We plan to translate the *e*DERRI™ into an application embedded in an EHR system that will automatically generate readmission risk predictions for hospitalized patients with diabetes.

In addition to the generalizability of the CDM-based dataset, the current study has other notable strengths. The dataset is sampled from patients with a hospitalization between 7/1/2010 and 12/31/2020, which is much more recent than the datasets used for other currently published readmission risk models in diabetes patients. Also in contrast to the most used dataset, which only included hospital encounters with an associated diagnosis of diabetes and a length of stay less than 15 days, the current dataset included all encounter types regardless of the associated diagnosis, capturing both inpatient and outpatient data. Lastly, the sample size of 36,563 patients with 2,836,569 encounters provided ample data to develop DL models and conduct experiments up to 100 prior encounters.

There are some limitations worth acknowledging. The data were sampled from a single urban, academic health system. Therefore, generalizability of the models to other populations is unknown and requires testing. The lack of both patient and hospital zip code precludes estimating the distance between a patient's home zip code and the hospital, which is known to be associated with readmission risk.[8,11] Lastly, readmissions to other hospitals were not captured.

Conclusion

An LSTM model with very good performance predicting unplanned, all-cause 30-day readmission among patients with diabetes was developed and internally tested. LSTM models outperform traditional models at predicting readmission in this population. LSTM model performance initially increases as the number of prior encounters increases then plateaus. Carefully selected laboratory features can yield predictive models with performance equal to that of models based all available laboratory studies. Additional study is needed to externally validate the model.

Acknowledgements

This research was supported by the National Health Institute (NIH) under grant number R01DK122073.

References

1. Benbassat J, Taragin M. Hospital readmissions as a measure of quality of health care: Advantages and limitations. Archives of Internal Medicine. 2000;160:1074-1081.
2. Rubin DJ. Hospital readmission of patients with diabetes. Current Diabetes Reports. 2015;15:1-9.
3. Ostling S, Wyckoff J, Ciarkowski SL, Pai C-W, Choe HM, Bahl V, et al. The relationship between diabetes mellitus and 30-day readmission rates. Clinical Diabetes and Endocrinology. 2017;3:3.
4. Enomoto LM, Shrestha DP, Rosenthal MB, Hollenbeak CS, Gabbay RA. Risk factors associated with 30-day readmission and length of stay in patients with type 2 diabetes. J Diabetes Complications. 2017;31:122-127.
5. AHRQ, Healthcare cost and utilization project (hcup) national inpatient sample (nis). , 2018.
6. ADA. Economic costs of diabetes in the u.S. In 2017. Diabetes Care. 2018;41:917-928.
7. Rubin DJ, Shah AA. Predicting and preventing acute care re-utilization by patients with diabetes. Current Diabetes Reports. 2021;21.
8. Rubin DJ, Handorf EA, Golden SH, Nelson DB, McDonnell ME, Zhao H. Development and validation of a novel tool to predict hospital readmission risk among patients with diabetes. Endocr Pract. 2016;22:1204-1215.
9. Rubin DJ, Recco D, Turchin A, Zhao H, Golden SH. External validation of the diabetes early re-admission risk indicator (derri()). Endocr Pract. 2018;24:527-541.
10. Alamer AA, Patanwala AE, Aldayyen AM, Fazel MT. Validation and comparison of two 30-day re-admission prediction models in patients with diabetes. Endocr Pract. 2019;25:1151-1157.
11. Karunakaran A, Zhao H, Rubin DJ. Predischarge and postdischarge risk factors for hospital readmission among patients with diabetes. Med Care. 2018;56:634-642.

12. Alloghani M, Aljaaf A, Hussain A, Baker T, Mustafina J, Al-Jumeily D, et al. Implementation of machine learning algorithms to create diabetic patient re-admission profiles. BMC Medical Informatics and Decision Making. 2019;19:253.
13. Alturki L, Aloraini K, Aldughayshim A, Albahli S. Predictors of readmissions and length of stay for diabetes related patients. in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). 2019.
14. Bhatt V, Chakravorty T, Chakraborty S, *Re-admission rate prediction of diabetes patient: Health analytics-based approach*, 2022, Springer Singapore. p. 743-754.
15. Dinh Phu Cuong L, Wang D. A comparison of machine learning methods to predict hospital readmission of diabetic patient. Studies of Applied Economics. 2021;39.
16. Cui S, Wang D, Wang Y, Yu PW, Jin Y. An improved support vector machine-based diabetic readmission prediction. Comput Methods Programs Biomed. 2018;166:123-135.
17. Grampurohit S, *Diabetes patients hospital re-admission prediction using machine learning algorithms*, 2021, Springer Singapore. p. 485-497.
18. Neto C, Senra F, Leite J, Rei N, Rodrigues R, Ferreira D, et al. Different scenarios for the prediction of hospital readmission of diabetic patients. J Med Syst. 2021;45:11.
19. Ramírez JC, Herrera D, *Prediction of diabetic patient readmission using machine learning*, 2019, Springer International Publishing. p. 78-88.
20. Shang Y, Jiang K, Wang L, Zhang Z, Zhou S, Liu Y, et al. The 30-days hospital readmission risk in diabetic patients: Predictive modeling with machine learning classifiers. BMC Medical Informatics and Decision Making. 2021;21.
21. Shibly MMA, Tisha TA, Mazumder MMI. Predicting early readmission of diabetic patients: Toward interpretable models. in International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020. 2021. Springer.
22. Shih D-H, Huang F-C, Weng C-L, Shih P-Y, Yen DC. Thirty-day re-hospitalization rate prediction of diabetic patients. Journal of Internet Technology. 2020;21:2065-2074.
23. Hammoudeh A, Al-Naymat G, Ghannam I, Obied N. Predicting hospital readmission among diabetics using deep learning. Procedia Computer Science. 2018;141:484-489.
24. Hu P, Li S, Huang Y, Hu L. Predicting hospital readmission of diabetics using deep forest. in 2019 IEEE International Conference on Healthcare Informatics (ICHI). 2019.
25. Reddy SS, Sethi N, Rajender R. Evaluation of deep belief network to predict hospital readmission of diabetic patients. in 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, IEEE.
26. Sarthak, Shukla S, Prakash Tripathi S, *Embpred30: Assessing 30-days readmission for diabetic patients using categorical embeddings*, 2021, Springer Singapore. p. 81-90.
27. Welchowski T, Schmid M. A framework for parameter estimation and model selection in kernel deep stacking networks. Artificial Intelligence in Medicine. 2016;70:31-40.
28. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, et al. Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. Biomed Res Int. 2014;2014:781670.
29. Elixhauser A SC PL. Clinical classifications software (ccs): Agency for healthcare research and quality. 2014 Available at: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed 12/27/2021.
30. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical Care. 1998;36:8-27.
31. Centers for medicare & medicaid services (cms). 2016 all-cause hospital-wide measure updates and specifications report. Hospital-level 30-day risk-standardized readmission measure – version 5.0, 2016.
32. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. Neural computation. 1989;1:270-280.
33. Cho K, Merrienboer Bv, G¸lÁehre a, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using rnn encoder decoder for statistical machine translation. in EMNLP. 2014.
34. Powers DMW. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. Journal of Machine Learning Technologies. 2011;2:37-63.
35. Povalej Brzan P, Obradovic Z, Stiglic G. Contribution of temporal data to predictive performance in 30-day readmission of morbidly obese patients. PeerJ. 2017;5:e3230.

36. Zhao H, Tanner S, Golden SH, Fisher SG, Rubin DJ. Common sampling and modeling approaches to analyzing readmission risk that ignore clustering produce misleading results. BMC Medical Research Methodology. 2020;20:281.