Although most seriously ill patients wish to avoid aggressive and burdensome care, many receive invasive procedures and therapies in the final months of their lives with limited benefit.\textsuperscript{1-3} Such aggressive care near the end of life is associated with reduced quality of life for patients, excessive end-of-life medical spending, and a high proportion of patients dying in the hospital or other medical facility rather than at home.\textsuperscript{4,5}

Palliative care is a holistic approach to care administered at any point in a serious illness that offers patients support with intensive symptom management and complex medical decision-making, whereas hospice care utilizes a palliative interdisciplinary approach for patients with a prognosis of having less than 6 months to live; these patients typically desire less aggressive care.\textsuperscript{6} Palliative care is associated with improved patient quality of life,\textsuperscript{7,8} reduced hospital readmissions,\textsuperscript{9} decreased hospital length of stay,\textsuperscript{10} and reduced total cost of care.\textsuperscript{11,12} Despite multiple studies demonstrating that the majority of patients in the last of year of life experience a high and escalating symptom burden and considerable need for support with advanced care planning,\textsuperscript{13-16} providers and health care systems have struggled to proactively identify and support these high-risk patients and their families.\textsuperscript{17-19}

Improving the use of palliative care has been identified as a key driver of success for value-based reimbursement models such as accountable care organizations (ACOs), Medicare Advantage plans, and capitated insurance plans.\textsuperscript{20-24} One study of managed care patients provided with home palliative care support found a mean savings of $3908 per member per month in the last 3 months of life compared with propensity-matched controls, with hospice participation increasing from 25% to 70% and median hospice length of stay increasing from 9 to 34 days.\textsuperscript{25} Currently, the majority of patients who would benefit from palliative care services are either referred late in their disease course or not at all,\textsuperscript{26,27} so there has been intense interest in developing and implementing prognostic tools to proactively identify patients near the end of life who may benefit from palliative care services.\textsuperscript{28,29}

Prognostic models are commonly used in clinical practice to identify high-risk populations who may or may not benefit from the use of palliative care.\textsuperscript{30-32} Using Electronic Health Records and Claims Data to Identify High-risk Patients Likely to Benefit From Palliative Care

Aixia Guo, PhD; Randi Foraker, PhD, MA; Patrick White, MD, HMDC; Corey Chivers, PhD; Katherine Courtwright, MD, MSHP; and Nathan Moore, MD

ABSTRACT

OBJECTIVES: Palliative care has been demonstrated to have positive effects for patients, families, health care providers, and health systems. Early identification of patients who are likely to benefit from palliative care would increase opportunities to provide these services to those most in need. This study predicted all-cause mortality of patients who could benefit from palliative care.

STUDY DESIGN: Claims and electronic health record (EHR) data for 59,639 patients from a large integrated health care system were utilized.

METHODS: A deep learning algorithm—a long short-term memory (LSTM) model—was compared with other machine learning models: deep neural networks, random forest, and logistic regression. We conducted prediction analyses using combined claims data and EHR data, only claims data, and only EHR data, respectively. In each case, the data were randomly split into training (80%), validation (10%), and testing (10%) data sets. The models with different hyperparameters were trained using the training data, and the model with the best performance on the validation data was selected as the final model. The testing data were used to provide an unbiased performance evaluation of the final model.

RESULTS: In all modeling scenarios, LSTM models outperformed the other 3 models, and using combined claims and EHR data yielded the best performance.

CONCLUSIONS: LSTM models can effectively predict mortality by using a combination of EHR data and administrative claims data. The model could be used as a promising clinical tool to aid clinicians in early identification of appropriate patients for palliative care consultations.
specific interventions.30-37 However, a recent study38 highlighted the limitations of such models, including the use of small data sets for derivation, limitations in number and complexity of variables considered, oversimplified models, and constraints on the population included in the models. To overcome these limitations, the authors applied a deep learning model on electronic health records (EHRs) for a large number of patients.38 However, a remaining limitation of the study was that it employed only EHR data, which may be incomplete for patients who receive care in multiple medical systems, as most patients do.

In this paper, we integrated administrative claims and EHR data into our models. Claims data provide a richer picture for each patient, although the information can lag by 3 to 6 months. We hypothesized that the integration of claims and EHR data would provide more complete and accurate data than either source alone. We searched the PubMed and Google Scholar databases for literature prior to August 2020. To our best knowledge, no studies have compared EHR and claims data for predicting mortality for patients in the United States.

We implemented a deep learning method to predict mortality of inpatients and outpatients as a proxy for identifying patients to be considered for palliative care. Our long-term goal is to provide a clinical referral system for daily use to assist physicians in identifying patients who could benefit from palliative care.

**METHODS**

**Study Purpose**

We hypothesized that combining claims data with EHR data could achieve better performance for mortality prediction than using EHR or claims data alone. To test this hypothesis, we applied and compared mortality prediction performance across 4 types of machine learning and deep learning models—long short-term memory (LSTM), deep neural networks (DNN), random forest (RF), and logistic regression (LR)—using EHR data, claims data, and a combination of EHR and claims data.

**Data Source**

In this study, we used EHR data and administrative claims data from the Medicare and Medicare Advantage ACOs of a large integrated health care system. The system operates in 2 states and uses EpicCare EHR software. We included all 59,639 patients with encounter records from January 2017 to February 2020.

We implemented a deep learning method to predict mortality of inpatients and outpatients as a proxy for identifying patients to be considered for palliative care. Our long-term goal is to provide a clinical referral system for daily use to assist physicians in identifying patients who could benefit from palliative care.

**TAKEAWAY POINTS**

We conclude that the long short-term memory model can be used as a promising clinical tool to aid clinicians in early identification of patients who are appropriate for palliative care consultations.

- It can update predicted probability of mortality daily for each patient.
- It does not require expert knowledge or hand selection to design features.
- The predicted mortality probabilities may assist physicians in identifying patients for palliative care consultations.

**Study Design**

A patient was considered a positive case if the patient had a recorded date of death. All other patients were then considered negative cases. We first included 59,639 total unique patients based on the patient having at least 1 EHR record or claim record (“claims + EHR”). Among these 59,639 patients, when we extracted only the claims data of patients, there were 56,598 unique patients (“claims only”). When we extracted only the EHR data of patients, there were 55,316 unique patients (“EHR only”).

For the above 3 cases, we followed a similar process of feature extraction, data splits, model training, and evaluation as described below.

**Feature Extraction**

For each patient, we extracted all available encounter records. The features included diagnosis codes, procedure codes, medication codes, problem lists, demographics, and social history. For the social history, we combined the questions and answers as a column of “codes.” For example, if a patient had a question of “drug use” and the corresponding answer is “no,” then we combined the question and answer as “DrugUseNo.” There were 84,315 unique codes included in the analyses.

To better capture the pattern and nature of our longitudinal data, all codes were sorted in a time-increasing order and then mapped to 32-dimensional vector space by the technique of Word2Vec.39 The Python Genism Word2Vec model was used with the following hyperparameters: size (embedding dimension) was 32, window (the maximum distance between a target word and all words around it) was 5, min_count (the minimum number of words counted when training the model) was 1, and sg (the training algorithm) was CBOW, or the continuous bag of words. Each feature was then represented by a 32-dimensional numerical vector and also associated with a time point (age at event), which was calculated by the difference in years between the corresponding visit date and the birth date of the patient. We also included race as a feature in our study. Ultimately, each individual patient had their own numerical vector to represent their codes.

The number of events for different patients may vary. We selected 500 events for all patients; this empirical choice was based on proximity to the mean number (mean = 466 for case of claims + EHR) for all patients. These 500 events were sorted in a time order. If a patient had more than 500 events, only their most recent 500 were included. If fewer than 500 were available, zero events padding to 500 was utilized.

The input of LSTM had a dimension of \((n, 500, 34)\), where \(n\) was the number of samples/patients, 500 was the number of codes included, and 34 was the dimensionality of features (32-dimensional vector by Word2Vec, plus age and race). The input of DNN, RF, and LR had a dimension of \((n, 500 \times 34)\).
Data Splits for Training, Validation, and Testing
We included 59,639 patients for claims + EHR, 56,598 patients for claims only, and 55,316 patients for EHR only analyses.

The selected patients were randomly split into training (80%), validation (10%), and testing (10%) data sets. The number of patients in each data set is shown in eAppendix Table 1 (eAppendix available at ajmc.com).

The models were trained on the training data set and then validated on the validation data set. The model with the best performance on the validation data set was selected as the final model to be tested in the testing data set.

The Training of Machine Learning and Deep Learning Models
The embedded vectors were the inputs of machine learning and deep learning models. Our model is an LSTM neural network, which is composed of an input layer, 1 hidden layer (128 dimensions), and a scalar output layer. A binary cross-entropy loss function was employed as the output layer and a Sigmoid function was used as the activation function for the hidden layer. The Adam optimizer was used to optimize the model with a minibatch size of 256 samples. We did an extensive hyperparameter search for activation functions (Sigmoid, tanh, SeLU, and ReLU), as well as the embedding dimensions (32, 64). We did not extensively search other hyperparameters, such as number of LSTM layers, number of recurrent units, and batch size, as these hyperparameters were of minor importance.

We also compared the LSTM with other models: DNN, RF, and LR. The DNN model was composed of an input layer, 4 hidden layers (with 256, 128, 64, and 32 dimensions, respectively), and a scalar output layer. We used the Sigmoid function as the output layer and ReLU function at each hidden layer. Binary cross-entropy was used as loss function, and the Adam optimizer was used to optimize the models with a minibatch size of 256 samples. The RF and LR models were configured by the default options in the Scikit-learn package in Python 3.

Model Evaluation
We implemented a variety of metrics to evaluate the model performance in order to identify the best models. The metrics included area under the receiver operator characteristic (ROC) curves (AUCs), precision recall curves, overall accuracy, sensitivity, specificity, precision, and F1-score.

We also tested the model under different discrimination thresholds or cutoffs. The values of cutoffs used were 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99.

Separate Model Testing for Inpatients and Outpatients
In the study design, we included all inpatients and outpatients. Inpatients were defined as patients who had at least 1 hospital admission. Patients without hospital admissions were defined as outpatients. We investigated inpatients and outpatients separately in the testing data set by our trained LSTM models for claims + EHR analyses.

In the testing data set, 3143 patients had an admission and 2821 patients had only outpatient visits recorded. Among these, 509 inpatients and 107 outpatients were deceased and 2634 inpatients and 2714 outpatients were alive at the end of follow-up.

Figure 1 is the flowchart of the overall study. Analyses were conducted in 2020 using the libraries of Scikit-learn, Keras, Scipy, and Matplotlib with Python version 3.6.5.

RESULTS
The study found 4 main results. In this section, we list an overview of the main results. First, we showed the characteristics of the overall study population (Table) and observed that the majority of patients were female and White, with a mean age of 74 years. Second, we evaluated model performance by ROC curves, precision recall curves, and other performance metrics (Figure 2 and eAppendix Table 2). In all modeling scenarios, LSTM models outperformed the other 3 models, and using claims + EHR data yielded the best performance. Third, we showed the distribution of predicted probabilities for the case of claims + EHR by LSTM model to intuitively understand prediction performance results and potential thresholds for a desired accuracy (Figure 3 and eAppendix Figure). The distribution of predicted probabilities of LSTM on patients in the testing data set showed a high true-negative rate (specificity) and a low false-positive rate. Finally, we showed the LSTM model performance on inpatients and outpatients to understand the difference in performance among different populations (Figure 4 and eAppendix Table 3). The trained LSTM models performed better on the inpatient sample compared with the outpatient sample (AUC, 0.97 vs 0.92; average precision [AP], 0.92 vs 0.72).

Characteristics of the Overall Study Population
The Table details the characteristics of our overall study population. Approximately 10% of patients died over the course of follow-up. The mean age of patients who died was 78 years, and the mean age was 73 years for those who were alive at the end of follow-up. The majority of the patients were female and White. Approximately 76% patients had a diagnosis of essential (primary) hypertension, and deceased patients had higher proportion (84%) of this diagnosis compared with the patients who were alive (75%).

Model Performance by AUC and Precision Recall Curves
The LSTM model outperformed the other 3 models in terms of AUC values and precision recall curves. All the models achieved the best performance by using a combination of EHR and claims data compared with using each data type alone.

Figure 2 shows the model performance by AUC and precision recall curves. The AUC for the LSTM model was 0.97 (0.95 for DNN, 0.92 for RF, and 0.88 for LR) for claims + EHR, 0.96 (0.93, 0.90, and
0.88, respectively) for claims only, and 0.94 (0.90, 0.84, and 0.77) for EHR only.

Meanwhile, the LSTM model achieved an AP score of 0.89 (0.82 for DNN, 0.71 for RF, and 0.60 for LR) for claims + EHR, 0.87 (0.76, 0.66, and 0.60, respectively) for claims only, and 0.77 (0.64, 0.47, and 0.37) for EHR only.

Both the AUC values and precision recall curves indicated that LSTM outperformed the other 3 models. The curves also demonstrated that LSTM models had a strong early recall behavior.

**Model Performance by Other Metrics**
LSTM models outperformed the other 3 models using additional performance metrics, and using claims + EHR data yielded the best performance.

Appendix Table 2 lists other metrics we used to evaluate model performance. The metrics include overall accuracy, sensitivity, specificity, precision, and F1-score.

In all cases, LSTM demonstrated the best performance. For claims + EHR, it achieved an accuracy of 0.97 (vs 0.96 for claims only and
0.95 for EHR only), a sensitivity of 0.77 (vs 0.72 and 0.60, respectively), a specificity of 0.99 (vs 0.99 and 0.98), a precision of 0.91 (vs 0.91 and 0.79), and an F1-score of 0.83 (vs 0.81 and 0.68).

Distribution of Predicted Probabilities

The LSTM model had high sensitivity and specificity. The eAppendix Figure shows the distribution of predicted probabilities for claims + EHR by the LSTM model on the testing data set. A majority of deceased patients were predicted by the model to have high probabilities of mortality, and a small proportion of them had lower predicted probabilities of mortality, which means the model had a high true-positive rate (sensitivity) and a low false-negative rate. Of the patients who were alive at the end of follow-up, only a few were predicted to have a high probability of mortality and most of them had very low predicted probabilities of mortality, which means the model had a high true-negative rate (specificity) and a low false-positive rate.

Figure 3 shows the model performance of LSTM under different cutoffs of predicted probabilities. When using lower cutoffs (eg, 0.1), the model had a higher sensitivity (88%) with lower precision (58%) and F1-score (70%). When using higher cutoffs, the sensitivity of the model decreased and specificity increased. The F1-score first increased and then decreased as the cutoffs became higher.

Performance on Separate Inpatient and Outpatient Cohorts by the Trained LSTM Model

The LSTM models achieved better performance on inpatient compared with outpatient samples. We also tested the trained LSTM model on the inpatients and outpatients separately. The performance results are summarized in eAppendix Table 3 and Figure 4.

The trained LSTM models performed better on inpatients compared with outpatients. The AUC value was 0.97 for inpatients (0.92 for outpatients), and the AP score was 0.92 for inpatients (0.72 for outpatients). The other metrics followed a similar trend.

DISCUSSION

In this study, we utilized EHR data and administrative claims data from a large health system to predict mortality of patients to better identify patients who might benefit from palliative care. We investigated 3 data sets, using both EHR data and claims data, only claims data, and only EHR data. We compared LSTM models with DNN, RF, and LR models.

Our results indicated that deep learning LSTM models can most effectively predict the mortality of patients compared with other models using all data sets. One possible reason for observing better performance of the LSTM model was that it can learn from experience based on the time-series data when the time steps are of arbitrary size.47 LSTM models can also handle vanishing gradient problems by retaining information (via memory unit) for an arbitrary time period.48 The memory unit can determine which information needs to be retained, forgotten, and output to the next computational unit.49

In addition, our framework for using a LSTM model for mortality prediction has additional advantages, as it (1) is suitable for multiple data format sources, (2) does not require expert knowledge or hand selection to design features, (3) can dynamically update predicted mortality probabilities based on newly added features of patients, and (4) is a more general framework that can be used in both outpatient and inpatient systems. Understanding the relative performance of these models for identifying high-risk patients may help health care systems and managed care organizations select the optimal model for their organization.

One possible reason for observing better performance of the LSTM model was that it can learn from experience based on the time-series data when the time steps are of arbitrary size. Therefore, LSTM models can also handle vanishing gradient problems by retaining information (via memory unit) for an arbitrary time period. The memory unit can determine which information needs to be retained, forgotten, and output to the next computational unit.

In addition, our framework for using a LSTM model for mortality prediction has additional advantages, as it (1) is suitable for multiple data format sources, (2) does not require expert knowledge or hand selection to design features, (3) can dynamically update predicted mortality probabilities based on newly added features of patients, and (4) is a more general framework that can be used in both outpatient and inpatient systems. Understanding the relative performance of these models for identifying high-risk patients may help health care systems and managed care organizations select the optimal model for their organization.

Our results also indicated that models achieved the best performance by using both EHR and claims data, the next best by using claims data only, and the worst by using only EHR data. This was not surprising, as the combined EHR data and claims data provided more information for the models to discover and identify more potential patterns, which may contribute to the accurate predictions. One
FIGURE 2. Prediction Performance for 4 Models by AUC (A, C, E) and Precision Recall Curves (B, D, F).

AUC, area under the receiver operating characteristic curve; DNN, deep neural networks; EHR, electronic health record; LR, logistic regression; LSTM, long short-term memory; RF, random forest.

*A results for “claims + EHR” are shown in the first row (A, B); for “claims only,” the second row (C, D); and for “EHR only,” the third row (E, F).*
possible reason for only claims data performing better than only EHR data is that claims data usually contain more complete information for patients than EHR data; most patients receive care in multiple medical systems, which can result in incomplete EHR data. A study by Zeltzer et al of 118,510 Israeli patients found that over longer time durations, the relative benefit of claims data appeared to increase compared with EHR data, which were associated with greater prognostic accuracy in short-term outcomes.49

There is a consensus that effective use of palliative care is critical to the success of value-based care organizations by improving quality and reducing costs for patients near the end of life, where spending is concentrated. One of the key barriers to widespread adoption of palliative care to date has been correctly identifying which patients to target. Smith et al report that “[s]ystematic identification of patients for whom specialist and/or primary palliative care interventions are likely to improve quality of life and reduce cost is essential in setting up a successful population management program. Unfortunately, widely accepted criteria for identifying patients who would most benefit from palliative care have not been developed.”50 Current approaches to identifying patients for palliative care often focus on individual disease states (eg, congestive heart failure, chronic obstructive pulmonary disease) or historical utilization (eg, hospitalizations in the past 12 months), which do not identify all patients at risk and often identify patients too late in the disease course. Typically, these programs focus on either inpatients or outpatients but do not include both.

The advances in this article demonstrate a more accurate, holistic approach to identifying patients at high risk of mortality using all available sources of patient information (claims and EHR) for both inpatients and outpatients. If this approach is validated in additional populations, it can be used to test the impact of palliative care for patients in different phases of care on quality, spending, and utilization. This would allow value-based care organizations to more accurately measure the added value of palliative care for various populations of patients.

To better understand the differing performance for inpatients and outpatients, we investigated the number of records (source of knowledge) for each group. The inpatients had more information, with a mean number of 848 records, compared with the outpatients, with a mean number of 136 records. Thus, we hypothesize that the larger amount of information provided by inpatients to the LSTM model enabled better prediction of mortality compared with outpatients. Moreover, we tested the LSTM model for the case of claims + EHR in 4 patient groups with different numbers of records: fewer than 150 (n = 1680 patients), between 150 and 300 (n = 1309 patients), between 301 and 500 (n = 1104 patients), and more than 500 (n = 1871 patients). We observed that the number of records had an impact on model prediction results. Specifically, prediction results became less accurate for patients with fewer records. Future studies will also investigate the model performance on other patient groups, including different disease groups.

Limitations

Our study has several limitations. First, our models were trained based on data from 1 site, and additional research is needed to determine their scalability to other health care settings. Second, we included only a subset of the population (Medicare/Medicare Advantage) who is older and has more comorbidities than the general population, so prediction for other age groups with fewer comorbidities may not be as accurate.51-53 Unknown model “fairness” with regard to protected attributes might be caused by machine/deep learning models.54,55 Third, we predicted the overall status of death rather than the timing of death from a clinical event, although more than 96% of deaths occurred over a 1-year follow-up. Future studies will use similar methods to predict time to death, in-hospital mortality, and 30-day mortality at admission, which may help clinicians decide on the timing of palliative care interventions. These models used comprehensive data sets that were assembled after the last follow-up for these patients. Using these models prospectively may be challenged by our ability to assemble, analyze, and interpret them in real time. Future studies will focus on the implementation of this proof-of-concept study.

Conclusions

We demonstrate that deep learning LSTM models can effectively predict mortality by using a combination of EHR data and administrative claims data. The promising predictions of all-cause mortality

FIGURE 3. Model Performance of LSTM Under Different Predicted Probability Cutoff Values

LSTM, long short-term memory.
could assist physicians in suggesting and providing palliative care consultations to appropriate patients.

**Acknowledgments**

The authors would like thank Karen Shakiba, Jacob Heidbrink, Ly Mettlach, and Ryan Soluade at BJC Medical Group for their research assistance.

**Author Affiliations:** Institute for Informatics (AG, RF) and Department of Internal Medicine (RF, PW), Washington University School of Medicine, St Louis, MO; University of Pennsylvania Health System (CC), Philadelphia, PA; Division of Pulmonary, Allergy, and Critical Care, and Palliative and Advanced Illness Research Center, Perelman School of Medicine at the University of Pennsylvania (KC), Philadelphia, PA; BJC Medical Group and BJC Accountable Care Organization (NM), St Louis, MO.

**Source of Funding:** Work included in this document was produced by the research team. This work was produced with the support of the Big Ideas Program, a BJC HealthCare and Washington University internal grant program, hosted by the Healthcare Innovation Lab and the Institute for Informatics.

**Author Disclosures:** The authors report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

**Address Correspondence to:** Aixia Guo, PhD, Washington University School of Medicine, 600 S Taylor Ave, Ste 102, St Louis, MO 63110. Email: aixia.guo@wustl.edu.

---

**FIGURE 4. LSTM Model Performance on Inpatients and Outpatients in the Testing Data Set**

A. Inpatient

B. Inpatient

C. Outpatient

D. Outpatient

- **AUC**, area under the receiver operating characteristic curve; LSTM, long short-term memory.
- The predicted probability distribution for inpatients [A] and outpatients [B] are shown. Also shown are area under the receiver operating characteristic curves [C] and precision recall curves [D] for inpatients and outpatients, respectively.
**eAppendix Table 1.** Data Split for Training, Validation, and Testing for Claims Only, EHR Only, and Both Claims and EHR

<table>
<thead>
<tr>
<th>Case</th>
<th>Class</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claims + EHR</td>
<td>Alive</td>
<td>42,822</td>
<td>5372</td>
<td>5348</td>
<td>53,542</td>
</tr>
<tr>
<td></td>
<td>Deceased</td>
<td>4889</td>
<td>592</td>
<td>616</td>
<td>6097</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>47,711</td>
<td>5964</td>
<td>5964</td>
<td>59,639</td>
</tr>
<tr>
<td>Claims only</td>
<td>Alive</td>
<td>40,570</td>
<td>5079</td>
<td>5046</td>
<td>50,695</td>
</tr>
<tr>
<td></td>
<td>Deceased</td>
<td>4708</td>
<td>581</td>
<td>614</td>
<td>5903</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>45,278</td>
<td>5660</td>
<td>5660</td>
<td>56,598</td>
</tr>
<tr>
<td>EHR only</td>
<td>Alive</td>
<td>39,875</td>
<td>4999</td>
<td>4992</td>
<td>49,866</td>
</tr>
<tr>
<td></td>
<td>Deceased</td>
<td>4377</td>
<td>533</td>
<td>540</td>
<td>5450</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>44,252</td>
<td>5532</td>
<td>5532</td>
<td>55,316</td>
</tr>
</tbody>
</table>

EHR, electronic health record.
### eAppendix Table 2. Metrics to Evaluate the Model Performance

<table>
<thead>
<tr>
<th>Case</th>
<th>Model</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claims + EHR</td>
<td>LSTM</td>
<td>0.97</td>
<td>0.77</td>
<td>0.99</td>
<td>0.91</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>DNN</td>
<td>0.95</td>
<td>0.73</td>
<td>0.98</td>
<td>0.78</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.90</td>
<td>0.02</td>
<td>1.00</td>
<td>1.00</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.91</td>
<td>0.57</td>
<td>0.95</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Claims only</td>
<td>LSTM</td>
<td>0.96</td>
<td>0.72</td>
<td>0.99</td>
<td>0.91</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>DNN</td>
<td>0.94</td>
<td>0.64</td>
<td>0.98</td>
<td>0.78</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.90</td>
<td>0.05</td>
<td>1.00</td>
<td>0.94</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.90</td>
<td>0.60</td>
<td>0.94</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td>EHR only</td>
<td>LSTM</td>
<td>0.95</td>
<td>0.60</td>
<td>0.98</td>
<td>0.79</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>DNN</td>
<td>0.93</td>
<td>0.58</td>
<td>0.97</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.90</td>
<td>0.02</td>
<td>1.00</td>
<td>0.92</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>0.88</td>
<td>0.45</td>
<td>0.92</td>
<td>0.39</td>
<td>0.42</td>
</tr>
</tbody>
</table>

DNN, deep neural networks; EHR, electronic health record; LR, logistic regression; LSTM, long short-term memory; RF, random forest.
**eAppendix Table 3.** Metrics to Evaluate the LSTM Model Performance on Inpatients and Outpatients Separately

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM inpatient</td>
<td>0.97</td>
<td>0.77</td>
<td>0.99</td>
<td>0.91</td>
<td>0.83</td>
</tr>
<tr>
<td>LSTM outpatient</td>
<td>0.95</td>
<td>0.73</td>
<td>0.98</td>
<td>0.78</td>
<td>0.76</td>
</tr>
</tbody>
</table>

LSTM, long short-term memory.
**eAppendix Figure.** Prediction Probabilities on the Testing Data Set by LSTM Model for the Case of “Claims + EHR”

EHR, electronic health record; LSTM, long short-term memory.