Mining a large-scale EHR with machine learning methods to predict all-cause 30-day unplanned readmissions

Haijun Zhai\textsuperscript{1}, Srikanth Iyer\textsuperscript{2,3}, Yizhao Ni\textsuperscript{1}, Todd Lingren\textsuperscript{1}, Eric Kirkendall\textsuperscript{1}, Huaxiu Tang\textsuperscript{1}, Qi Li\textsuperscript{1}, Imre Solti\textsuperscript{1,3*}  
\textsuperscript{1}Division of Biomedical Informatics, \textsuperscript{2}Division of Emergency Medicine, \textsuperscript{3}James M. Anderson Center for Health Systems Excellence, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA  
* Senior and corresponding author  
Imre.Solti@cchmc.org

ABSTRACT

This study proposed a novel hybrid approach based on Natural Language Processing (NLP) and Machine-Learning (ML) to predict all-cause 30-day unplanned readmission by automatically analyzing a large-scale EHR. In the approach, we considered both the structured data and the clinical narrative text (physician notes) of the electronic health records (EHR). Six feature sets were created on different information sources. Logistic regression plus LASSO was used to create the most predictive models on the six different feature sets. The predictions of the six classifiers were fed into SVM to make a final prediction. Our approach achieved 0.764 AUC on the test set, which was statistically significantly higher (over 11 percent) than two previously published methods implemented in many institutions.

Decision support tools can predict the probability of readmissions within 30 days based on predictive factors, can help identify inpatient admissions that are at high-risk of subsequent readmission, and can prioritize resources to reduce this risk. A CDS tool of this type could lead to a reduction of costs and an improvement in outcomes of patients admitted to the hospital.

Various readmission prediction models have been developed in the past [6]. All these models were manually designed with expert knowledge to determine predictive factors and were proposed before widespread implementation of electronic health records (EHR). They therefore leveraged only a small fraction of the content available in current EHR systems. However, readmission is a complex phenomenon, and there are many associated factors. Despite this knowledge, the detection and utility of such factors has not been studied widely due to data accessibility issues, methodological issues and so on.

The aim of this study is to evaluate a proposed hybrid approach based on Natural Language Processing (NLP) and Machine-Learning (ML) to predict all-cause 30-day unplanned readmission by automatically analyzing a large-scale EHR. Our approach considered both the structured data and the clinical narrative text (physician notes) of the EHR.

To our knowledge, we were the first to conduct prediction of readmission by automatically analyzing a large-scale EHR. Instead of creating a prediction model on a small number of manually collected variables in previous studies, all clinical factors (which involve an extremely large number of variables) were taken into consideration in this study. A hybrid approach was proposed to
accommodate different types of clinical factors. In summary, the major contributions of this study:

• We proposed a hybrid approach to accommodate different types of clinical factors.

• We were the first to conduct readmission prediction by automatically analyzing a large-scale EHR.

• We applied systematic methods to do the feature selection and modeling, which generated the optimal combination of small number of (several hundred) features from an extremely large number (hundreds of thousands) of candidates.

• We were the first to explore using clinical notes to predict readmission by applying NLP.

• In addition to considering clinical information of inpatients at admission, we proposed a novel feature extraction method to capture dynamic characteristics of clinical variables to improve the performance.

II RELATED WORK

There has been much work conducted on readmission prediction and the reduction of readmission rates in the past several decades. Various readmission prediction models have been developed and evaluated. In 2011, Kansagara et al. [6] conducted a systematic review of all the models created up-to-date. As early as 1985, Smith et al. conducted analyses on 14 variables of patients at the time of discharge and found that three variables were significantly associated with readmissions. A score was calculated based the five variables and used to risk-stratify patients into groups according to levels of risk for readmission. The derived score provided a way to recognize high-risk patients for clinical interventions. The most popular model was “LACE” [7]. It was created by Walraven et al. using retrospective data from medical and surgical patients in Canada. The “LACE” index was derived from four independent variables, including length of stay (“L”), acuity of the admission (“A”), comorbidity of the patient (measured with the Charlson comorbidity index score) (“C”), and emergency department use (measured as the number of visits in the six months before admission) (“E”). Gruneir et al. [8] successfully applied the LACE model to a Canadian population and adapted it to real-time automation. More models created on different small set of variables can be found in the literature [9-14].

In addition to the studies on general population, some studies focused on specific populations. For example, Coleman et al. [15] used variables collected from data available at the time of hospital discharge to predict the risk of readmission for patients aged 65 and older. Furthermore, they enhanced the performance by adding a questionnaire that was not typically completed until after an index hospitalization. A few additional models [16-20] predicted the risk of readmission for this same population using different methods. There are also studies focused on disease-specific readmissions. Philbin et al. [21] developed a method for identifying a patient’s risk for hospital readmission for congestive heart failure (CHF) using information derived exclusively from administrative data sources and available at the time of an index hospital discharge.

In the last two years other studies have been performed on prediction of readmission [22-26]. For example, Baillie et al. [23] discovered that a single risk factor, inpatient admissions in the past 12 months, had strong predictive ability and they implemented a prediction model based on that single factor. In addition to clinical data, studies have been expanded to explore a wide range of information. For example, Beck et al. [27] explored the usage of geographic information to predict the readmission (readmission and emergency department revisit). They concluded that the risk index generated on geographic social information can help identify asthmatic children likely to return to the hospital.

III DATA

Our study population was obtained from an urban tertiary care pediatric hospital between January 1, 2010 and August 31, 2012. There were 71,909 admissions (encounters) to the inpatient wards during this time period. We identified 5,942 case and 57,656 control encounters, which summed to 63,598 encounters in total after removing planned readmissions and those admissions that would not be eligible for readmission within 30 days of the end of our study period (Figure 1). Cases are defined by encounters which have an unplanned readmission within 30 days.

For each encounter, we collected all structured entries of the EHR, including vital signs, labs, clinical assessments, medications, diagnoses and procedures, as well as all clinical notes (Table 1 shows the details of the EHR).
We chronologically split the 63,598 encounters into two groups. The first part included all encounters occurred in the first 24 months, which had 48,605 encounters (4,444 cases and 44,161 controls) and was used for training and evaluating the algorithms in a 10-fold cross-validation setting. The second part included all the encounters that occurred in the latter eight months and contained 14,993 encounters (1,498 cases and 13,495 controls), was kept unseen and used as test set.

Figure 1: Steps to generate cases and controls (Note 1: Distinguishing planned vs unplanned readmission was conducted based on time difference between the readmission schedule time and the exact readmission time. If a readmission was scheduled 10 hours earlier than the exact readmission time, the readmission was recognized as planned readmission. The threshold of 10 hours was determined by the analysis of readmission distribution on the time difference. The results of this distinguishing strategy were manually verified by our clinical experts, which shows over 90 percent accuracy. Note 2: 1,865 encounters in the bottom-right part of this figure may be readmitted after 8/31/2012 so all of them were excluded from the control set)

IV METHODS

Figure 2 shows the diagram of our approach that consists of three main components:

1. Feature Extraction: discretizing continuous variables, categorizing/grouping nominal variables, extracting CUls (Concept Unique Identifiers) from clinical notes.

2. Feature Selection and Modeling: identifying the most predictive features for each source and build the source-specific classifiers.

3. Classifiers Integration: integrating classifiers created on different information sources.

<p>| Table 1: EHR |</p>
<table>
<thead>
<tr>
<th>Description</th>
<th>Number of Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medications</td>
<td>The medications taken by the patients in the hospital, e.g., lincomycin, opioid agonists, benzodiazepines, ...</td>
</tr>
<tr>
<td>Diagnoses</td>
<td>Billing diagnoses, e.g., cellulitis, leg pain, asthma with exacerbation, ...</td>
</tr>
<tr>
<td>Labs</td>
<td>Lab test results, e.g., CBC with differential, glucose, ...</td>
</tr>
<tr>
<td>Vital signs and clinical assessments</td>
<td>For example, Temperature, blood pressure, patient pain experience, patient behaviors, ...</td>
</tr>
<tr>
<td>Demographics</td>
<td>For example, birthday, name, zipcodes, ...</td>
</tr>
<tr>
<td>Procedures</td>
<td>For example, CT abdomen/pelvis with contrast, wound culture and gram stain, cervical spine x-rays, ...</td>
</tr>
<tr>
<td>Clinical Notes</td>
<td>Clinical text written by the care providers</td>
</tr>
</tbody>
</table>

1 FEATURE EXTRACTION

Feature extraction is a major challenge in this study. We have a large amount of available information which contains an extremely large number of variables. Most of the variables are not well formatted and need to be processed. Considering the different sources, viewpoints and available time of information, we created six distinct feature sets instead of one on different kinds of information. The six feature sets are 9Var, ADInfo, DynInfo, ACUI, PCUI and DCUI.

1.1 9VAR

As initial work, we collected predictive factors for readmission from the literature [7-26] and included the factors that can also be found in our dataset. We then built a feature set by processing these collected factors. There were two reasons for creating this feature set. The first reason is that previous studies have shown that these factors perform well, which was also demonstrated by our pre-study experimental results. Another reason is that most of these factors are proxies (i.e., Number of admissions in the last year (Admissions), Number of emergency department visits in the last year (ED visits), Length of Stay (LOS), Number of medications taken in the last two days of hospitalization (Medications), Number of diagnoses (Diagnoses), and number of procedures (Procedures)), which can only be included manually. Following the methodology from previous studies [7, 20, 21], we performed discretization on the numeric factors (i.e., Admissions, ED visits, LOS,
Medications, Diagnoses) and did feature selection using the chi-square statistical test. Nine variables were used in the feature set; age, sex, race, ED visits, Admissions, LOS, Medications, Diagnoses, and Procedures.

![Diagram](image)

**Figure 2: The diagram of our approach**

### 1.2 ADINFO

Processing and modeling structured clinical information is resource-intensive. The status of an admitted patient changes dynamically due to the various interventions and treatments conducted during a hospital stay. All information such as measurements, clinical assessments, medications, procedures, diagnoses, etc., is captured in our EHR. Structured entries are critical part of all EHR, however it is quite challenging to process this data.

One common issue encountered with structured data is that the same variable may have multiple measurements over time. For example, a patient’s temperature can be measured every few hours. To address this issue we created a feature set called “Admission & Discharge Information” (ADInfo). As the status of a patient at admission is widely considered in previous studies, we followed previous studies in setting up a time window to collect the earliest measurement of each clinical variable to represent the patient status at admission. In this study, we set the time window to four hours (i.e., two hours before and after the admission point), which is much less than those used in previous studies. The reason is that variables in this dataset [28] have a much higher availability than previous studies and the narrower window intuitively leads to more accurate status of patients at admission. In addition to status at admission, we also considered patient status before discharge by including the most recent measurement of each variable in a time window before discharge. Considering the decreased availability of variables at discharge, we increased the time window to six hours. In summary, two distinct factors with the measurements for each clinical variable to represent the inpatients’ status at admission and discharge were included in ADInfo, respectively. Figure 3 (top) shows an example of the feature extraction for ADInfo.

We then conducted discretization on the two types of variables included in ADInfo, continuous and nominal variables. For each nominal variable, we used the original values as categories. For continuous variables we used two different methods to perform the discretization. The first is an unsupervised method learnt from the literature [28, 29] in which each variable was discretized into three categories (“Low”, “Medium” and “High”). The cut-off points of 5% and 95% quantiles are based on different age...
groups (i.e., 1-3 months, 4-12 months, 1-4 years, 4-12 years, >12 years). The cut-off points were generated by gathering all the measurements of each variable in the dataset. This method is widely applied in pediatric analysis. Another method is a supervised method based on the Chi-square test called “ChiMerge” [30, 31]. The unsupervised method was used to create the final model since the pre-study experimental results showed that there was no significant difference between the results of unsupervised method and supervised method, and the unsupervised method is more computationally efficient and stable. Finally, we had a binary-valued feature set for ADInfo where 1/0 indicated presence/absence of the feature for the encounter.

1.3 DYNINFO

Although ADInfo could describe patient’s clinical status at admission and discharge, it could not capture patient status changes during an encounter. Hence, we created another feature set called “Dynamic Information” (DynInfo). In DynInfo, all measurements during an encounter for a variable were taken into consideration. Similar to what we did for ADInfo, we conducted discretization on all variables. Instead of binary-valued feature in ADInfo, we used Term-Frequency (TF) to represent the feature values in DynInfo. In addition, we included the number of changes in two sequential measurements for nominal variables (i.e., number of changes from a value to a different value in two sequential measurements), as well as number of increases and decreases in two sequential measurements for continuous variables (i.e., number of increases from “Low” to “Medium” or “Medium” to “High” or “Low” to “High”, number of decreases from “High” to “Medium” or “Medium” to “Low” or “High” to “Low”, in two sequential measurements). The pre-study experimental results showed statistically significant improvement by including the number of changes in the feature set. Finally, we conducted the linear scaling normalization on the DynInfo features:

\[ x' = \frac{x - \min(x)}{\max(x) - \min(x)} \]

which is commonly used in the machine learning literature [32].

---

**Figure 3 (top): Example of feature generation for ADInfo**

**Features for temperature**
- At admission: High
- At discharge: Medium

**Temperature records of an encounter**
- Admission point: 2 hours
- Discharge point: 6 hours

**Temperature with value “H” at admission**
- 2 hours

**Temperature with value “M” at discharge**
- 6 hours

---

**Figure 3 (bottom): Example of feature generation for DynInfo**

**Features for temperature**
- High: 3
- Medium: 4
- Low: 2
- Increased: 3
- Decreased: 1

**Figure 3 Examples of feature extraction for ADInfo and DynInfo** (Note 1: “H”, “M” and “L” are abbreviations of “High”, “Medium” and “Low”, which are the discretization results of temperature based on the unsupervised method. Note 2: The black arrows indicate the time point when patients were admitted/discharged. Note 3: Red arrows indicates the measurements used for ADInfo. Note 4: In the figure the admission point is not located at the beginning of the bar since typically clinical assessments and measurements were taken before patients got admitted.)
1.4 CLINICAL NOTES

Clinical notes include valuable information about the patients’ clinical status. In recent years many studies have been conducted on using clinical notes to build automated clinical support system for diagnosis [33, 34] and patient status assessment [35]. In this study, we also explored the utilization of clinical notes for predicting readmission. To extract the features from clinical notes, we first performed a linguistic processing on the notes using the Clinical Text Analysis and Knowledge Extraction System (cTAKES) toolkit [36]. cTAKES includes a lexicon lookup component, which matches the tokens of the notes against concepts from the Unified Medical Language System (UMLS) metathesaurus [37] and assigns UMLS concept identifiers (CUIs) to the tokens when a match is found. It can also perform negation recognition such as non-observed signs/symptoms. We used this semantic information (i.e., CUIs) and the negation information (i.e., negative CUIs) to create a note-based binary-valued feature set (i.e., 1/0 indicates the presence/absence of the corresponding CUI or negative CUI). In addition, we split the clinical notes into three major types, admission notes, discharge notes, and progress notes (i.e., “H&P”, “D/C Summaries”, and other notes), and build feature sets on individual note types. The intention of the split was to build feature sets based on the availability of the notes at different time points and compare the corresponding predictive performance.

2. FEATURE SELECTION AND MODELING

After feature extraction we had six feature sets, where over 300,000 features were included in ADInfo and DynInfo and around 20,000 features in each CUI set. To filter out redundant features and noise, we pre-filtered the data using statistical tests. Chi-square test and t-test were applied to the binary-valued and real-valued features respectively, where the features with p-value >0.1 were ruled out. After pre-filtering, we had a set of refined features (around 15,000 features for ADInfo and 7,000 for DynInfo and 9,000 for each CUI set).

In order to find the most predictive model, we attempted standard machine learning classifiers such as Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM) and Logistic Regression (LR) with the top N most significant features (Top N). We also used logistic regression with least square shrinkage and selection operator (LASSO) to automatically select features.

3. INTEGRATION

Six classifiers were created on the six different feature sets. To make a final decision, the predictions (i.e., probabilities) of the six classifiers were fed into an additional classifier to perform the final prediction. In our pre-study experiment we tried NB, DT, LR, SVM, Majority Voting (MV) to integrate the predictions. Experimental results showed that SVM (with RBF kernel) worked the best and was thus used as the integrating classifier in our approach.

4. BASELINE METHODS

For comparison, we implemented two previously published methods. The first method was LACE which is widely used and evaluated in many institutions [7]. The second method was from a recent publication of a readmission study [23] which utilized a single risk factor – inpatients admissions in the previous 12 months (Admissions) – to predict readmission.

5. EXPERIMENTAL PLATFORM

In this study, we used Weka version 3.6.8 and R version 2.15.3 as our experimental tools. In addition, we used LIBSVM [38] for the implementation of our SVM classifier. For the implementation of LR plus LASSO classifier, we used the R package implemented by Goeman [30, 40], which utilized a computationally efficient gradient-ascent algorithm.

6. EVALUATION METHODS

We used standard 10-fold cross-validation on the training set to evaluate our approach and determine parameters. The finalized model was further evaluated on the test set. We used standard metrics to measure the algorithm’s predictive performance: Positive Predictive Value (PPV), Sensitivity and F-measure (F). We also measured specificity, created ROC curves and computed the area under the curve (AUC). In
addition, we used a paired t-test to perform a significance test on AUC between different classifiers.

V RESULTS

In our experiments, we first compared the results of different machine learning algorithms on each feature set. We only showed the results on DynInfo in Table 2, but the same conclusions have been drawn from the other feature sets. Table 2 shows the performances of NB, LR, DT, SVM combined with top N most-significant features, where the parameter N and the parameters of the algorithms were tuned by cross-validation based on AUC. The performance of LR with LASSO can also be found in Table 2. We observed that LR with LASSO significantly outperformed all the other methods by calculating paired t-test on AUC at 0.05 p-value. LR with LASSO algorithm was therefore used to train the prediction models in the rest of our experiments.

Figure 4 shows the effect of the parameter lambda of LASSO on the AUC and number of selected features. We observed that LASSO achieved the best results on DynInfo when lambda was equal to 16 and around 320 features were selected.

Table 3 shows the performance of LR+LASSO on each feature set. We observed that the classifier with any of these feature sets significantly outperformed the two baselines. In addition, the classifier with DynInfo achieved the best performance. All the comparisons were conducted by paired t-test on AUC at 0.05 p-value.

Table 4 presents the performance on different combination of the feature sets. We observed that as more feature sets were included, the performance was improved significantly.

Finally, we compared the performance of our approach (integrating the six classifiers created on each feature set) against the two baselines in Table 5. Our approach achieved more than 13 percent improvement on AUC. The improvement was statistically significant with paired t-test. The PPV-sensitivity curves presented in Figure 4 also shows that our approach performed consistently better than the two baselines.

To verify the generalizability of our approach (i.e. LR+LASSO on different feature sets and SVM for integrating the resulting predictions), we trained the model on the training set using the parameters determined with ten-fold cross-validation and evaluated it on the test set. The results are presented in Table 6. All the algorithms worked similarly on the training set, with only slight changes in the AUC performances. Our approach achieved significant improvement of 11 percent in AUC over the two baselines. The evaluation on an independent test set provided us strong evidence that our approach would work well on unseen data.

Table 2: Performance of Different algorithms on DynInfo (Note that PPV, Sensitivity and Specificity were reported according to the best F)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>F</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.162</td>
<td>0.358</td>
<td>0.250</td>
<td>0.849</td>
<td>0.604</td>
</tr>
<tr>
<td>+Top N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>0.163</td>
<td>0.518</td>
<td>0.248</td>
<td>0.732</td>
<td>0.640</td>
</tr>
<tr>
<td>+Top N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.297</td>
<td>0.381</td>
<td>0.334</td>
<td>0.909</td>
<td>0.750</td>
</tr>
<tr>
<td>+Top N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>0.245</td>
<td>0.376</td>
<td>0.297</td>
<td>0.884</td>
<td>0.691</td>
</tr>
<tr>
<td>+Top N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR</td>
<td>0.329</td>
<td>0.331</td>
<td>0.329</td>
<td>0.932</td>
<td>0.754</td>
</tr>
<tr>
<td>+LASSO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: The effect of the parameter lambda of LASSO on the AUC and number of selected features on DynInfo
Table 3: Performance of each individual classifier (Note that the threshold was selected according to the best F)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Threshold</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>F</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>9Var</td>
<td>0.2</td>
<td>0.300</td>
<td>0.327</td>
<td>0.313</td>
<td>0.923</td>
<td>0.742</td>
</tr>
<tr>
<td>ADInfo</td>
<td>0.2</td>
<td>0.313</td>
<td>0.330</td>
<td>0.321</td>
<td>0.927</td>
<td>0.745</td>
</tr>
<tr>
<td>DynInfo</td>
<td>0.2</td>
<td>0.329</td>
<td>0.331</td>
<td>0.329</td>
<td>0.932</td>
<td>0.754</td>
</tr>
<tr>
<td>AUC</td>
<td>0.2</td>
<td>0.318</td>
<td>0.281</td>
<td>0.299</td>
<td>0.939</td>
<td>0.712</td>
</tr>
<tr>
<td>PCUI</td>
<td>0.2</td>
<td>0.311</td>
<td>0.308</td>
<td>0.309</td>
<td>0.931</td>
<td>0.734</td>
</tr>
<tr>
<td>DCUI</td>
<td>0.2</td>
<td>0.317</td>
<td>0.281</td>
<td>0.298</td>
<td>0.928</td>
<td>0.728</td>
</tr>
<tr>
<td>LACE</td>
<td>9.0</td>
<td>0.206</td>
<td>0.378</td>
<td>0.267</td>
<td>0.903</td>
<td>0.681</td>
</tr>
<tr>
<td>Admissions</td>
<td>5.0</td>
<td>0.215</td>
<td>0.408</td>
<td>0.281</td>
<td>0.883</td>
<td>0.685</td>
</tr>
</tbody>
</table>

Table 4: Performance of integrating classifiers (Note that the threshold was selected according to the best F)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Threshold</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>F</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>9Var + DIfo</td>
<td>0.4</td>
<td>0.278</td>
<td>0.462</td>
<td>0.347</td>
<td>0.880</td>
<td>0.764</td>
</tr>
<tr>
<td>DynInfo</td>
<td>0.5</td>
<td>0.326</td>
<td>0.411</td>
<td>0.363</td>
<td>0.944</td>
<td>0.776</td>
</tr>
<tr>
<td>DCUI</td>
<td>0.5</td>
<td>0.326</td>
<td>0.411</td>
<td>0.363</td>
<td>0.944</td>
<td>0.776</td>
</tr>
<tr>
<td>LACE</td>
<td>9.0</td>
<td>0.206</td>
<td>0.378</td>
<td>0.267</td>
<td>0.903</td>
<td>0.681</td>
</tr>
<tr>
<td>Admissions</td>
<td>5.0</td>
<td>0.215</td>
<td>0.408</td>
<td>0.281</td>
<td>0.883</td>
<td>0.685</td>
</tr>
</tbody>
</table>

Table 5: Performance comparison (Note that the threshold was selected according to the best F)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Threshold</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>F</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our approach</td>
<td>0.5</td>
<td>0.326</td>
<td>0.411</td>
<td>0.363</td>
<td>0.944</td>
<td>0.776</td>
</tr>
<tr>
<td>LACE</td>
<td>9.0</td>
<td>0.206</td>
<td>0.378</td>
<td>0.267</td>
<td>0.903</td>
<td>0.681</td>
</tr>
<tr>
<td>Admissions</td>
<td>5.0</td>
<td>0.215</td>
<td>0.408</td>
<td>0.281</td>
<td>0.883</td>
<td>0.685</td>
</tr>
</tbody>
</table>

Table 6: Performance on test set (Note that the threshold was determined by the training set)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>PPV</th>
<th>Sensitivity</th>
<th>F</th>
<th>Specificity</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>9Var</td>
<td>0.317</td>
<td>0.356</td>
<td>0.335</td>
<td>0.915</td>
<td>0.741</td>
</tr>
<tr>
<td>ADInfo</td>
<td>0.340</td>
<td>0.352</td>
<td>0.346</td>
<td>0.924</td>
<td>0.738</td>
</tr>
<tr>
<td>DynInfo</td>
<td>0.362</td>
<td>0.343</td>
<td>0.352</td>
<td>0.933</td>
<td>0.745</td>
</tr>
<tr>
<td>AUC</td>
<td>0.333</td>
<td>0.324</td>
<td>0.329</td>
<td>0.928</td>
<td>0.706</td>
</tr>
<tr>
<td>PCUI</td>
<td>0.333</td>
<td>0.293</td>
<td>0.312</td>
<td>0.935</td>
<td>0.711</td>
</tr>
<tr>
<td>DCUI</td>
<td>0.358</td>
<td>0.308</td>
<td>0.331</td>
<td>0.939</td>
<td>0.721</td>
</tr>
<tr>
<td>Our approach</td>
<td>0.345</td>
<td>0.426</td>
<td>0.381</td>
<td>0.910</td>
<td>0.764</td>
</tr>
<tr>
<td>LACE</td>
<td>0.232</td>
<td>0.395</td>
<td>0.293</td>
<td>0.907</td>
<td>0.689</td>
</tr>
<tr>
<td>Admissions</td>
<td>0.231</td>
<td>0.442</td>
<td>0.303</td>
<td>0.872</td>
<td>0.684</td>
</tr>
</tbody>
</table>

VI DISCUSSION

As a major contribution of this study, we proposed a hybrid approach to predict readmission. In our approach, we created six feature sets, built classifiers on each feature set and integrated the predictions with an additional classifier. There are several considerations for employing this hybrid approach. First, the variables came from different sources such as structured EHR and clinical notes. Modeling them separately can provide deep insight of different sources and adapt to the diverse application requirements. For example, some applications with fast-response requirement may ignore the clinical notes that are not instantly available. Second, separating DynInfo from the commonly used ADInfo provided us a novel method to model the dynamic characteristics of clinical variables and it showed superior performance to ADInfo. Third, splitting clinical notes into three groups (i.e., admission notes, progress notes and discharge notes) allows one to consider the availability of the data in a pragmatic sense, in different stages during an encounter. Fourth, we created a unique classifier using the variables identified by previous studies. Besides the reasons described previously, we wanted to compare the automated models with the manually created model. The results showed that our automated model "DynInfo" performed significantly better than the manually created model. Meanwhile, this manually created classifier allows us to include more proxies in the future without much modification of the whole
process. Finally, the hybrid approach used in this study suggests that performance can be further improved by applying advanced model selection and integration techniques on more comprehensive classifiers in the future.

As shown in Table 3 and 6, the classifiers created on CUIs performed significantly better than the baseline methods, which demonstrated the promise of using clinical notes to predict readmission. However, the use of CUIs is just an initial exploration. There may be more predictive features in the clinical notes. Future work such as identifying predictive phrases and linkages needs to be conducted to improve the performance.

An extremely large number of candidate features were generated in this work. Selecting an optimal subset of these features was a major challenge and constitutes an important contribution of the study. Two different feature selection strategies were compared in Table 2. It shows that LASSO worked better than the Top N method since LASSO generated the best combination by solving an optimization problem other than heuristically used top N significant-most significant features. Meanwhile, using the same feature selection method (i.e., Top N), SVM performed better than NB, DT and LR since SVM has superior predictive ability compared to other methods. Using pre-filtering with significance test and LASSO, we obtained the optimal subset (several hundred) of features from an extremely large number (hundreds of thousands) of candidates. LR+LASSO performed statistically significantly better than all the other methods.

In this study, we used the internal categorizations provided by our dataset for medications, diagnoses and procedures instead of directly using the medication names, ICD9 codes or diagnosis names. According to our pre-study experiments, the performances of using categorizations were better than directly using the names. However, as Hosseinzadeh et al. [25] stated, there was a variety of conceptual level medical classification for medications, diagnoses, and medical-surgical procedures. Finely defined granularities for medications, diagnoses and procedures may be created and tested in the future.

A readmission reduction project has been started in our institution. We are developing a readmission prediction tool by implementing the algorithm created in this study. The readmission tool will assist our pediatric clinicians to perform timely interventions on inpatients at high risk for readmission, improve outcomes of inpatients and reduce readmission rate. However, the specific actions linked to the probabilities of readmission need to be investigated in prospective trials.

VII CONCLUSION

In this work, we studied the prediction of all-cause 30-day unplanned readmissions. We proposed a novel hybrid approach based on NLP and ML by automatically analyzing a large-scale EHR. Our approach achieved statistically significantly higher AUC (over 11 percent) than the two previously published methods used in many institutions.

VIII Acknowledgement

The authors were supported by internal funds from Cincinnati Children’s Hospital Medical Center. HZ, YN, TL, QL, HT, and IS were partially supported by grants 5R00LM010227-05, 1R21HD072883-01, and 1U01HG006828-01.

References

validated predictive algorithm. Open Medicine, 5(2), e104, 2011.


