Latent Topic Ensemble Learning for Hospital Readmission Cost Reduction

Christopher Baechle, Ankur Agarwal, Ravi Behara, Xingquan Zhu
Dept. of Computer Science & Engineering, Florida Atlantic University, Boca Raton, FL 33431, USA

***Abstract*—Unplanned hospital readmission is a costly problem in the United States. Patients treated and readmitted within 30 days cost tax payers up to $26 billion annually. In 2013 the U.S. federal government began to reduce payments to hospitals with excessive patient readmissions. Predictive modeling using machine learning can be a useful tool to help identify patients most likely to need readmission. However, current systems have several shortcomings. When creating predictive models for hospital readmission, existing methods either build models using data from a single hospital or naively combining data from multiple hospitals. Because hospitals often have different data distributions, models created from a single hospital's data are often biased. Additionally, models created from combined data overlook local data distributions. In this paper, we propose, LTEL, which uses an ensemble of topic specific models to leverage data from multiple hospitals. LTEL creates models based on latent topics derived from different hospitals. Models are built and evaluated incorporating federal financial penalties. The dataset contains data collected from 16 regional hospitals. Compared to baseline methods, LTEL significantly outperforms the best performing baseline method for cost reduction.**

Keywords— Ensemble learning; latent topics; hospital readmission

I. INTRODUCTION

The Center for Health Information and Analysis estimates unplanned hospital readmissions cost tax payers $26 billion annually [1]. Up to $17 billion of these readmissions may be avoidable. Recent federal legislation has brought many sweeping changes to the healthcare industry through modifications of Medicare reimbursement criteria. The Hospital Readmissions Reduction Program (HRRP) financially penalizes hospitals that incur high rates of unplanned Medicare patient readmission [2]. Many patients with chronic diseases poorly manage their illness and use emergency care services when symptoms become aggravated. Often these readmissions are preventable with proper post-discharge care and follow-up. The Centers for Medicare and Medicaid Services (CMS) has been tasked with identifying hospitals which have excessive rates of readmissions and providing target readmission rates consistent with expectations of the healthcare industry. As a result, medical facilities have begun to implement programs to reduce readmissions in order to lower financial penalties.

One potential solution would assign a home healthcare professional to all discharged hospital patients. The practical concerns of cost and staff availability make this plan unrealistic. This has caused many hospitals to look toward machine learning models for the identification of patients most likely to be readmitted. Patients are intelligently assigned post-discharge care based on predictive readmission models, thereby lowering readmission rates and optimally using limited resources. Although many systems exist to predict hospital readmission using computational approaches [3], they commonly suffer from several major shortcomings.

Localized Models: Current readmission models must be localized for performance reasons and combining readmission data amongst separate hospitals has yet to produce significantly better results [4]. Hospitals often differ in patient demographics resulting in disparate readmission rates and distribution of diseases. Naively incorporating data from many sources causes poor model performance. The resulting models are often biased due to differing class and feature distribution amongst available hospitals. Fig. 1 shows the vast difference in feature distributions for a large primary hospital plotted against a combined auxiliary dataset of 15 other available hospitals. A line overlain with an origin of (0,0) and slope=1 shows the ideal distribution of features. A primary and auxiliary dataset whose feature distribution closely follows this line could be considered for a naively merged dataset with little risk of decreased model performance. However, as shown in Fig. 1, this is clearly not the case.

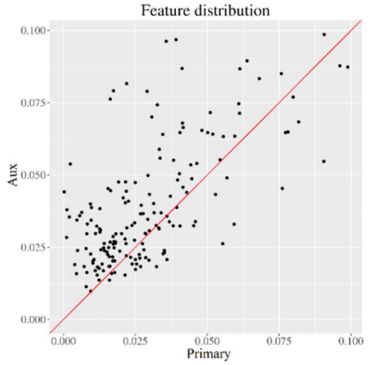

Fig. 1. Scatterplot of feature value distribution for a primary hospital *vs* all available source hospitals combined as a single dataset.

Many localized models only use a small portion of available data. For example, a model may be built using only instances that contain diabetes as a primary diagnosis. Such models result in better performance because diabetes patient's class distributions are better aligned than patients of differing diseases between hospitals. Creating a disease specific model may result in better performance at the cost of discarding many potentially useful instances. Additionally, patients may suffer from multiple diseases to varying degrees. Current models often assign the patient a primary diagnosis and use a single model for readmission classification. However, patients may belong to several models of varying degrees, requiring an ensemble of models to fully represent the patient.

While a good predictive system should have the ability to intelligently incorporate all available data into model creation and improve overall performance, no existing hospital readmission system has shown statistically significant

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

performance improvements when attempting to incorporate all available auxiliary data. In this paper, we propose to use Latent Dirichlet Allocation (LDA) topic modeling to find related groups of patients. For example, topic modeling may discover a group of patients suffering from substance abuse. Additionally, patients often belong to several groups and LDA assigns each instance a group membership weight. This allows all available data to be used to create many models.

Unstructured Data Sources: Few existing systems use unstructured clinical notes for model creation. Electronic Health Record (EHR) systems store patient data in two forms: structured and unstructured. Structured data may include demographic information such as age and gender. It may also include medical billing codes such as ICD9 or ICD10. Unstructured data often occurs in the form of clinical notes written in natural language by the attending medical staff. Much of this information is not be expressed in structured form, suggesting potentially valuable information is often not utilized. Our system has chosen unstructured data as the primary data source for these reasons.

More specifically, we represent clinical notes as a *bag-of-words* for predictive modeling. This representation is generally highly dimensional, so we use topic modeling, Latent Dirichlet Allocation LDA [5], to handle highly dimensional clinical notes and allow data from multiple sources to be carefully combined for readmission predictive modeling.

Readmission Costs: The demand for readmission systems largely relies on the assumption that hospitals want to reduce CMS penalties. Although penalty formulations and criteria have been made public, few researchers have investigated using cost when building models or as a readmission model evaluation performance metric. Cost sensitive learning [6], [7] has been previously studied, all existing solutions require the cost for False Positive (FP) and False Negative (FN) to be constant. However, CMS penalization formulas [2] are based on rates and always changing. This additional concern needs addressing to incorporate CMS penalties.

In summary, LTEL attempts to address all of these shortcomings. LTEL uses unstructured clinical notes with LDA to extract topics and assign instances to those topics. A single instance may belong to multiple topics and membership is represented as a weighted value. Models are created per-topic rather than per-hospital and new instances classified using a weighted membership ensemble with soft majority voting. Evaluation is performed using CMS cost formulas. Though many hospital readmission systems exist [3], [8]–[12], to our knowledge this is the first published system which addresses all of these concerns successfully.

II. Preliminary & Related works

A. Cost Sensitive Modeling and Evaluation

Many predictive hospital readmission models ignore misclassification cost and assume FP and FN misclassification cost to be equal. Cost sensitive modeling and evaluation assigns cost to misclassifications. Table II shows a confusion matrix which assigns misclassification cost, where μ is the cost of a FP and λ is the cost of a FN.

TABLE I. CONFUSION MATRIX

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

TABLE II. COST-SENSITIVE CONFUSION MATRIX

	Predicted Positive	**Predicted Negative**
Actual Positive	0	λ
Actual Negative	μ	0

Cost-sensitive data mining can be categorized into cost-sensitive learning, cost-sensitive classification, and cost-sensitive evaluation.

Cost-sensitive learning builds a predictive model using λ and μ misclassification cost variables. Accordingly, cost sensitive classification is defined by Eq. (1), where p_+ is probability of an instance (determined by the classifier) being positive and $\frac{\mu}{\lambda+\mu}$ represents the threshold of classification. When the threshold of classification is reached, the instance is classified as positive.

$$class = \begin{cases} positive, & if\ p_+ > \dfrac{\mu}{\lambda+\mu} \\ negative, & otherwise \end{cases} \quad (1)$$

According to HRRP, which dictates the criteria for excessive hospital readmissions and associated financial penalties [2], the base operating Diagnostic Related Group (DRG) payment amount and the Excess Readmission Ratio (ERR) are the two primary components of this calculation. The base operating DRG payment is calculated using many criteria, including case mix index, labor share, wage index, non-labor share, cost of living adjustments, technology payments, and total number of Medicare cases [13]. This research assumes these variables outside the scope of control for most hospitals.

The second component to readmission penalty is the ERR. ERR is defined as

$$ERR = \frac{Predicted\ readmissions\ rate}{Expected\ readmissions\ rate} \quad (2)$$

Expected readmission rate is the expected rate of readmission given the hospital's patient population. CMS determines expected readmission rate using regression models based on national readmission statistics [2]. Predicted readmission rate is related to the actual readmission rate. However, this rate incorporates the hospital's risk adjusted readmission statistics as to minimize uncontrollable risk factors. Predicted readmissions rate is the rate of hospital readmissions for which a given hospital is responsible. Expected readmission rate can be treated as uncontrollable because hospitals have little control over their patient population and demographics. Predicted readmissions rate may be improved by hospitals through the reduction of readmissions. The equation for 30-day all-cause readmission

penalty for a given DRG is defined as Eq. (3), where DRG is the sum of payments made for the group and ERR is the Excess Readmission Ratio for the group.

$$DRG(ERR - 1) \quad (3)$$

B. Clinical Natural Language Processing

Recent advances have chosen clinical notes as a primary data source for readmission prediction [10], [14]. Clinical notes are often written to be directly read by other medical staff in order to facilitate further treatment of the patient or understand the patient's medical history. This means that clinical notes will often contain important information that may not be encoded as an ICD code. Structured data such as ICD codes may be difficult to extract from an EHR and require additional IT involvement. Clinical notes are often easily exported as Microsoft Word or text documents.

The Clinical Text Analysis and Extraction System (cTAKES) was created by researchers at the Mayo clinic beginning in 2006 to annotate clinical notes [15]. cTAKES handles clinical NLP tasks such as annotation of diseases, medications, and symptoms. cTAKES utilizes the Unified Medical Language System (UMLS). UMLS provides a normalized database of several medical vocabularies and dictionaries. ICD-9, SNOMED-CT, NCI Thesaurus, MeSH, and RxNorm dictionaries are enabled in cTAKES by default.

C. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised machine learning algorithm which attempts to group a set of training instances into topics. LDA is often used in the field of text mining and can offer insights into the structure and relationships of data. A predefined number of topics must be provided as a parameter to LDA and often requires experimentation as to the number of optimal topics for a given dataset. Additionally, topics may be difficult to interpret because latent relationships exist and require a domain expert to assign semantic meaning. LDA has been used extensively in the field of text mining, but has seen little use in text mining of clinical notes and no published work as of yet has used LDA as a method for improving hospital readmission model performance.

D. Auxiliary Data

Often times hospital readmission data from multiple medical facilities is available when creating a new readmission model. Intuitively, one might attempt to combine all available data into a single classifier. However, classification distribution, patient population, and data entry methods vary amongst hospitals. Naively combining all available instances will often result in a model with poor results.

The primary hospital is the hospital which contains instances for which classification is desired. Auxiliary data from available hospitals contains instances which we would like to use to strengthen the performance of the primary hospital. The primary and auxiliary hospitals often have different classification distributions and distribution among features. An algorithm which uses instances from auxiliary hospitals to build a predictive model on the primary hospital must address this or many times the result is a model which performs worse than a model which used only data from the primary hospital. When this occurs it is known as model degradation.

E. Hospital Readmission Prevention

Hospital readmission reduction programs can be divided into two categories: (1) patient education, follow-up, and outreach and (2) data mining techniques to identify those most likely in need of readmission. This research focuses on the second method.

Predictive modeling may be used to identify which patients are most likely to need readmission and to spend a disproportionate amount of resources trying to intervene. This level of intervention may not be practical for every patient. However, due to the high financial penalty imposed by CMS for excess readmissions, the benefit will often exceed the cost. High quality statistical models which are able to predict hospital readmissions with a high discriminative ability have been an area of active research since the introduction of the HRRP. However, progress has been slow and many models have been published which have relatively similar predictive power [3].

Many approaches to statistical patient readmission prediction exist. General approaches which are not hospital specific are constructed using clinical research and expert knowledge. A model known as LACE is a popular index created to analyze readmission risk [8]. This model was created by analyzing almost 5,000 patients over the period of four years. The resulting 30-day readmission rate for those patients was 8%. Researchers extracted four categories of variables which independently explained readmission rates. Length of stay (L), acuity of admission (A), comorbidity (C), and emergency room frequency (E) were found to explain a great deal of readmissions. Unlike many other computationally intensive models, LACE is easy to calculate and easy to interpret. The primary evaluation metric for this research was the c-statistic. LACE obtained a c-statistic of 0.7 using the data in the original study, which can often be considered to have moderate discriminative ability. Studies which use LACE on different datasets are often unable to obtain a c-statistic near 0.7 [11], and can be as low as 0.55 [12]. Other models which use similar methods also perform poorly [17].

Research shows models improve significantly when they are institution or disease specific, rather than a single model [9]. Increases of almost 20% in c-statistic were possible by using this methodology. Although these models often perform better than LACE, additional Information Technology (IT) infrastructure may be required extract, load, and transform data. Extracting structured data from Electronic Health Record (EHR) systems can be labor intensive and require software experts to facilitate. Structured data in EHR systems is often in the form of ICD-9 and ICD-10 billing codes which are assigned for reimbursement purposes and may not convey a complete picture of the patient. Due to these practical concerns, LACE remains extremely popular in the clinical setting. Using clinical notes as a primary data source may be considerably cheaper and less time consuming. This may

potentially increase adoption of machine learning based readmission models.

III. LTEL Algorithm

In this section, we first explain the cost calculation as per CMS formula, which is indeed a much more complicated procedure than existing cost-sensitive learning methods consider. Then we will explain the detailed LTEL framework.

A. CMS Cost

CMS calculates penalties based on Eq. (3) and is the foundation of our cost evaluation model. Table III outlines definitions for variables used in our cost sensitive model. The individual components of the DRG amount is outside the scope of our model and treated as a single cost variable. Many components of DRG amount are difficult or impractical to influence administratively. Expected rate is calculated based on national readmission statistics and is also treated as a single variable with which the medical facility has no control. However, the predicted rate can be improved. CMS uses regression modeling to determine the number of readmissions with which a hospital is held responsible.

TABLE III. MODEL VARIABLES

C	Total cost of Diagnostic Related Group (DRG)
C_{np}	Total cost of DRG for new patient(s) under analysis
ω	Risk adjustment factor
R	Number of readmissions for current fiscal period
R_{fn}	Number of FN readmissions in new patient analysis
P	Number of total patients in DRG for current fiscal period
ρ	Expected rate
$\hat{\rho}$	Predicted rate
p_r	Probability of needing readmission
N	Number of new patients under consideration

$$cost_{cms} = C * \left(\frac{\omega\hat{\rho}}{\rho} - 1\right) \quad (4)$$

In the case of cost sensitive classification for hospital readmission, we can define μ as the cost of intervention. Intervention is often in the form of a home health care nurse. The patient would not have been readmitted to the hospital, therefore any investment towards preventing readmission is lost. λ is defined as the cost the hospital will incur due to increased readmission rates. The misclassification cost for a single isolated FN instance is defined in Eq. (5).

$$\lambda = \left(C + C_{np}\right)\left(\frac{\omega\left(\frac{R+1}{P+1}\right)}{\rho} - 1\right) - cost_{cms} \quad (5)$$

The total cost of FN misclassification is not the sum of all individual FN misclassifications, however. CMS calculates cost using ERR, as defined in Eq. (2). The increase in readmissions are the patients which needed but did not receive intervention. The increase in the total number of patients are the number of patients evaluated. The difference between total FN evaluator misclassification performance and the currently calculated CMS cost is represented by λ_{total}. Eq. (6) represents this total cost. The total misclassification cost for new patients under analysis is the sum of λ_{total} and μ_{total} as shown in eq 8.

$$\lambda_{total} = (C + C_{np})\left(\frac{\omega\left(\frac{R+R_{fn}}{P+N}\right)}{\rho} - 1\right) - cost_{cms} \quad (6)$$

$$\mu_{total} = \sum \mu \quad (7)$$

$$cost_{total} = \lambda_{total} + \mu_{total} \quad (8)$$

B. Cost-sensitive classification

Patient readmission prediction can occur during two time periods: (1) During the discharge process and (2) Nightly batch processing. Predicting patient readmission probability during discharge is advantageous in that the patient is still in close contact with the facility. Intervention can be setup via scheduling of a home health care nurse or other medical professional. When a patient has left the facility, contacting them again may be difficult.

Nightly batch processing loses the advantage of immediate patient contact, but gains the advantage of a non-fixed FN cost. Since a batch of instances are available, λ can be updated during each classification iteration. For each iteration, the number of patients increases by one, but the number of readmissions increases by the probability of the current patient resulting in readmission. Eq. (9) defines an updatable λ, which may more accurately represent cost over the previously proposed method.

$$\lambda_n = \left(C + C_{np}\right)\left(\frac{\omega\left(\frac{R+\sum_{i=1}^{n} p_r}{P+n}\right)}{\rho} - 1\right) - \sum_{i=0}^{n-1} \lambda_i \quad (9)$$

$$\lambda_0 = cost_{cms} \quad (10)$$

1) A Running Example for CMS Cost Calculation

Given the following starting assumptions: C is $10,000,000; ω is 1; R is 202; P is 1,000; ρ is .200; $\hat{\rho}$ is .202, first, the base ERR and cost must be calculated.

$$ERR_{cms} = \frac{\omega\hat{\rho}}{\rho} - 1 = \frac{.202}{.200} - 1 = 0.01$$

$$cost_{cms} = C * \left(\frac{\omega\hat{\rho}}{\rho} - 1\right)$$
$$= C * ERR_{cms}$$
$$= \$10{,}000{,}000 * 0.01 = \$100{,}000$$

This represents the current cost in CMS penalties. These are patients for which ground truth label can be applied as the 30-day readmission window has passed. At this point, the classification status of these patients cannot be affected through intervention or otherwise.

If a single new patient were presented and patient is a readmission, the cost of the single patient is as follows. The cost of each patient treatment for this DRG is assumed to be C_{np} = $10,000.

$$ERR_1 = \frac{\omega\left(\frac{R+1}{P+1}\right)}{\rho} - 1 = \frac{\left(\frac{202+1}{1000+1}\right)}{.200} - 1 = 0.013986$$

$$\lambda_1 = (C + C_{np})(ERR_1) - cost_{cms}$$
$$= (\$10{,}010{,}000)(0.01398) - \$100{,}000 = \$40{,}000$$

The cost of this patient arriving and later needing readmission is $40,000. When analyzed in isolation, the cost of a single readmission will cost this amount. However, when calculating the subsequent cost of additional readmissions, the ground truth label of this readmission may not be known for up to 30 days. A more accurate representation of the potential CMS penalty would use Eq. (9). Using predictive modeling, a probability of readmission can be assigned until the ground truth label is known for this instance. This may lead to a closer cost estimate rather than simply assuming all new readmission instances to cost $40,000. Assuming the instance was classified with $p_r = 0.6$ and n = 1 new instances, the previous example could then be rewritten as the following.

$$ERR_1 = \frac{\omega\left(\frac{R+p_r}{P+n}\right)}{\rho} - 1 = \frac{\left(\frac{202+.6}{1000+1}\right)}{.200} - 1 = 0.01198$$

$$\lambda_1 = (C + C_{np})ERR_1 - cost_{cms}$$
$$= (\$10{,}010{,}000)(0.01198) - \$100{,}000 = \$20{,}000$$

When the next patient is ready have misclassification cost calculated, it is assumed we do not know the ground truth classification of the first patient. Assuming the model reported $p_r = 0.8$ for the next patient, $\sum_{i=1}^{n} p_r = 1.4$ and n = 2.

$$ERR_2 = \frac{\omega\left(\frac{R+\sum_{i=1}^{n} p_r}{P+n}\right)}{\rho} - 1 = \frac{\left(\frac{202+1.4}{1000+2}\right)}{.200} - 1$$
$$= 0.01497$$

$$\lambda_2 = (C + C_{np})ERR_2 - (\lambda_1 + cost_{cms})$$
$$= (\$10{,}020{,}000)(0.01497) - (\$120{,}000) = \$30{,}000$$

When the third patient is ready have misclassification cost calculated, assume the readmission model reported $p_r = 0.9$, $\sum_{i=1}^{n} p_r = 2.3$, and n = 3.

$$ERR_3 = \frac{\omega\left(\frac{R+\sum_{i=1}^{n} p_r}{P+n}\right)}{\rho} - 1 = \frac{\left(\frac{202+2.3}{1000+3}\right)}{.200} - 1$$
$$= 0.01844$$

$$\lambda_3 = (C + C_{np})ERR_3 - (\lambda_2 + \lambda_1 + cost_{cms})$$
$$= (\$10{,}030{,}000)(0.01844) - (\$150{,}000) = \$35{,}000$$

As ground truth labels become available, these calculations may be updated with different starting assumptions to reflect the new information that has become available.

C. LTEL Framework

The LTEL algorithm is described below. First, the classifiers are trained. Topics are created using the combined primary training and auxiliary instances. Each instance is then assigned a membership weight to each topic. These weights are between [0,1] and sum to 1 for each instance. A classifier is then built using all instances and weights. Instances with a greater weight have a proportionally greater influence over classifier creation. The number of trained classifiers and topics are equal and have 1-to-1 relationship.

Algorithm 1 LTEL Model Training Process

procedure LTEL_TRAIN(D_{prm},D_{aux}, k)
▷ D_{prm}: Primary hospital data; D_{aux} data from other hospitals;
▷ k: number of latent topics;
  $D_{train} \leftarrow D_{prm} \cup D_{aux}$ ▷ Combining primary and auxiliary data;
  $T \leftarrow LDA(D_{train}, k)$; ▷ Find latent topics;
  for instance $x_i \in D_{train}$ **do** ▷ Find instance-topic weight;
    $x_i^w \leftarrow Instance\ Topic\ Weight(T, x_i)$
    $D_{train}^w = D_{train}^w \cup x_i^w$
  end for
  for topic $t_j \in T$ **do** ▷ Learn topic specific classifiers;
    $D_{train}^{w_{t_j}} \leftarrow Weighted\ Instances\ to\ Topic\ (D_{train}^w, t_j)$;
    $C^{t_j} \leftarrow Train\ Classifier(D_{train}^{w_{t_j}})$ ▷ Topic specific classifier;
  end for
  return (C, T)
end procedure

When an unseen instance needs classification, topic membership weights for that instance are calculated using the previously discovered topics. Posterior probability is found using the previously trained classifiers, weighted by the instance's membership of that topic. FN misclassification cost (λ) is then calculated using either fixed or updatable cost equations and the instance is then classified using Eq. (1).

Algorithm 2 LTEL classification algorithm

▷ Classify unseen test instances
procedure LTEL_CLASSIFY(D_{test}, C, T, μ)
▷ D_{test}: Primary hospital test data; ▷ C: Trained classifier ensemble;
▷ T: Latent topics; ▷ μ: FP misclassification cost;
  for instance $x_i \in D_{test}$ **do**
    $x_i^w \leftarrow LDA(T, x_i)$ ▷ Find LDA topic weights for instance
    for Topic Classifier $C^{t_j} \in C$ **do**
      $\Sigma_+ \leftarrow x_i^{w_{t_j}} * Posterior(x_i, C^{t_j})$ ▷ Sum of weighted posterior
    end for
    $\lambda \leftarrow \lambda(\ldots)$; ▷ Update λ using Eq. (9)
    if $\Sigma_+ > \frac{\mu}{\lambda+\mu}$ **then** ▷ Classify using Eq. (1)
      $\hat{y}_i \leftarrow (+)$
    else
      $\hat{y}_i \leftarrow (-)$
    end if
  end for
  return R ▷ Class predictions of D_{test}
end procedure

Algorithm 3 LTEL performance evaluation

▷ Performance evaluation of test instances
procedure LTEL_EVALUATE(D_{test}, R)
▷ D_{test}: Primary hospital test data;
▷ R: Classified results from LTEL_Classify;
  $R_{fp} \leftarrow 0$; $R_{fn} \leftarrow 0$
  $X \leftarrow D_{test} \cup R$ ▷ Merge predictions and labels
  for instance $x_i \in X$ **do**
    ▷ Compare ground truth *vs.* predicted label
    if $\hat{y}_i == (+)$ *and* $y_i == (-)$ **then**
      $R_{fp} \leftarrow R_{fp} + 1$
    else if $\hat{y}_i == (-)$ *and* $y_i == (+)$ **then**
      $R_{fn} \leftarrow R_{fn} + 1$
    end if
  end for
  $\lambda_{total} \leftarrow \lambda(\ldots)$; ▷ Calculate FN cost using Eq. (6)
  $\mu_{total} \leftarrow \mu(\ldots)$; ▷ Calculate FP cost using Eq. (7)
  return $\lambda_{total} + \mu_{total}$; ▷ Calculate total cost using Eq. (8)
end procedure

Performance evaluation compares the ground truth labels to the classification predictions, calculating cost using CMS equations based on whether the classification is correct. As FP and FN have different cost implications, these are tracked separately. The total misclassification cost is then reported.

IV. EXPERIMENTS

A. Benchmark Data

TABLE IV. DESCRIPTION OF ALL HOSPITALS

Hospital	Instances	Readmission Rate
A	15991	0.0958
B	193	0.0000
C	342	0.0146
D	1056	0.0643
E	13589	0.0884
F	82	0.0243
G	1983	0.0927
H	2172	0.0465
I	3767	0.0501
J	8698	0.0756
K	209	0.0095
L	3704	0.0534
M	6790	0.0602
N	2222	0.0567
O	57	0.0175
P	1859	0.0268

The dataset for this research contains data collected from 16 regional hospitals. Data is split using 10-fold cross validation. Hospitals vary greatly in size, patient demographics, and readmission rates as shown in Table IV. An additional challenge presented by this dataset is several hospitals only have limited data available. Rather than discard hospitals with limited data, these are included as auxiliary instances when training LTEL.

TABLE V. TOPICS DISCOVERED BY LDA FOR ALL HOSPITALS.

Topic	Important Terms	Description
1	Hypertension; Penicillin; Asthmas; Accident	Combination of diseases, drugs, and terms.
2	Hypertension; Diabetes; Coronary Disease; Plavix Lopressor	Heart disease.
3	COPD; Lung Disease; Albuterol; Advair	Diseases and medications related to chronic lung disease.
4	Hypertension; Heart Attack; Aspirin	Heart attack.
5	Cancer; Liver Diseases; Percocet	Combination of diseases and pain medication.
6	Percocet; Bone Fracture; Arthritis; Vicodin	Bone diseases and pain medications.
7	Heart Fibrillation Congestive Heart Failure Coronary Disease	Heart conditions known to co-occur.
8	Communicable Disease Urinary Tract Infection Vancomycin	Bacterial infections and anti-biotics due to communicable disease.
9	Kidney Disease; Diabetes; Hypertension	Combination of diseases.
10	Alcohol Abuse; Hepatitis; Liver Disease; Drug Habituation	Liver diseases related to drug and alcohol abuse.

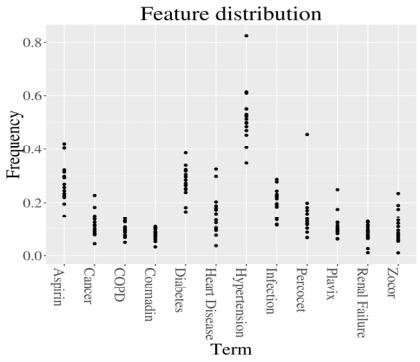

Fig. 2. Distribution of Common Diseases for each Hospital

B. Cost

For comparison purposes, all hospitals were given initial starting assumptions. The fixed misclassification cost of a FP was $800. After consulting with domain experts, this was found to be a reasonable average cost for sending a home healthcare professional to a patient's home for several hours and a possible second visit. Expected readmissions rate was set to .14 and current predicted readmissions rate was .143. The fixed FN misclassification cost was then calculated to be $61,428.60 assuming an average of 1,000 ground truth labels being available.

C. Baseline Methods

In our experiments, we implemented the following baseline methods for comparisons.

Primary: This method uses only the data from the primary hospital to train the classifier for readmission prediction.

Prm+Aux: This method aggregates all data from the primary hospital and the auxiliary hospitals to form one dataset. A single model is then trained from the aggregated dataset.

Bagging: This method is an ensemble learning approach which treats each hospital separately. A classifier is trained from each single hospital, and all classifiers equally vote to predict a test instance.

TrAdaBoost: This method is a popular transfer learning algorithm addressing differences in distribution [18]. TrAdaBoost is a modification of the AdaBoost algorithm which creates an ensemble of classifiers and uses a weighted voting mechanism to classify instances. AdaBoost works by creating so-called expert classifiers that perform well on a certain subset of training data. Each classifier is weight based upon how many instances that expert can correctly classify. TrAdaBoost iteratively builds an ensemble of classifiers, lowering the weights of diff-distribution instances when incorrectly predicted during the construction phase of the AdaBoost algorithm. The assumption is that these instances are too different from the target domain and offer little usable information.

LTEL: This represents the proposed method which uses latent topic ensemble learning for hospital readmission prediction.

Fig. 3 plots AUC against fixed cost evaluation using hospitals containing at least 2,500 instances and NB classifier across 10-folds. As shown in the scatter plot there is low correlation between AUC and cost (correlation = -0.51), suggesting may not be a good metric for this domain when misclassification has been made available.

C. D. Feature Extraction

In our experiments, discharge summaries are annotated using Apache cTAKES. Annotations containing diseases & disorders, medications, and anatomical site are used. Annotations are normalized to a UMLS CID to increase the quality of features. The corpus contains 7,112 extracted features and non-cTAKES features are not included. Features are represented using the *bag-of-words* model. Instances with multiple occurrences of the same CID are limited to a single representation thereby limiting each feature to the binary values {0,1}.

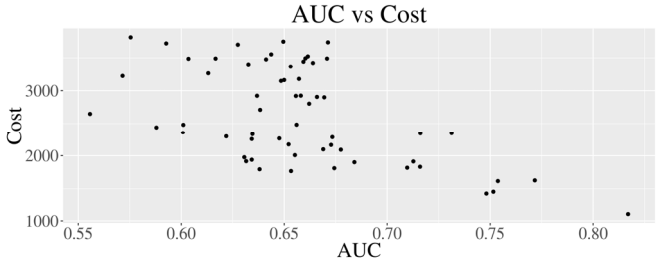

Fig. 3. Scatterplot showing the weak correlation between AUC and cost.

D. Learning Algorithms

The algorithms chosen for this research are Naïve Bayes (NB), k-nearest neighbors (kNN), Linear Regression (LR), and Support Vector Machine (SVM). Weka 3.6 is used for the implementations of all algorithms. NB is computationally quick and performs well for highly dimensional data. kNN uses distance functions to find the most similar instances to the instance under consideration. Similar to NB, kNN is computationally fast. LR forms a linear decision boundary between instances. SVM forms a similar linear boundary, but attempts to maximize the margin between classes.

E. Experimental Results

Table VI shows that the base classifier with the lowest average cost using fixed cost methodology is LR. However, in practice NB may still be a good choice as it is considerably faster than LR. Additionally, the results were not statistically significant ($p = 0.37$) and therefore possible NB to be as good base classifier.

TABLE VI. COMPARISON OF FIXED MISCLASSIFICATION COST AVERAGED OVER 10 FOLDS FOR EACH BASE CLASSIFIER USING LTEL FOR PRIMARY HOSPITAL A.

Classifier	Cost
NB	796.36
kNN	929.08
LR	770.59
SVM	5885.19

Fig. 4 shows baseline methods plotted against LTEL. Each point represents a hospital and cross-validation fold cost. NB was the only classifier tested due to time and computational resource limitations. NB builds classifiers relatively quickly and Table VI provides evidence that NB is an acceptable choice. Hospitals containing more than 2,500 instances are plotted, however all available instances are used in building classifiers to maximize data usage. Hospitals containing fewer than 2,500 instances often did not contain enough positive instances to reliably perform 10-fold cross validation. A line starting at origin (0,0) with slope=1 overlays the plot. Points below the plot have a higher baseline cost when compared to LTEL and are considered to have performed better than the baseline methodology.

TABLE VII. COMPARISON OF BASE CLASSIFIERS AND BASELINE METHODOLOGIES FOR HOSPITAL A USING UPDATABLE COST. NOTE THAT NEGATIVE COST MEANS THE MODEL EXCEEDED CMS EXPECTATIONS.

		Method				
		Primary	**Prm+Aux**	**TrAdaBoost**	**Bagging**	**LTEL**
Classifier	**NB**	-7493	-7284	-5997	-7320	-7998
	kNN	-5152	-5330	-4611	-5396	-5633
	LR	-6278	-6412	-3849	-6382	-6341
	SVM	-3187	-3165	-4612	-3165	-3156

LTEL performed better than the baseline methods in most instances. In the few instances LTEL did not perform better than primary baseline, results are often close to the linear overlay, suggesting little performance degradation occurred. LTEL performed better than TrAdaBoost for all data points, suggesting that LTEL is more appropriate than a popular existing transfer learning method for this domain.

The updatable cost methodology showed similar gains as fixed cost. Compared to primary, LTEL often had lower cost. For the data points where cost was not lower, often times the point was very close to the linear overlay, suggesting performance degradation to be a relatively rare occurrence. All data points were located below the linear overlay for TrAdaBoost, as was the case with fixed cost classification. Table VII shows a comparison of classifiers and baseline methodologies for hospital A. When compared to all available methods and base classifiers, LTEL using NB has the lowest cost. These results were statistically significant ($p < 0.01$). NB has additional desirable qualities such as fast classification performance. A hospital which implements the LTEL system using NB can potentially significantly lower CMS penalties when compared to other known methods.

V. Conclusions

In this paper, we proposed a latent topic based ensemble learning framework. We argued that when building a hospital re-admission prediction model, the data distributions across different hospitals vary significantly. Existing methods often combine all data together, or select a small subset of samples from all available data, which result in biased or ineffective models. Alternatively, we proposed LTEL to derive latent topics from different hospitals, and use topics to align data across hospitals and determine weight accordingly. The

weighted instances from different hospitals are then combined to build classifiers to predict instances from a primary hospital. The experiments and validation from data collected from 16 regional hospitals demonstrated significant cost reduction compared to best performing baseline available.

VI. References

[1] S. Reardon, "Preventable Readmissions Cost CMS $17 Billion," 2015. [Online]. Available: http://bit.ly/1nL8k7g. [Accessed: 11-Oct-2016].

[2] Centers for Medicare and Medicaid Services, "Readmissions Reduction Program," 2014. [Online]. Available: http://go.cms.gov/1gLbnoa. [Accessed: 15-Jun-2015].

[3] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani, "CLINICIAN ' S CORNER Risk Prediction Models for Hospital Readmission A Systematic Review," *Jama*, vol. 306, no. 15, pp. 1688–1698, 2011.

[4] L. Turgeman and J. H. May, "A mixed-ensemble model for hospital readmission," *Artif. Intell. Med.*, vol. 72, pp. 72–82, 2016.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[6] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, 2001, vol. 17, no. 1, pp. 973–978.

[7] B. Krawczyk, "Cost-sensitive one-vs-one ensemble for multi-class imbalanced data," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, 2016, pp. 2447–2452.

[8] C. van Walraven, I. A. Dhalla, C. Bell, E. Etchells, I. G. Stiell, K. Zarnke, P. C. Austin, and A. J. Forster, "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community," *Can. Med. Assoc. J.*, vol. 182, no. 6, pp. 551–557, 2010.

[9] S. Yu, F. Farooq, A. van Esbroeck, G. Fung, V. Anand, and B. Krishnapuram, "Predicting readmission risk with institution-specific prediction models," *Artif. Intell. Med.*, vol. 65, no. 2, pp. 89–96, 2015.

[10] R. Duggal, S. Shukla, S. Chandra, B. Shukla, and S. K. Khatri, "Predictive risk modelling for early hospital readmission of patients with diabetes in India," *Int. J. Diabetes Dev. Ctries.*, 2016.

[11] H. Wang, R. D. Robinson, C. Johnson, N. R. Zenarosa, R. D. Jayswal, J. Keithley, and K. A. Delaney, "Using the LACE index to predict hospital readmissions in congestive heart failure patients," *BMC Cardiovasc. Disord.*, vol. 14, no. 1, pp. 1–8, 2014.

[12] P. E. Cotter, V. K. Bhalla, S. J. Wallis, and R. W. S. Biram, "Predicting readmissions: Poor performance of the LACE index in an older UK population," *Age Ageing*, vol. 41, no. 6, pp. 784–789, 2012.

[13] J. Hoffman, "Overview of CMS Readmissions Penalties for 2016," 2015. [Online]. Available: http://www.besler.com/2016-readmissions-penalties/. [Accessed: 25-Sep-2016].

[14] A. Agarwal, R. S. Behara, S. Mulpura, and V. Tyagi, "Domain Independent Natural Language Processing -- A Case Study for Hospital Readmission with COPD," *2014 IEEE Int. Conf. Bioinforma. Bioeng.*, pp. 399–404, 2014.

[15] G. K. Savova, J. J. Masanz, P. V Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *J. Am. Med. Informatics Assoc.*, vol. 17, no. 5, pp. 507–513, 2010.

[16] D. Goodman, E. Fisher, and C. Chang, "The Revolving Door: A Report on US Hospital Readmissions," *Princeton, NJ Robert Wood Johnson Found.*, 2013.

[17] P. S. Keenan, S. L. T. Normand, Z. Lin, E. E. Drye, K. R. Bhat, J. S. Ross, J. D. Schuur, B. D. Stauffer, S. M. Bernheim, A. J. Epstein, Y. Wang, J. Herrin, J. Chen, J. J. Federer, J. A. Mattera, Y. Wang, and H. M. Krumholz, "An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure," *Circ. Cardiovasc. Qual. Outcomes*, vol. 1, no. 1, pp. 29–37, 2008.

[18] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 193–200.

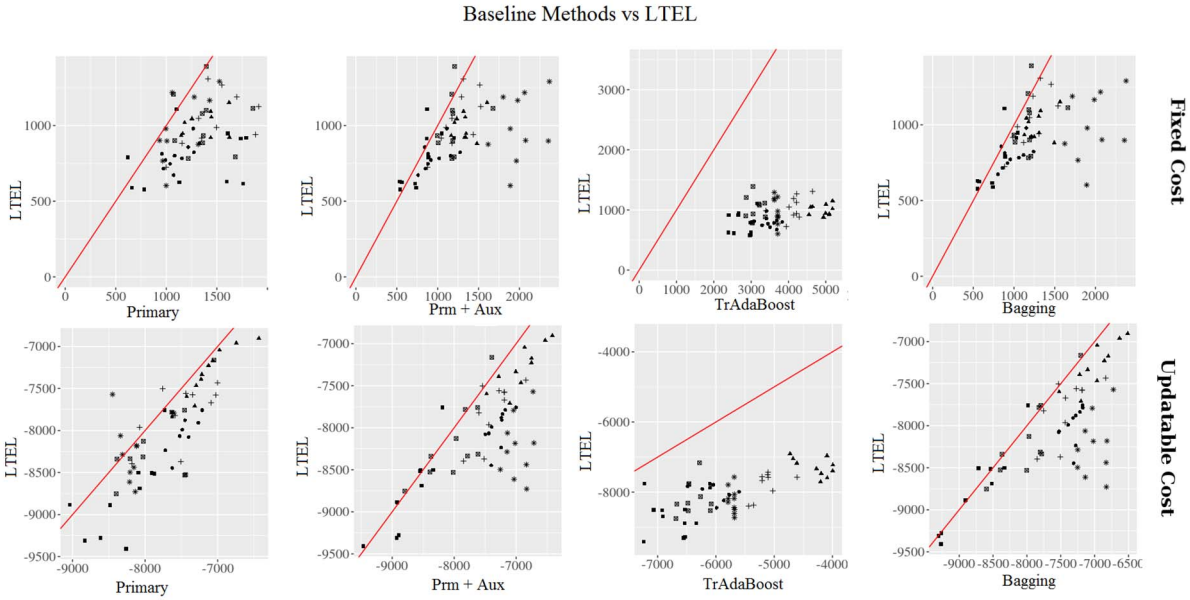

Fig. 4. Scatterplots of LTEL compared to baseline methods for hospital A using NB. A linear overlay with origin (0,0) and slope=1 is shown. Points below this line have a lower LTEL cost than the comparable baseline method. Due to axis scaling this line may not appear 45 degrees. Note that negative cost means the model exceeded CMS expectations.