Baechle and Agarwal *J Big Data* (2017) 4:37
DOI 10.1186/s40537-017-0098-z

Journal of Big Data

RESEARCH Open Access

A framework for the estimation and reduction of hospital readmission penalties using predictive analytics

Christopher Baechle* and Ankur Agarwal

*Correspondence:
cbaechle@fau.edu
Department of Computer & Electrical Engineering and Computer Science, College of Engineering, Florida Atlantic University, Boca Raton, FL, USA

Abstract

Background: Recent US legislation imposes financial penalties on hospitals with excessive patient readmissions. Predictive analytics for hospital readmissions have seen an increase in research due to the passage of this legislation. However, many current systems ignore the formulas used by the Centers for Medicare and Medicaid Services for imposing penalties. This research expands upon current methodologies and directly incorporates federal penalization formulas when selecting patients for which to dedicate resources.

Methods: Hospital discharge summaries are structured using clinical natural language processing techniques. Naïve Bayes classifiers are then used to assign a probability of readmission to each patient. Hospital Readmission Reductions Program formulas and probability of readmission are applied using four readmission scenarios to estimate the cost of readmission. The highest cost patients are identified and readmission mitigation efforts are attempted.

Results: The results show that the average penalty savings over currently employed binary classification to be 51.93%. Binary classification is also shown to select more patients than necessary for readmission intervention. Additionally, intervening in only high-risk patients saved an average of 90.07% compared to providing all patients with costly aftercare.

Conclusion: Focusing resources toward the potentially most expensive patients offers considerably better results than unfocused efforts. Utilizing direct calculation to estimate readmission costs has shown to be a more efficient use of resources than current readmission reduction methods.

Keywords: Scientific algorithms of big data, Big data applications, Big data tools, Natural language processing, Naïve Bayes classification

Introduction

Millions of Americans are admitted to hospitals each year. A significant number of these individuals are readmitted within 30-days, and many of those readmissions are avoidable. The Center for Health Information and Analysis estimates unplanned readmissions to cost $26 billion annually [1]. Recent US legislation has begun to penalize hospitals which have excess readmissions. The Hospital Readmission Reduction Program (HRRP) aims to reduce payments to hospitals which have an excess of avoidable readmissions

[2]. The Centers for Medicare and Medicaid Services (CMS) are tasked with identifying avoidable readmissions and penalizing hospitals using a set of defined formulas. These penalties often exceed potential reimbursement and have motivated hospitals to work towards readmission reduction [3].

Many strategies for reducing unplanned readmissions exist. One potential strategy is to provide patient education and follow-up. This method can be applied equally to all patients. Researchers at a US hospital found that patients often do not fill prescription medications prescribed during hospital visits [4]. To address this problem, patients are now encouraged to have their prescriptions filled directly at hospital pharmacies. Researchers found this method to drastically increases patient compliance. Although simple methods have been found to be effective, the number of hospitals penalized for excess readmissions has held steady for several years [5]. Clearly, there is a need for more research into reducing unplanned readmissions.

Patients often require a home healthcare professional to largely mitigate the risk of all cause 30-day readmission. Ideally, all patients would receive a home healthcare professional after hospital discharge. This would be prohibitively expensive and an ineffective use of limited resources, so this is not a feasible option. A more effective use of resources would be through identification of potential patient readmissions using statistical analysis. This has been an area of active research since the introduction of HRRP [6]. Statistical analysis allows resources to be used more effectively. Current systems often produce binary classification and are not well suited for Decision Support Systems (DSS). If resources are available to send a home healthcare professional to a single patient, but there are two patients classified as potential readmissions, additional information is required. A probability of readmission may be desirable in this instance. However, binary classification systems often do not inherently offer this additional information.

Many current systems are additionally known to have poor discriminative ability. LACE, a popular readmission system, has been found to produce a c-statistic as low as 0.55 [7]. Systems trained using localized hospital data often fare better, but rarely produce c-statistic scores greater than 0.7 [6]. Although c-statistic is a popular statistical measure for Hospital Readmission Prediction Systems (HRPS), it may not be the most appropriate. C-statistic traditionally assumes equal misclassification cost, which is rarely true for hospital readmission. When cost formulas and criteria are available, the quality of the model can be evaluated using these formulas directly rather than using a proxy measure. Research from Baechle et al. [8] found the correlation between c-statistic and hospital readmission cost to be low ($\text{cor} = -0.21$). C-statistic serves as a poor proxy for cost, yet few researchers have incorporated cost into readmission models [9, 10].

Our proposed HRPS directly incorporates HRRP cost formulas. HRRP penalties are not constant per patient, but instead based on a rate. Sorting patients by probability of readmission allows hospitals to choose a threshold for which to intervene based on available resources. Hospitals attempting to reduce HRRP penalties may use target readmission rates set by CMS to decide how to allocate resources. Exceeding CMS target rates does not result in a refund or negative penalty [2]. Probability of readmission and potential costs are presented, forming the basis of a Clinical Decision Support System (CDSS). This methodology allows hospital staff to incorporate domain knowledge into the final

determination of resource allocation. The proposed methodology, MinCost, may allow hospitals to optimize available resources to reduce HRRP penalties.

Background

Current HRPS utilizing statistically driven methods fall into two categories. The first category are systems which build hospital agnostic models. These models are built once and may predict readmission for any patient at any hospital. A popular system by researchers at Yale University is freely available online at readmissionscore.org. This system consists of a simple questionnaire to be completed by medical staff. Using c-statistic as the primary evaluation metric, the benchmark score for heart failure (HF) patients is 0.61 [11]. This system is considered to have poor discriminative ability, but has been popular with clinical staff due to its simplicity. LACE is a similar readmission scoring system which accounts for length of stay (L), acuity of admission (A), comorbidity (C), and emergency room frequency (E) [12]. Although LACE initially showed improvements over systems devised by Yale (c-statistic $= 0.7$), additional research has found this model to vary in quality when presented with differing hospitals. Few researchers have obtained a c-statistic near 0.7 [13] and some researchers report results as low as 0.55 [7]. Clearly, models created given a set of data and assumptions do not perform well when those assumptions change.

A second expanding area of research uses machine learning models tailored to each hospital [14]. While these methods may require more work to implement, they often perform better due to localization issues. Feature distribution may differ greatly among hospitals and localizing models eliminates those concerns.

Machine learning

Many machine learning algorithms have been used for the creation of HRPS. Logistic regression (LR) is a regression model whose dependent variable is categorical. Boulding et al. [15] have used LR to predict readmission using patient satisfaction as independent variables. Greenwald et al. [16] also successfully used LR for predicting readmission using physical function, cognitive status, and psychosocial support. Support vector machines (SVM) are another binary linear classifier which attempt to maximize the margins of classification. Research by Braga et al. used SVM to predict readmission for intensive care patients, while Sushmita et al. have used SVM for prediction of all-cause readmission [9, 17]. Decision trees have also been successfully used for the prediction of patient readmission [18]. Naïve Bayes (NB) classifiers are simple probabilistic classifiers which use Bayes' theorem to classify instances. NB assumes conditional independence between features. Although this assumption is often not true, NB remains a useful classifier and is often used in text classification [19]. Researchers using unstructured text as a data source have seen good results using NB for predicting readmission [10, 20].

HRPS utilizing machine learning often produce a binary classification. Algorithms use training data to create a model which will classify new instances as either readmission or non-readmission, based on the evidence provided. This may be problematic, as patients with a high risk of readmission may be assigned the same classification prediction as those whose readmission risk is considerably lower. If resources are available to send a home healthcare professional to a single patient, but there are two patients classified as

potential readmissions, additional information is required. However, binary classification systems often do not inherently offer additional information such as probability of readmission. This shortcoming limits the use of a HRPS within a CDSS.

Performance evaluation

The most common evaluation metric of HRPS performance is c-statistic [6]. C-statistic is defined as the area under a receiver operating characteristic (ROC) curve. ROC curves are graphical plots that illustrate the performance of a binary classifier as its discrimination threshold is varied. The plot compares performance of the true positive rate (TPR) and false positive rate (FPR) for various thresholds of classification. A survey by Kansagara et al. [6] reviewed many HRPS using c-statistic as the primary performance metric and few perform better than 0.7. Small data samples are often cited as a primary reason for poor model performance. However, models created using The United Kingdom's National Health Service (NHS), using millions of patients, [21] performed similarly to a model created using 1029 patients [22].

The current use of c-statistic as an evaluation metric for HRPS has shown to produce inconsistent results. Researchers have argued that c-statistic is often used inappropriately to measure the performance of classification systems [23]. Assumptions about misclassification cost and uniform distribution within classes may not be true and comparison of classification systems using c-statistic may produce incoherent results [24]. Although cost as a performance metric may offer a clear alternative to c-statistic, its use in HRPS has been limited.

Data sources

Data sources for readmission models can be categorized into structured and unstructured data [25]. Structured data is generally stored in a relational database and contains information such as demographics and ICD-9 or ICD-10 codes. Current systems described by Kansagara et al. [6] commonly use structured data as the primary data source. The main advantage to using structured data is that once extracted from a system, little work needs to be done to convert the data into a form usable by supervised machine learning algorithms. Unstructured data is often represented in the form of clinical notes or discharge summaries. These are often written in natural language such as English and allow the medical professional to completely describe their thoughts regarding patient status. Systems using structured data may potentially lose information if there is no predefined input for an observation. However, unstructured data is often difficult to convert to a format usable by supervised machine learning algorithms. The field of natural language processing (NLP) can often be of assistance in structuring natural language to a usable format. Although unstructured data as a primary data source for hospital readmission systems has historically seen little adoption, advances in clinical NLP have helped increase adoption among researchers [10, 20, 26–29].

Clinical NLP software

Apache cTAKES is an open source Clinical NLP tool created and maintained by the Mayo Clinic [30]. Apache cTAKES annotates clinical notes using domain specific dictionaries and clinically trained NLP models. Core to cTAKES is the Unified Medical

Language System (UMLS), a set of dictionaries and vocabularies maintained by the National Library of Medicine (NLM) [31]. UMLS allows cTAKES to annotate notes using vocabularies assembled by domain experts and enables cTAKES to separate clinical and non-clinical terms. Many diseases, symptoms, and medications have variations in spelling, abbreviations, and usage. UMLS provides a normalization ID known as the Concept ID (CID) which allows terms to be reduced to their base components. Many variants of the same feature will often confuse machine learning models and merging terms that have the same semantic meaning strengthens the model and increases performance. Figure 1 illustrates many lexical variants of asthma normalized to a single CID.

Methodology

Data

The data for this research consists of 1248 discharge summaries containing chronic obstructive pulmonary disease (COPD) as a primary diagnosis. Unstructured data was chosen as a primary data source, due to the ease of extraction from Electronic Health Record (EHR) systems. Discharge summaries and classification labels were available as Microsoft Word documents. Apache cTAKES was used to annotate discharge summaries and normalize medications, diseases, and symptoms to a UMLS CID. Figure 2 illustrates the most frequently occurring terms in this dataset. A total of 1018 UMLS normalized terms have been identified and extracted from this dataset using Apache cTAKES.

Annotations are converted to a *bag-of-words* representation, where each annotation in the corpus is a feature and the presence of an annotation in an instance is the value of that feature. Given the discharge summaries. *The patient has asthma and diabetes* and *Jane Doe was found to have COPD and asthma*, Table 1 illustrates the bag-of-words representation using medical terms discovered by cTAKES. In this example, the classification label is true for the first instance (the patient was readmitted within 30 days) and false for the second instance (patient was not readmitted within 30 days).

The bag-of-words representation of unstructured text is inherently unable to detect the presence of a missing value. A missing value in the context of a clinical note would

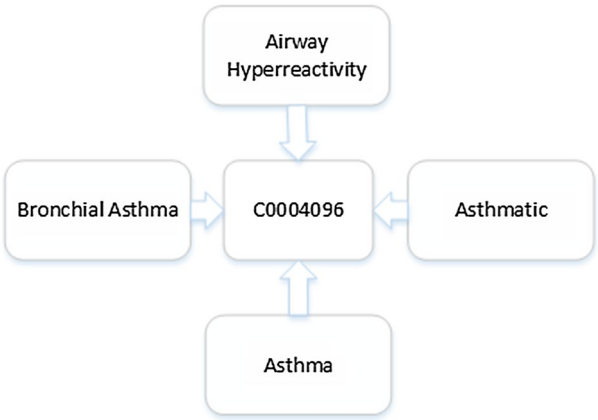

Fig. 1 Illustration of similar terms normalizing to a UMLS Concept ID

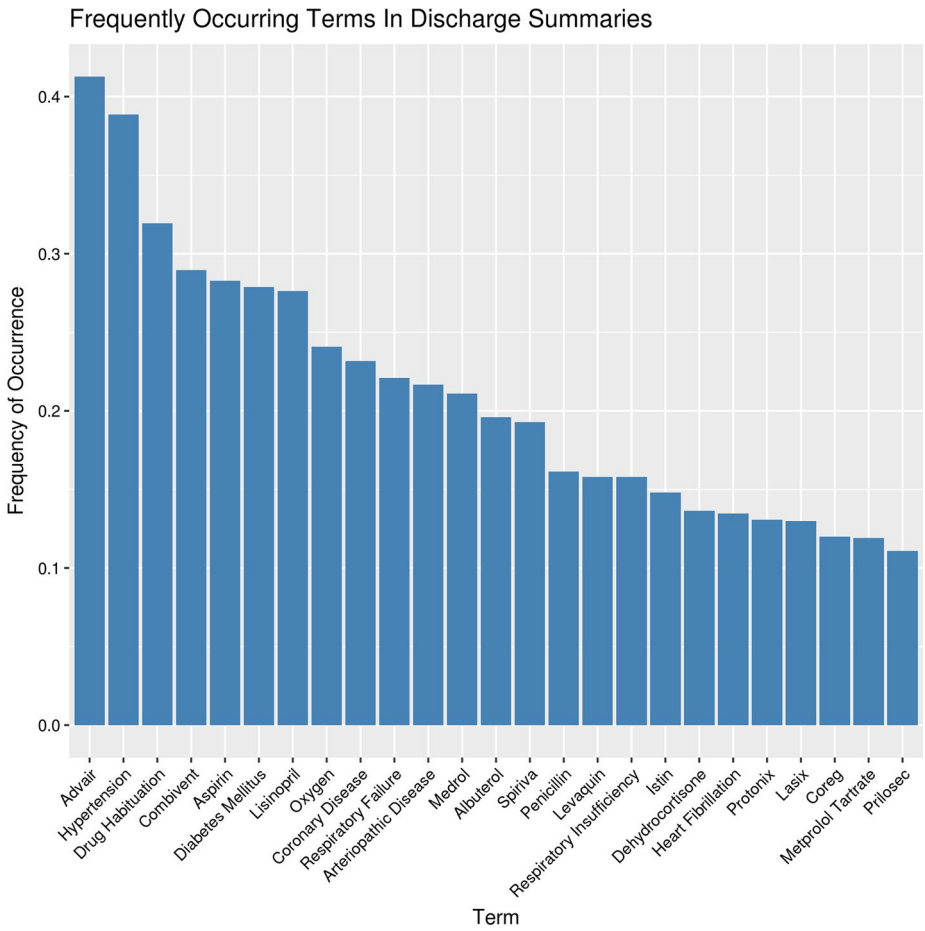

Fig. 2 Frequently occurring terms in discharge summaries

Table 1 Example bag-of-words representation

	Asthma	Diabetes	COPD	Readmission
Discharge summary #1	1	1	0	True
Discharge summary #2	1	0	1	False

imply a medical professional omitted text (either intentionally or intentionally). The value of a term which is present is be represented as the number one and the value of a term which is not present is represented as the number zero. The omission of a term in unstructured text is assumed to be purposeful and missing values, whether intentional or unintentional, are represented by the number zero.

Machine learning

Many HRPS use binary classification labels when predicting readmission [9, 14, 17, 32–35]. However, staffing resources are often limited and a probability of readmission is potentially more useful. Sorting a group of patients by readmission probability allows patients with the highest readmission probability to be allotted the greatest number of

resources. Though many supervised machine learning algorithms have the ability to coerce classification distribution, the NB machine learning algorithm naturally produces posterior probabilities without coercion [19]. Therefore, NB can be used to sort patients by likelihood of readmission. This is useful when resources are limited and an optimal subset of patients must be selected to minimize penalties.

Additional classifiers were considered, but early experimentation often resulted in models which classified all instances as positive. This may be due to the high cost of readmission or resulting coercion of posterior probabilities. Since these classifiers offered no useful predictive abilities, NB was chosen as the primary classifier.

HRRP cost

CMS has made available the formulas for which HRRP penalties are calculated [2]. The cost consists of two primary components: Diagnostic Related Group (DRG) amount and excess readmissions ratio (ERR). DRG is calculated using many variables, including case mix index, labor share, wage index, non-labor share, cost of living adjustments, technology payments, and total number of Medicare cases. Few of these variables can be affected administratively and for modeling purposes considered unchangeable. ERR is defined as follows

$$ERR = \frac{Predicted\ readmissions\ rate}{Expected\ readmissions\ rate} - 1. \tag{1}$$

Expected readmissions rate is the target readmissions rate which CMS has assigned a given hospital. Using national readmission statistics and regression models, this is the rate which the average hospital would obtain, given a hospital's patient demographics and disease distribution. Predicted readmissions rate is related to the actual readmissions rate obtained by a hospital. A risk adjustment is performed and predicted readmissions rate is the product of the actual rate and risk adjustment. Expected readmissions rate is considered constant as it is set by CMS. However, predicted readmissions rate can be lowered by lowering the actual readmission rate. The final penalty for a given DRG is calculated by CMS as follows

$$\text{Cost} = (DRG)(ERR). \tag{2}$$

Modeling cost

Table 2 defines the variables used in cost modeling.

ERR can be expanded and risk adjustment factored out. Risk adjustment is the fraction of actual readmissions between [0,1] for which a hospital is responsible.

$$\text{Cost} = (C)\left(\frac{\omega\hat{\rho}}{\rho} - 1\right) \tag{3}$$

Each hospital has a set of patients for which the ground truth readmissions status is known. This is due either to 30-days having lapsed or the patient having experienced readmission. These patients are denoted P and R respectively. By definition, $\hat{\rho} = \frac{R}{P}$ and can be expanded in our equation.

$$\text{Cost} = (C)\left(\frac{\omega\left(\frac{R}{P}\right)}{\rho} - 1\right) \tag{4}$$

Table 2 Variable definitions in cost model

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
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
N_s	Number of new patients to select for intervention
p_s	Probability of intervention success
$\bar{c}_t$	Average cost of patient intervention (i.e. home healthcare professional)

When a new patient is added to our model and the ground truth readmission status is known to be a readmission, R increases by 1, P increases by 1, and the DRG increases by the cost of that patient. The expected rate ρ remains unaffected as this is set by CMS. The total cost is then calculated as follows.

$$\text{Cost} = \left(C + C_{np}\right)\left(\frac{\omega\left(\frac{R+1}{P+1}\right)}{\rho} - 1\right) \tag{5}$$

As new patients enter the hospital, the ground truth readmission label will not be known for up to 30 days. However, an estimate can be calculated using the probability of readmission. Therefore, a new patient entering the current cost estimation model can be assigned a probability of readmission p_r.

$$\text{Cost} = \left(C + C_{np}\right)\left(\frac{\omega\left(\frac{R+p_r}{P+1}\right)}{\rho} - 1\right) \tag{6}$$

Additional patients may be admitted before ground truth readmission status is known for previous patients. As patients enter the hospital, the model increases by N patients and the sum of readmission probabilities $\sum_{i=1}^{N} p_r$. The final cost estimation for HRRP penalties is below:

$$\text{Cost} = \left(C + C_{np}\right)\left(\frac{\omega\left(\frac{R+\sum_{i=1}^{N} p_r}{P+N}\right)}{\rho} - 1\right). \tag{7}$$

ERR may be useful in some instances and is defined by:

$$ERR = \frac{\omega\left(\frac{R+\sum_{i=1}^{N} p_r}{P+N}\right)}{\rho} - 1. \tag{8}$$

Optimal patient intervention

Given a set of N patients, it is desirable to choose the smallest subset for which to send a home healthcare professional. This analysis can often be done as a nightly batch process to gather a sufficiently large number of patients for calculation. Before processing begins, patients must have readmission probability assigned using NB, then sorted by probability of readmission. Patients with the highest probability of readmission are analyzed first.

The net cost of patient intervention must include the possibility that intervention may not work. Internal statistics regarding the effectiveness of patient intervention may be collected and is represented by p_s. Assuming N patients, of which we intervene for N_s, we can define $ERR^{(N_s)}$ as follows:

$$ERR^{(N_s)} = \frac{\omega\left(\frac{R+(1-p_s)N_s+\sum_{i=N_s}^{N} p_r^{(i)}}{P+N}\right)}{\rho} - 1 \tag{9}$$

where $(1 - p_s)N_s$ represents the probable number of readmissions that will still occur even though a home healthcare professional has been assigned. N_s can be increased iteratively until either ERR ≤ 0 or there are no additional patients to analyze. If ERR ≤ 0, no additional HRRP penalties are incurred and resources for preventing readmission can be diverted elsewhere. Figure 3 describes the process for finding the optimal number of patients in which to select for intervention.

Cost estimates may additionally incorporate the average cost of a home healthcare professional, represented as $\bar{c}_t$. Assuming N_s interventions, the current total cost of intervention will be the following:

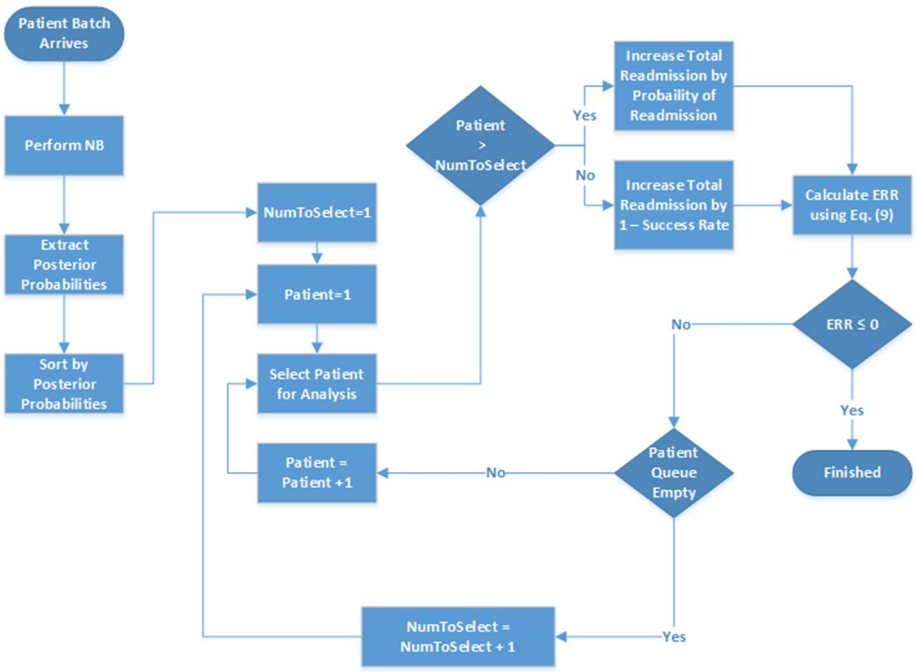

Fig. 3 Flowchart of selecting optimal number of patients for which to provide post-discharge care

$$\text{Cost}^{(N_s)} = (C + C_{np})\left(ERR^{(N_s)}\right) + \bar{c}_t N_s. \tag{10}$$

This represents the total cost of intervention given N patients and N_s interventions. Since the patients have been sorted by decreasing order of readmission probability and number of patients in analysis is constant, it is possible to iteratively increase the number of patients to intervene until a minimum cost is achieved. When $ERR^{(N_s)} \leq 0$, no additional cost savings are possible as CMS does not refund medical facilities for exceeding expected rates. No penalties will be incurred, but the cost of sending a home healthcare professional remains. If the cost of $\bar{c}_t$ is very high or set of patients very unlikely to require readmission, a minimum cost where $ERR^{(N_s)} > 0$ is possible. This scenario indicates that patients are unlikely to need readmission and cost of intervention high. In this case, it is less expensive to pay HRRP penalties than to intervene. Due to the generally high cost of HRRP penalties, this scenario is rare.

Framework

An overview of the framework is shown in Fig. 4. Discharge summaries are gathered and sent to Apache cTAKES for annotation. Once annotated, features are extracted using the bag-of-words representation. This representation allows discharge summaries to be used with the NB machine learning algorithm, which predicts the probability of readmission. These probabilities are then sent to the MinCost algorithm which attempts to identify the costliest patients.

Results

Two baseline methodologies were chosen for comparison. The first baseline uses the NB classifier ignoring cost and performing traditional classification. Most systems reviewed by Kansagara et al. use a similar method of classification which ignores cost. The second baseline method assumes to intervene on all patients using a home healthcare professional. This method is known as all-intervention (AI). Stratified tenfold cross validation is performed on all comparative methodologies. Cost is reported as total cost for each stratified fold analysis, not per patient. Due to sampling methods, these costs are relative and meant to be compared with baseline methods. These are not to be taken as absolute costs of a typical hospital.

Obtaining a low penalty by intervening in a small number of patients is desired. This allows limited staffing resources to be used elsewhere. Success rate of intervention and other starting assumptions are shown in Tables 3 and 4. These starting assumptions were reached using domain expertise of typical home healthcare costs and the estimated

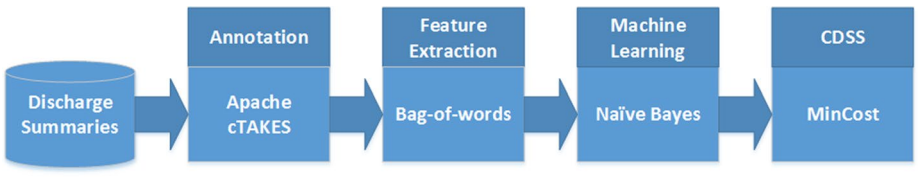

Fig. 4 Block diagram of framework

Table 3 Initial variable assumptions for all scenarios

C	\$10,000,000
C_{np}	\$10,000 * N
ω	1
P	1000
$\bar{c}_t$	\$800

Table 4 Variable values used for each assumption scenario

Assumption	R	ρ	$\hat{\rho}$	p_s
A	143	0.14	0.143	0.90
B	143	0.14	0.143	0.97
C	205	0.20	0.205	0.90
D	205	0.20	0.205	0.97

readmission rates from two hospitals with differing patient demographics. These starting assumptions have been used successfully in previous research [8].

As shown in Fig. 5, baseline classification selects many more patients than necessary for readmission intervention. Average ERR for MinCost is − 0.001, however average ERR for baseline classification is − 0.04. Many patients for baseline classification would have received a home healthcare professional, while not actually lowering penalties. Compared to binary NB classification, MinCost significantly lowers net cost when all factors are taken into consideration (shown in Fig. 6). These results are statistically significant using a paired t test, where $p < 0.01$ in all instances. The AI baseline methodology is shown in Table 5 to have significantly larger costs than all other methods. Table 6 illustrates the cost savings of MinCost vs baseline methods. The average penalty for MinCost is 51.93% lower than classification and 90.07% lower than AI. Assumptions C and D are shown to have the greatest cost savings, suggesting that hospitals with high readmission rates may benefit most from MinCost.

In some cases, reaching a zero ERR may not be possible due to a high initial ERR or small number of new patients under analysis. Assumption A is modified to use a high initial ERR ($\hat{\rho} = 0.148$) and under this assumption, Fig. 7 shows NB classification to stop classifying patients for intervention far before optimal. When reaching a zero ERR is not possible, it may be most reasonable to send follow-up care to all or most of those patients in the DRG due to high costs of penalty. In this case, MinCost is reduced to the AI baseline. In practice, a medical facility may choose to initially only intervene in extremely high risk patients, while accumulating a pool of medium-to-high risk patients for calculation.

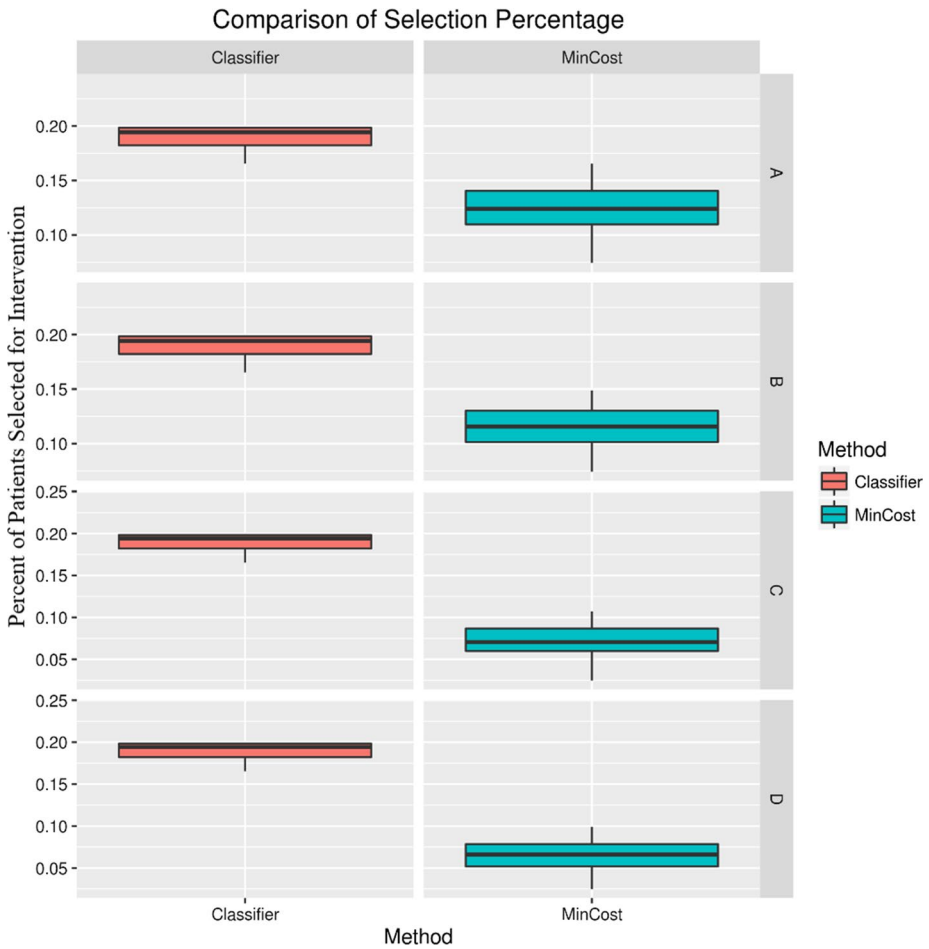

Fig. 5 Comparison of MinCost and binary classification patient selection percentage for each assumption scenario

Conclusions

Our system for minimizing HRRP penalties has shown that simply using binary readmission classification systems is often not sufficient. Though readmission classification systems may provide some insight to medical facilities with no statistical readmission reduction strategy in place, integrating cost into machine learning models has shown to significantly reduce cost by optimally selecting only those patients in greatest need of intervention. Our system also gives additional control and insight to staff in determining which patients will receive valuable resources. Probability of readmission and potential cost are often more useful than binary class labels which lack the ability to prioritize. Future work intends to further analyze the effect of cost analysis during various stages of care and improve patient readmission probability models.

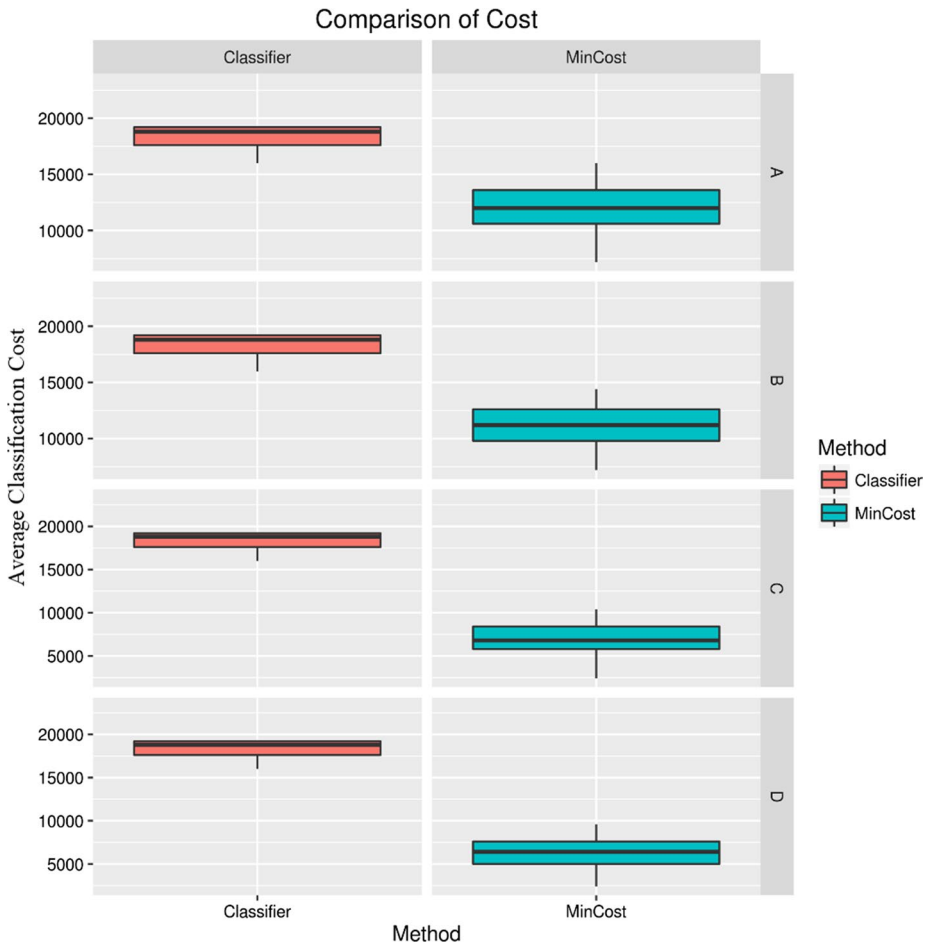

Fig. 6 Comparison of MinCost and binary classification patient cost for each assumption scenario

Table 5 Cost results averaged over tenfold for each assumption scenario

Assumption	MinCost	Classification	AI
A	$11,920	$18,640	$96,720
B	$11,040	$18,640	$96,720
C	$6720	$18,640	$96,720
D	$6160	$18,640	$96,720

Table 6 Percentage cost difference for MinCost vs baseline methodologies

Assumption	Classification (%)	AI (%)
A	− 36.05	− 87.67
B	− 40.77	− 88.58
C	− 63.94	− 93.05
D	− 66.95	− 93.63
Average	− 51.93	− 90.07

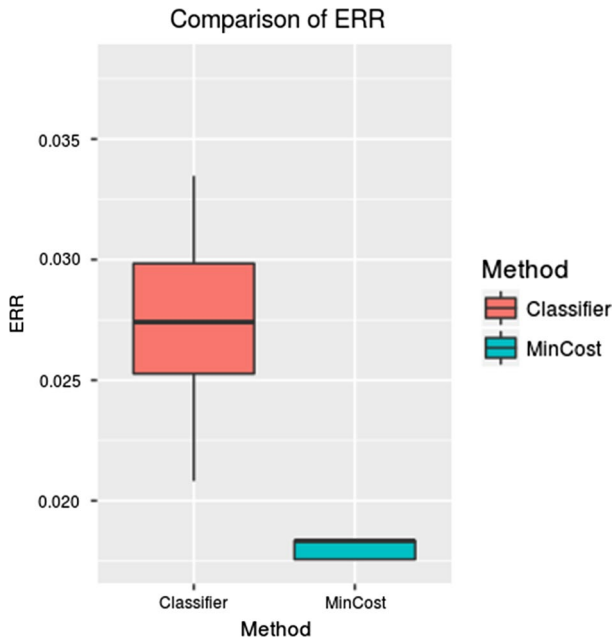

Fig. 7 Comparison of final ERR when initial ERR very high

Authors' contributions
CB carried out the conception, design, and implementation of this research as well as interpretation of results. AA made substantial contributions to the conception and design of this research as well as critically reviewing and interpreting results. CB carried out the drafting of manuscript. AA critically reviewed the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
This research was supported in part by NSF Grants IIP-1444949 and IIP-1624497.

Publisher's Note

Received: 22 August 2017 Accepted: 19 October 2017
Published online: 02 November 2017

References

1. Reardon S. Preventable readmissions cost CMS $17 Billion. 2015. http://bit.ly/1nL8k7g. Accessed 11 Oct 2016.
2. Centers for Medicare and Medicaid Services. Readmissions reduction program. 2014. http://go.cms.gov/1gLbnoa. Accessed 15 Jun 2015.
3. Hoffman J. Overview of CMS readmissions penalties for 2016. 2015. http://www.besler.com/2016-readmissions-penalties/. Accessed 25 Sep 2016.
4. Goodman D, Fisher E, Chang C. The revolving door: a report on US Hospital readmissions. Robert Wood Johnson Found.: Princeton; 2013.

5. Boccuti C, Casillas G. Aiming for fewer hospital U-turns: the Medicare Hospital Readmission Reduction Program. 2017. http://kff.org/medicare/issue-brief/aiming-for-fewer-hospital-u-turns-the-medicare-hospital-readmission-reduction-program/. Accessed 01 Apr 2017.
6. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S. CLINICIAN' S CORNER risk prediction models for hospital readmission a systematic review. JAMA. 2011;306(15):1688–98.
7. Cotter PE, Bhalla VK, Wallis SJ, Biram RWS. Predicting readmissions: poor performance of the LACE index in an older UK population. Age Ageing. 2012;41(6):784–9.
8. Baechle C, Agarwal A, Behara R, Zhu X. A cost sensitive approach to predicting 30-day hospital readmission in COPD patients. In: 2017 IEEE EMBS International conference on Biomedical & Health Informatics (BHI), p. 317–20. 2017.
9. Sushmita S, Khulbe G, Hasan A, Newman S, Ravindra P, Roy SB, De Cock M, Teredesai A. Predicting 30-day risk and cost of 'all-cause' hospital readmissions. In: Expand boundaries health inform using AI, p. 453–61. 2015.
10. Duggal R, Shukla S, Chandra S, Shukla B, Khatri SK. Predictive risk modelling for early hospital readmission of patients with diabetes in India. Int J Diabetes Dev Ctries. 2016;36(4):519–28.
11. Keenan PS, Normand SLT, Lin Z, Drye EE, Bhat KR, Ross JS, Schuur JD, Stauffer BD, Bernheim SM, Epstein AJ, Wang Y, Herrin J, Chen J, Federer JJ, Mattera JA, Wang Y, Krumholz HM. 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. 2008;1(1):29–37.
12. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, Austin PC, Forster AJ. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Can Med Assoc J. 2010;182(6):551–7.
13. Wang H, Robinson RD, Johnson C, Zenarosa NR, Jayswal RD, Keithley J, Delaney KA. Using the LACE index to predict hospital readmissions in congestive heart failure patients. BMC Cardiovasc Disord. 2014;14(1):1–8.
14. Yu S, Farooq F, van Esbroeck A, Fung G, Anand V, Krishnapuram B. Predicting readmission risk with institution-specific prediction models. Artif Intell Med. 2015;65(2):89–96.
15. Boulding W, Glickman SW, Manary MP, Schulman KA, Staelin R. Relationship between patient satisfaction with inpatient care and hospital readmission within 30 days. Am J Manag Care. 2011;17(1):41–8.
16. Greenwald JL, Cronin PR, Carballo V. A novel model for predicting rehospitalization risk incorporating physical function, cognitive status, and psychosocial support using natural language processing. Med Care. 2016;0:1–6.
17. Braga P, Portela F, Santos MF, Rua F. Data mining models to predict patient's readmission in intensive care units. In: ICAART 2014-Proceedings of the 6th International conference on agents and artificial intelligence, vol. 1, p. 604–10. 2014.
18. Futoma J, Morris J, Lucas J. A comparison of models for predicting early hospital readmissions. J Biomed Inform. 2015;56:229–38.
19. Lewis DD. Naive (Bayes) at forty: the independence assumption in information retrieval. In: European conference on machine learning 1998, p. 4–15.
20. Agarwal A, Baechle C, Behara R, Zhu X. A Natural language processing framework for assessing hospital readmissions for patients with COPD. IEEE J Biomed Health Inform. 2017;PP(99):1–1. doi:10.1109/JBHI.2017.2684121
21. Bottle A, Aylin P, Majeed A. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. J R Soc Med. 2006;99(8):406–14.
22. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA, Zhang S, Reed WG, Swanson TS, Ma Y, Halm EA. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care. 2010;48(11):981–8.
23. Hand DJ, Anagnostopoulos C. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognit Lett. 2013;34(5):492–5.
24. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn. 2009;77(1):103–23.
25. Agarwal A, Behara RS, Mulpura S, Tyagi V. Domain independent natural language processing—a case study for hospital readmission with COPD. In: 2014 IEEE international conference on bioinformatics and bioengineering (BIBE), p. 399–404. 2014.
26. Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro VM, McCoy TH, Perlis RH. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Transl Psychiatry. 2016;6(10):e921.
27. Topaz M, Radhakrishnan K, Blackley S, Lei V, Lai K, Zhou L. Studying associations between heart failure self-management and rehospitalizations using natural language processing. West J Nurs Res. 2017;39(1):147–65.
28. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an Electronic Health Record Study. PLoS ONE. 2015;10(8):1–11.
29. Evans RS, Benuzillo J, Horne BD, Lloyd JF, Bradshaw A, Budge D, Rasmusson KD, Roberts C, Buckway J, Geer N, Garrett T, Lapp DL. Automated identification and predictive tools to help identify high-risk heart failure patients: Pilot evaluation. J Am Med Inform Assoc. 2016;23(5):872–8.
30. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.
31. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl 1):D267–70.
32. Turgeman L, May JH. A mixed-ensemble model for hospital readmission. Artif Intell Med. 2016;72:72–82.
33. Yang C, Delcher C, Shenkman E, Ranka S. Predicting 30-day all-cause readmissions from hospital inpatient discharge data. In: 2016 IEEE 18th International conference on e-Health networking, applications and services (Healthcom), p. 1–6. 2016.
34. Choudhry SA, Li J, Davis D, Erdmann C, Sikka R, Sutariya B. A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. Online J Public Heal Inf. 2013;5(2):219.
35. Hosseinzadeh A, Izadi MT, Verma A, Precup D, Buckeridge DL. Assessing the predictability of hospital readmission using machine learning. In: Proc twenty-fifth innov appl artif intell conf assess, p. 1532–8. 2013.