Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Comparison of different Machine Learning Models for Predicting Chronic Obstructive Pulmonary Disorder Hospital Readmissions

Mr. Abhinav Parameshwaran, Mr. Paritosh Kadam, Mr. Adarsh Gadekar, Mr. Kevin Sherla
Student,
Department of Computer Engineering,
Vidyalankar Institute of Technology, University of Mumbai, India

***Abstract*: Chronic Obstructive Pulmonary Disease (COPD) is a prevalent chronic pulmonary condition that affects hundreds of millions of people all over the world. Being progressive in nature, Chronic Obstructive Pulmonary Disease (COPD) patients require frequent hospital readmission. Readmission can be avoided if additional attention is paid to patients with high readmission risk. Machine learning (ML) based prediction models proved to be fast, accurate, and free from human errors with capabilities to address pressing problems in healthcare. In this research we compare the relative performance of different modeling paradigms to find the best model for this task.**

Keywords : COPD, hospitalization, readmission, Planned care, Integrated care, predictive models

I. INTRODUCTION

A hospital readmission is an episode when a patient who had been discharged from a hospital is admitted again within a specified time interval. The Centers for Medicare and Medicaid Services (CMS) administers oversight of the Medicare Program and the federal portion of the Medicaid Program. The CMS also ensures that the beneficiaries of the program are aware of the services for which they are eligible and that those services are accessible, are of high quality and develops health and safety standards for providers of health care services authorized by Medicare and Medicaid legislation. The standard benchmark used by the CMS is the 30-day readmission rate.

Readmission rates have increasingly been used as an outcome measure in health services research and as a quality benchmark for health systems. Hospital readmissions can indicate a breakdown in caregiving, whether in the act of transferring a patient from one care environment to another, or between a facility and home. The cost of hospital readmissions is enormous, estimated to be in the vicinity of $26 billion annually (Wilson, 2019), so it's no wonder Medicare is working to reduce this amount. Study found that many readmissions were easily preventable and the reason for the readmissions included discharge timing, follow-up, home health and skilled services. It also found that in 49% of the readmissions the hospital system had some amount of opportunity to improve the discharge process.

Decreasing hospital readmissions – defined as inpatient stays that occur within 30 days of discharge from an initial inpatient hospitalization – is a high priority for the Centers for Medicare & Medicaid Services (CMS). The 30-day risk standardized unplanned readmission measures include:

- Unplanned readmissions that happen within 30 days of discharge from the index (i.e., initial) admission.
- Patients who are readmitted to the same hospital, or another applicable acute care hospital for any reason.

Understanding the drivers of readmissions disparities can help to improve health outcomes for Medicare beneficiaries, particularly for those who are vulnerable, and in containing readmissions-related costs. This study aims to understand the factors, whether demographic or clinical, that are associated with the possible readmission of a patient within 30 days and build and test different predictive models that help analyze and avoid future readmissions. For our study, we have worked with people suffering from Chronic obstructive pulmonary disease (COPD).

Chronic obstructive pulmonary disease (COPD) is a chronic inflammatory lung disease that causes obstructed airflow from the lungs. Symptoms include breathing difficulty, cough, mucus (sputum) production and wheezing. COPD is a leading cause of hospital admission, the fifth leading cause of death in North America, and is estimated to cost $49 billion annually in North America by 2020. The majority of COPD care costs are attributed to hospitalizations; yet, there is limited data to understand the drivers of high costs among hospitalized patients with COPD. Patients with COPD typically have multiple comorbidities, many that share risk factors. In particular, smoking cigarettes and obesity are well-documented causes of inflammation. A few commonly known comorbidities include Cardiovascular Disease, Diabetes and Metabolic Syndrome, Osteoporosis, Lung Cancer, Depression, Sleep Disorders, Medication Reconciliation, etc.

II. METHODOLOGY

A. Data Collection

The DE-SynPUF was created with the goal of providing a realistic set of claims data in the public domain while providing the very highest degree of protection to the Medicare beneficiaries' protected health information. DE-SynPUF is used to develop and create software and applications that may eventually be applied to actual CMS claims data. The data structure of the Medicare DE-SynPUF is very similar to the CMS Limited Data Sets, but with a smaller number of variables.

The DE-SynPUF contains five types of data – Beneficiary Summary, Inpatient Claims, Outpatient Claims, Carrier

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Claims, and Prescription Drug Events.

Although the DE-SynPUF has very limited inferential research value to draw conclusions about Medicare beneficiaries due to the synthetic processes used to create the file, the Medicare DE-SynPUF does increase access to a realistic Medicare claims data file in a timely and less expensive manner to spur the innovation necessary to achieve the goals of better care for beneficiaries and improve the health of the population.

B. Feature Selection

Before feature selection was carried out a cumulative ratio of the people readmitted to people who were not readmitted was found to be as 25434 to 137. This showed a class difference which later turned out to be problematic while working with our different models that we built. To overcome this problem we resorted to the application of SMOTE analysis which is later explained in detail.

In feature selection the important matrix of features are selected which contribute most to the prediction variable or output. Filter-based feature selection methods use statistical measures to score the correlation or dependence between input variables that can be filtered to choose the most relevant features. All the redundant columns such as ID columns, the disease code (ICD9) columns and date columns (which were used to create length of stay, discharge date) were removed. Then the missing data value was calculated and the columns were given values depending upon the missing data and those with missing data over 95% were removed in this step.

In the next step, Lasso Regression was used to further scale down the feature selection. Lasso Regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and be eliminated from the model. Larger penalties result in coefficient values closer to zero, which is ideal for producing simpler models. On the other hand, L2 regularization doesn't result in elimination of coefficients or sparse models. After applying Lasso Regression we get some of the feature coefficients as True(1) or False(0). True is for the features that Lasso thought are important (non-zero features) while False is for the features whose weights were shrunk to zero. A total of 41 features were selected through Lasso Regression.

#	Features Selected by Lasso
1	SP_ALZHDMTA
2	SP_CHF
3	SP_CHRNKIDN
4	SP_CNCR
5	SP_COPD
6	SP_DEPRESSN
7	SP_DIABETES
8	SP_ISCHMCHT
9	SP_OSTEOPRS
10	SP_RA_OA
11	SP_STRKETIA
12	AGE_18_45
13	AGE_45_65
14	LOS_1
15	OtherUpperRespiratoryInfections_Comorbidity
16	RespiratoryFailureInsufficiencyArrest_Comorbidity
17	CancerOFBronchusLung_Comorbidity
18	OtherUpperRespiratoryDisease_Comorbidity
19	OtherLowerRespiratoryDisease_Comorbidity
20	Asthama_Comorbidity
21	AcuteBronchitis_Comorbidity
22	Bronchiectasis_Comorbidity
23	Bronchitis_not_specified_as_acute_or_chronic_Comorbidity
24	DiabetesMellitusWithComplication_Comorbidity
25	EssentialHypertension_Comorbidity
26	HypertensionwithComplications_Comorbidity
27	Osteoporosis_Comorbidity
28	CardiacArrestAndVentricular_Comorbidity
29	MentalHealthRelatedDisorders_Comorbidity
30	SubstanceRelatedDisorders_Comorbidity
31	CardiacDysrhythmias_Comorbidity
32	CPT_Spirometry
33	CPT_Theophylline
34	CPT_Chest_CTScan
35	CPT_Arterial_blood_gas_test
36	BENE_SEX_IDENT_CD_F
37	BENE_SEX_IDENT_CD_M
38	BENE_RACE_CD_Black
39	BENE_RACE_CD_Hispanic
40	BENE_ESRD_IND_1
41	BENE_ESRD_IND_0

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

C. Models Applied

The input for the models are the features selected in Lasso Regression and using that we have predicted the output that is the patient is readmitted or not. We applied 3 models that are XGBoost, Gradient Boosting, and Logistic Regression using SMOTE. We used XGBoost because our dataset was very large, imbalanced and the data had both numerical and categorical features. Also the number of features which were to be used were very less compared to the training dataset. Gradient boosting was used so that we could predict the categorical data. The reason for using Logistic regression is to get categorical results i.e readmitted or not. As the dataset was imbalanced we decided to use SMOTE to increase the number of cases in a balanced way.

XGBoost model is an ensemble learning algorithm based on the gradient-boosted tree algorithm. XGBoost model processes sparse data via a sparsity-aware learning algorithm and weights quantile sketch to approximate tree learning.

Logistic Regression, a machine learning algorithm which is used for classification problems, is a predictive analysis algorithm based on the concept of probability. The hypothesis of logistic regression tends to limit the cost function between 0 and 1.The ratio of readmitted people to people who were not readmitted was observed to be very high. This resulted in a severe class imbalance. The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important. Thus, we decided to use the same models but with SMOTE analysis. Synthetic Minority Oversampling Technique, or SMOTE for short, oversamples the examples in the minority class. This is achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This balances out the class distribution but does not provide any additional information to the model. The minority class for our research is beneficiaries who are not readmitted.

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.

The recall score is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives.

Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

F1 Score is the weighted average of Precision and Recall. This score takes both false positives and false negatives into account.

ROC - AUC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

III. RESULTS AND CONCLUSION

A basic cohort was created which shows us the population distribution ratio of COPD patients to the total number of patients. The list of selected comorbidities for COPD patients, observed in the last 3 months from the index date were as follows.

Characteristics (total number of patients = 25571)	**Number of COPD patients**	**Percentage of total number of patients**
Age		
18-45	945	0.03
45-65	3348	0.13
>65	21278	83.21
		0.00
Sex		0.00
Male	10894	42.60
Female	14677	57.40
		0.00
Race		0.00
White	24163	83.93
Black	2717	10.63
Hispanic	840	3.28
Others	551	2.15
Comorbidities:		0.00
Other Upper Respiratory Infections	158	0.62
Respiratory Failure Insufficiency arrest	142	0.56
Cancer of Bronchus Lung	384	1.50
Other Upper Respiratory Disease	279	1.09
Other Lower Respiratory Disease	1813	7.09
Asthma	414	1.62
Acute Bronchitis	184	0.72
Respiratory Disease Syndrome	0	0.00
Cancer Other	7	0.03

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

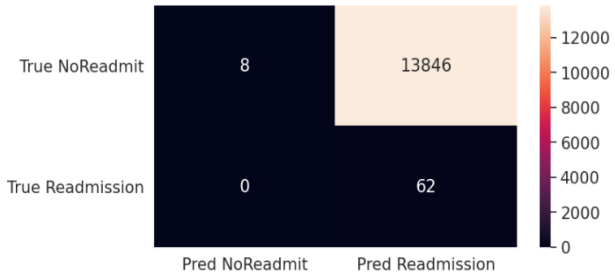
respiratory and Intrathoracic		
Diabetes mellitus without complication	2814	11.00
Essential Hypertension	4344	16.99
Hypertension with complication	509	1.99
Osteoporosis	450	1.76
Cardiac arrest and Ventricular	33	0.13
Mental Health related disorders	235	0.92
Substance related disorders	477	1.87
Cardiac Dysrhythmias	2267	8.87
		0.00
Procedures		0.00
Spirometry test	176	0.69

After analyzing the above cohort we observed that 4 out of 5 people above the age of 60 were suffering from COPD and the majority of this population included white males. Out of all the comorbidities, we also found that 17 percent of the people suffering from hypertension were also suffering from COPD. After applying the models on the features selected the following data and accuracies were acquired for all the different models and we compared the efficiency of the models without SMOTE and with SMOTE analysis.

Weighted Logistic Regression:

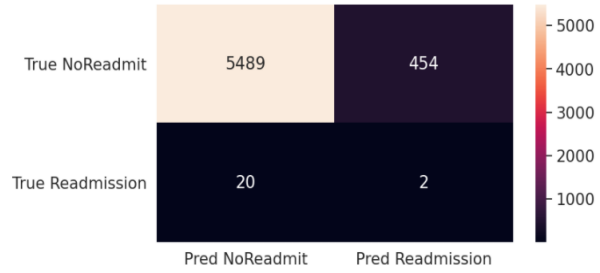
Train set:

Confusion Matrix:



Test set:

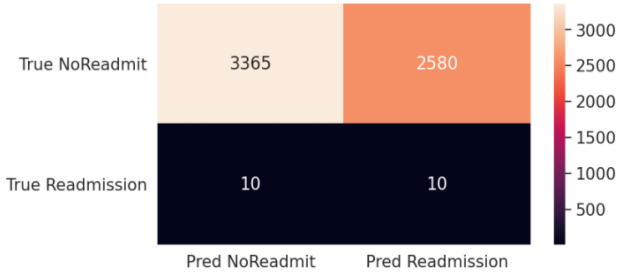
Confusion Matrix:



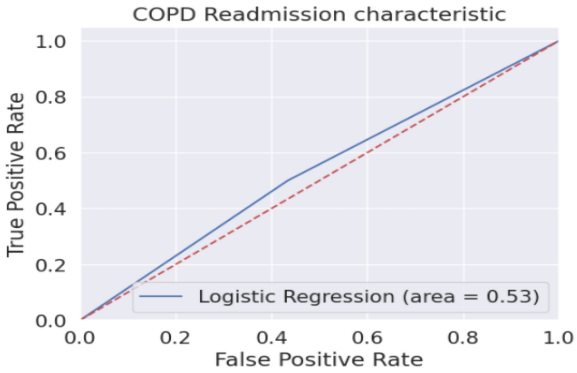
XG-Boost:

Train set:

Confusion Matrix:

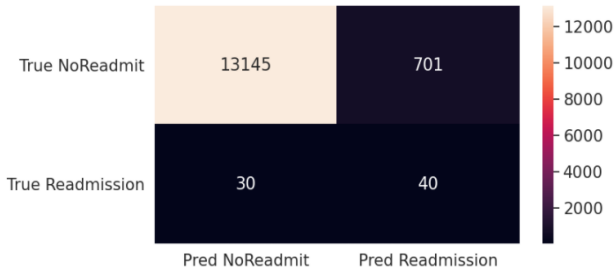


ROC graph:

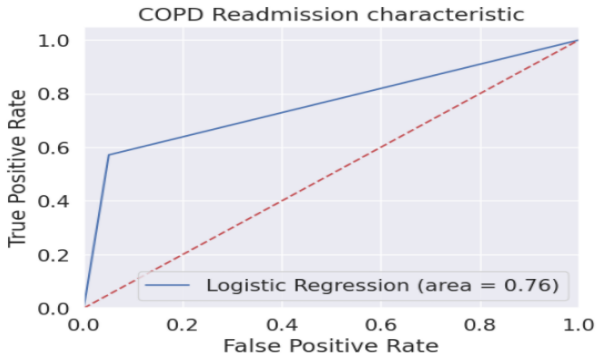


Test Set:

Confusion Matrix:



ROC Graph:

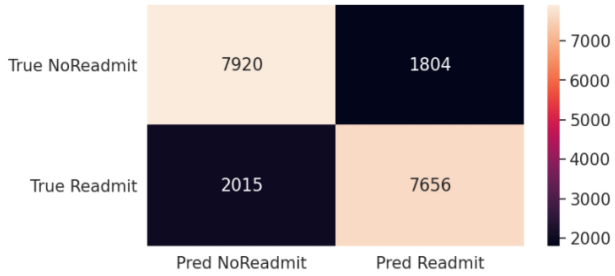


International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

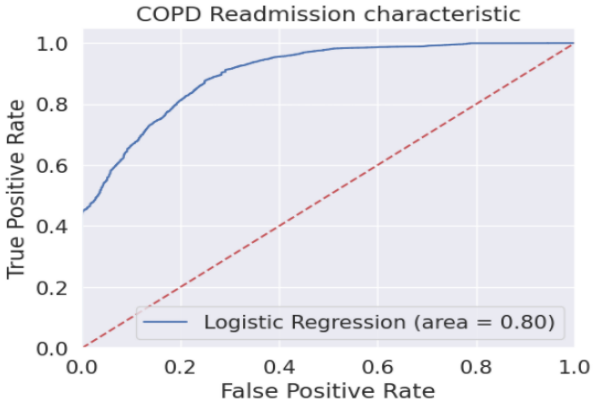
Logistic Regression with SMOTE:

Train Set:

Confusion Matrix:



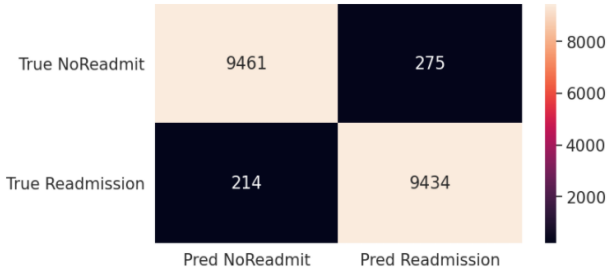
ROC Graph:



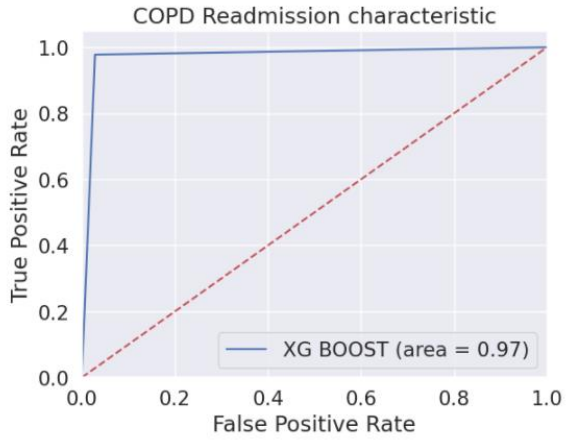
XG-Boost with SMOTE:

Train Set:

Confusion Matrix:

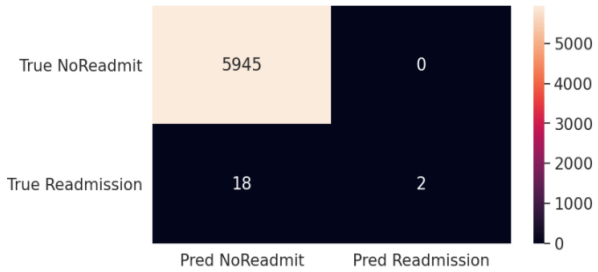


ROC Graph:

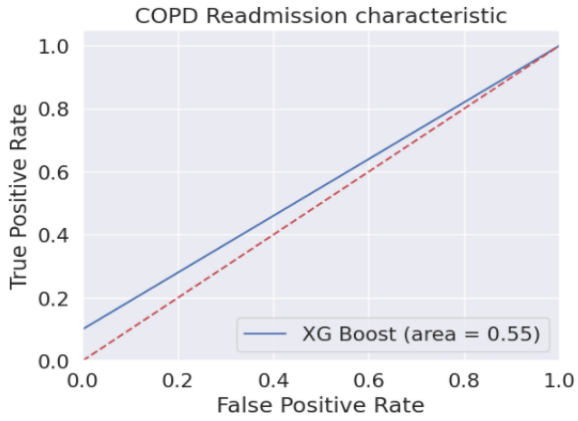


Test Set:

Confusion Matrix:



ROC Graph:



	Train				
Model Name	**Precision**	**Recall**	**Accuracy**	**F1**	**ROC_AUC _Score**
Weighted Logistic Regression	0.004	0.98	0.085	0.009	0.53
XG Boost	0.12	0.53	0.98	0.19	0.76
Gradient Boosting	0.001	0.001	0.99	0	0.5
	Test				
Weighted Logistic Regression	0.005	0.96	0.13	0.01	0.54
XG Boost	0.01	0.46	0.65	0.01	0.53
Gradient Boosting	0.001	0.001	0.99	0	0.5

This study evaluates the different models built to predict readmission occurrence in the case of COPD patients. The dataset worked with i.e DE-SynPUF 2008-2010 after some analysis observed it to be highly imbalanced. We compared three machine learning models; Weighted Logistic regression, XG-Boost, and Gradient Boosting. These Models gave poor accuracy and precision scores when made to work with imbalanced data as shown in the figure above. This imbalance

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

was eliminated by oversampling the minority class with the help of SMOTE analysis. Poor sampling practices can still lead to false conclusions about the quality of a model.

Taking this into consideration, we applied SMOTE to the training tests created for each model and then tested the models on the validation sets created previously. This gave us the following results.

	Train				
Model Name (SMOTE)	**Precision**	**Recall**	**Accuracy**	**F1**	**ROC_AUC _Score**
XG Boost	0.97	0.97	0.97	0.97	0.76
Gradient Boosting	0.001	0.001	0.99	0	0.5
Logistic Regression with SMOTE	0.84	0.92	0.87	0.87	0.87
	Test				
XG Boost	1	0.1	0.99	0.181	0.59
Gradient Boosting	0	0	0.96	0	0.55
Logistic Regression with SMOTE	0.87	0.84	0.92	0.87	0.87

In the case of an imbalanced dataset, the accuracy scores generated can be ignored. Thus the more results that we must focus on metrics like precision, recall, and F1-score. From the above, it can be seen on the actual imbalanced dataset, all 3 models were not able to generalize well on the minority class compared to the majority class. As a result, most of the negative class samples were correctly classified. Due to this, there were fewer false positives compared to more false negatives. Comparing the results with SMOTE and without SMOTE, the models perform as expected with oversampled train data. Amongst the test data results, Logistic regression was observed to have better overall precision, recall and f1 score.

For this project we have used synthetic data for our model testing. Synthetic data is the kind of data that is used in the medical and healthcare sector for which real data is not available. It is also used in the finance sector to test for new fraudulent cases which are examined using synthetic data. It is difficult to create high quality synthetic data if the model is complex. It's important that the synthetic data created is close to real world data, because if it isn't nearly identical then it can compromise the decision making quality of the model hence causing faults in the accuracy. Also, SMOTE overgeneralizes the minority class area concerning the majority class area. In a highly skewed class distribution dataset such as the DE-SynPUF dataset, due to the scanty distribution of the minority class concerning the majority class, there is a greater chance of class mixture. Also, there is no control over the number of synthetic samples generated. Our study thus verifies the importance of handling imbalanced datasets in the predictive modeling process and also discusses the significance of better patient data.

IV. ACKNOWLEDGMENT

Presentation inspiration and motivation have always played a key role in the success of any venture.

It gives us immense pleasure to present this section as a tribute to those who always stood by us as a strong and acted as torchbearers for us.

We would like to express our deep gratitude to Professor Dilip Motwani, our research supervisor, for his patient guidance, enthusiastic encouragement and useful critiques of this report.

We also wish to extend my thanks to Prof. Dilip Motwani and other colleagues for attending our project meetings and for their insightful comments and constructive suggestions to improve the quality of this project work.

V. REFERENCES

[1] Mehdi J, Aleksandr N, Evrett W, Sylvia S, Eric L. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. PLoS ONE. 2017;12(7):e0181173.
[2] Bernard F, Jayasree B. The rate and cost of hospital readmissions for preventable conditions. Med Care Res Rev. 2004;61(2):225240.
[3] Sunil K, Frank L. Deficits in communication and information transfer between hospital based and primary care physicians. JAMA. 2007;297(8):831841.
[4] Patricia H, Yves E, Isaline P. Validation of the potentially avoidable hospital readmission rate as a routine indicator of the quality of hospital care. Med Care. 2006;44(11):972981.
[5] Mark M. Statement of executive director of the Medicare Payment Advisory Commission, before the Subcommittee on Health, Committee on Energy and Commerce. US House of Representatives. April 18, 2007.
[6] Tina S, Matthew MC, Marcelo CP, Tamara K. Understanding why patients with COPD get readmitted: a large national study to delineate the Medicare population for the readmissions penalty expansion. Chest. 2015;147:12191226.
[7] Christopher JLM, Alan DL. Alternative projections of mortality and disability by cause 1990–2020: global burden of disease study. Lancet. 1997;349:1498–504.
[8] Devan K, Honora E, Amanda S, David K, Cecelia T, Michele F, Sunil K. Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306(15):168898.
[9] David WB, Suchi S, Lucila O, Anand S, Gabriel E. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. 2014;33(7):1123.
[10] Carolyn MC. Commentary: reducing hospital readmissions: aligning financial and quality incentives. Am J Med Qual. 2012;27(5):441.
[11] Robert PK, Eli YA. Hospital readmissions and the Affordable Care Act: paying for coordinated quality care. JAMA. 2011;306(16):1794.
[12] Tobias F, Cornelia UK, Dominik O, Frank P. Patterns of Multimorbidity in primary care patients at high risk of future hospitalization. Popul Health Manag. 2012;15(2):119.
[13] Christopher B, Ankur A. A framework for the estimation and reduction of hospital readmission penalties using predictive analytics. J Big Data. 2017;4:37.