Developing a Model to Predict Hospital Encounters for Asthma in Asthmatic Patients: Secondary Analysis

Gang Luo[1], PhD; **Shan He**[2], PhD; **Bryan L Stone**[3], MD, MS; **Flory L Nkoy**[3], MD, MS, MPH; **Michael D Johnson**[3], MD
[1]Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98195, USA
[2]Care Transformation, Intermountain Healthcare, World Trade Center, 16th floor, 60 East South Temple Street, Salt Lake City, UT 84111, USA
[3]Department of Pediatrics, University of Utah, 100 N Mario Capecchi Drive, Salt Lake City, UT 84113, USA
luogang@uw.edu, shan.he@imail.org, bryan.stone@hsc.utah.edu, flory.nkoy@hsc.utah.edu, mike.johnson@hsc.utah.edu

Corresponding author:
Gang Luo, PhD
Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98195, USA
Phone: 1-206-221-4596
Fax: 1-206-221-2671
Email: luogang@uw.edu

Abstract

Background: As a major chronic disease, asthma causes many emergency department visits and hospitalizations each year. Predictive modeling is a key technology to prospectively identify high-risk asthmatic patients and enroll them in care management for preventive care to reduce future hospital encounters including inpatient stays and emergency department visits. Yet, existing models for predicting hospital encounters in asthmatic patients are inaccurate. Usually, they miss over half of the patients who will incur future hospital encounters and incorrectly classify many others who will not. This makes it difficult to match the limited resources of care management to the patients who will incur future hospital encounters, increasing healthcare costs and degrading patient outcomes.

Objective: The goal of this study is to develop a more accurate model for predicting hospital encounters in asthmatic patients.

Methods: Via secondary analysis of 334,564 data instances, Intermountain Healthcare data from 2005 to 2018 were used to build a machine learning classification model to predict hospital encounters for asthma in the following year in asthmatic patients. The patient cohort included all asthmatic patients who resided in Utah or Idaho and visited Intermountain Healthcare facilities during 2005 to 2018. 235 candidate features were considered for model building.

Results: The model achieved an area under the receiver operating characteristic curve of 0.859 (95% CI: 0.846-0.871). When the cutoff threshold for conducting binary classification was set at the top 10% of asthmatic patients with the highest predicted risk, the model reached an accuracy of 90.31% (17391/19256; 95% CI: 89.86-90.70), a sensitivity of 53.69% (436/812; 95% CI: 50.12-57.18), and a specificity of 91.93% (16955/18444; 95% CI: 91.54-92.31). To steer future research on this topic, we pinpointed several potential improvements to our model.

Conclusions: Our model improves the state-of-the-art for predicting hospital encounters for asthma in asthmatic patients. After further refinement, the model could be integrated into a decision support tool to guide asthma care management allocation.

International Registered Report Identifier (IRRID): PRR2-10.2196/5039

Keywords: Asthma; patient care management; forecasting; machine learning

Introduction

Background

In the United States, asthma affects 8.4% of the population and leads to 2.1 million emergency department (ED) visits, 479,300 hospitalizations, 3,388 deaths, and US $50.3 billion in cost annually [1, 2]. Reducing hospital encounters including inpatient stays and ED visits is highly desired for asthmatic patients. For this purpose, using prognostic predictive models to prospectively identify high-risk asthmatic patients and enroll them in care management for tailored preventive care is deemed state of the art and has been adopted by health plans in 9 of 12 regions [3]. Once enrolled, care managers make regular phone calls to help patients book appointments and schedule health and related services. If done properly, this can cut the patients' future hospital encounters by up to 40% [4-7].

Unfortunately, the current high-risk patient identification methods have major gaps, leading to suboptimal outcomes. Care management typically enrolls only 1% to 3% of patients due to capacity constraints [8]. Existing models for predicting hospital encounters in asthmatic patients are inaccurate, reflected by the area under the receiver operating characteristic curve (AUC) ≤0.81 [9-22]. When used for care management, these models miss over half of the patients who will incur future hospital encounters and incorrectly classify many other patients as patients who will incur future hospital encounters. This makes it difficult to align care management enrollment with the patients who will actually incur future hospital encounters, increasing healthcare costs and impairing patient outcomes. If we could find 5% more of the asthmatic patients who would incur future hospital encounters and enroll them in care management, we could improve outcomes and avoid up to 9,850 inpatient stays and 36,000 ED visits each year [1, 4-7].

Objective

The goal of this study is to develop a more accurate model for predicting hospital encounters for asthma in asthmatic patients. The dependent variable is categorical with two possible values: whether future hospital encounter for asthma will occur or not. Accordingly, our model employs clinical and administrative data to perform binary classification, with the intention to better guide care management allocation and improve outcomes for asthmatic patients. A description of the development and evaluation of our model follows. A list of abbreviations used is provided at the end of the paper.

Methods

Study design and ethics approval

In this study, we conducted secondary analysis of retrospective data. The study was reviewed and approved by the institutional review boards of Intermountain Healthcare, University of Utah, and University of Washington Medicine.

Patient population

Our patient cohort was based on the patients who visited Intermountain Healthcare facilities during 2005 to 2018. Intermountain Healthcare is the largest healthcare system in the Intermountain region (Utah and southeastern Idaho), with 185 clinics and 22 hospitals providing care for ~60% of the residents in that region. The patient cohort included asthmatic patients identified as residents of Utah or Idaho, with or without a specific home address. A patient was defined as having asthma in a given year if the patient had at least one diagnosis code of asthma (International Classification of Diseases, Ninth Revision [ICD-9]: 493.0x, 493.1x, 493.8x, 493.9x; International Classification of Diseases, Tenth Revision [ICD-10]: J45.x) in that year in the encounter billing database [11, 23, 24]. Patients who died during that year were excluded. There were no other exclusions.

Prediction target (a.k.a. the dependent variable)

In the rest of this paper, we use hospital encounter for asthma to refer to inpatient stay or ED visit at Intermountain Healthcare with a principal diagnosis of asthma (ICD-9: 493.0x, 493.1x, 493.8x, 493.9x; ICD-10: J45.x). For each patient meeting criteria for asthma in a given year, we looked at any hospital encounter for asthma in the following year as outcome. In our modeling, we used each asthmatic patient's data by the end of each year to predict the patient's outcome in the following year.

Data set

The Intermountain Healthcare enterprise data warehouse provided a structured, clinical and administrative data set including all visits of the patient cohort at Intermountain Healthcare facilities during 2005-2018.

Features (a.k.a. independent variables)

Following the approach outlined in our study design papers [25, 26], we considered 235 candidate features derived from the structured attributes in our data set. These features came from four sources: the >100 potential risk factors for asthma exacerbations reported in the literature [9, 22, 27-33], features used in the existing models for predicting asthma exacerbations [9-22], factors impacting patients' general health status mentioned in the literature [34-36], and features suggested by the clinical experts in our team: MDJ, BLS, and FLN. Since the characteristics of the patient, the care provider, and the treating facility all impact the patient's outcome, we used patient features as well as provider and facility features [25, 26].

The 235 candidate features are listed in Table 1 in the appendix. There, each reference to the number of a specific type of items like medications counts multiplicity, unless the word "distinct" appears. A major visit for asthma is defined as an outpatient visit with a primary diagnosis of asthma, an ED visit with an asthma diagnosis code, or an inpatient stay with an asthma diagnosis code. An outpatient visit with asthma as a secondary diagnosis is defined as a minor visit for asthma. Intuitively, all else being equal and compared with a patient with only minor visits for asthma, a patient with one or more major visits for asthma is more likely to incur future hospital encounters for asthma.

Each input data instance for the predictive model includes the 235 candidate features, targets the unique combination of an asthmatic patient and a year (index year), and is used to predict the patient's outcome in the following year. For that patient and year combination, the patient's age, current primary care provider (PCP), and home address were determined based on the last day of the index year. The features of premature birth, bronchiolitis, the duration of asthma, the duration of chronic obstructive pulmonary disease, whether the patient had any drug or material allergy, whether the patient had any environmental allergy, whether the patient had any food allergy, and the number of allergies of the patient were derived from the historical data from 2005 to the index year. One feature was derived from the historical data in both the index year and the year before. This feature is: the proportion who incurred hospital encounters for asthma in the index year out of all asthmatic patients of the patient's current PCP in the year before. The remaining 226 features were derived from the historical data in the index year.

Data analysis

Data preparation

For every numerical feature, we checked the data distribution, adopted the following lower and upper bounds to spot invalid values, and replaced them with null values. Using the lower and upper bounds from the Guinness World Records [37], all body mass indexes (BMIs) <7.5 or >204, all weights <0.26 kilogram or >635 kilograms, and all heights <0.24 meter or >2.72 meters were deemed physiologically impossible and invalid. Using the lower and upper bounds provided by our team's clinical expert MDJ, all peripheral capillary oxygen saturation (SpO_2) values >100%, all temperatures <80 Fahrenheit or >110 Fahrenheit, all systolic blood pressure values ≤0 mm Hg or >300 mm Hg, all diastolic blood pressure values ≤0 mm Hg or >300 mm Hg, all heart rates <30 beats per minute or >300 beats per minute, and all respiratory rates >120 breaths per minute were deemed physiologically impossible and invalid.

To put all of the numerical features on the same scale, we standardized every numerical feature by first subtracting its mean and then dividing by its standard deviation. Since outcomes were from the following year, our data set provided 13 years of effective data (2005-2017) over a total of 14 years (2005-2018). To reflect model use in practice, the 2005-2016 data were used to train predictive models. The 2017 data were used to assess model performance.

Performance metrics

As shown in the following formulas and Table 1, we applied six standard metrics to gauge model performance: AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

$accuracy = (TP + TN)/(TP + TN + FP + FN)$,

$sensitivity = TP/(TP + FN)$,

$specificity = TN/(TN + FP)$,

$positive\ predictive\ value = TP/(TP + FP)$,

$negative\ predictive\ value = TN/(TN + FN)$.

Here, TP is true positive, TN is true negative, FP is false positive, and FN is false negative. For example, FN is the number of patients who will incur future hospital encounters for asthma and whom the model incorrectly projects to incur no future hospital encounter for asthma. Sensitivity shows the proportion of patients who will incur future hospital encounters for asthma found by the model. Specificity shows the proportion of patients who will incur no future hospital encounter for asthma found by the model.

Table 1. The confusion matrix.

Class	Future hospital encounters for asthma	No future hospital encounter for asthma
Predicted future hospital encounters for asthma	True positive	False positive
Predicted no future hospital encounter for asthma	False negative	True negative

For the six performance metrics, we obtained their 95% confidence intervals via 1,000-fold bootstrap analysis [38]. We calculated our final model's performance metrics on every bootstrap sample of the 2017 data. For each performance metric, we got 1,000 values, the 2.5th and 97.5th percentiles of which gave its 95% confidence interval. We drew the receiver operating characteristic curve to exhibit the sensitivity-specificity tradeoff.

Classification algorithms

We used Weka Version 3.9 [39] to construct machine learning classification models. Weka is a widely used, open-source machine learning and data mining package. It incorporates many standard machine learning algorithms and feature selection techniques. We considered the 39 native machine learning classification algorithms in Weka listed in the appendix, as well as the extreme gradient boosting (XGBoost) classification algorithm [40] implemented in the XGBoost4J package [41]. An XGBoost model is an ensemble of decision trees formed in a stage-wise way. As a scalable and efficient implementation of gradient boosting, XGBoost adopts a more regularized model formulation to help avoid over-fitting and improve classification accuracy. We used our previously developed automatic model selection method [42] and the 2005-2016 training data to automate the selection of the machine learning classification algorithm, feature selection technique, data balancing method for handling imbalanced data, and hyper-parameter values among all of the suitable ones. Our automatic model selection method [42] adopts the response surface methodology to automatically check many combinations of classification algorithm, feature selection technique, data balancing method, and hyper-parameter values, and conducts cross validation to choose the final combination to maximize the AUC. AUC has no reliance on the cutoff threshold used for deciding between projected future hospital encounters for asthma and projected no future hospital encounter for asthma. This gives AUC an advantage over the other five performance metrics accuracy, sensitivity, specificity, PPV, and NPV, whose values depend on the cutoff threshold used. For each classification algorithm, our automatic model selection method attempts to adjust all of the related hyper-parameters by testing many hyper-parameter value combinations. To expedite the search, our method does progressive sampling on the training set and uses test results on its subsets to quickly remove unpromising algorithms and hyper-parameter value combinations. As a result, with no need to find close-to-optimal hyper-parameter value combinations for almost all of the algorithms, our method can return a good combination of the algorithm, feature selection technique, data balancing method, and hyper-parameter values for building the final classification model. Compared with the Auto-WEKA automatic model selection method [43], our method can cut search time by 28 fold and model error rate by 11% simultaneously [42].

Results

Demographic characteristics of our patient cohort

Recall that each data instance targets a unique combination of an asthmatic patient and a year. Tables 2 and 3 exhibit the demographic characteristics of our patient cohort during 2005-2016 and 2017, respectively. The characteristics are relatively similar between the two time periods. During 2005-2016 and 2017, about 3.59% (11332/315308) and 4.22% (812/19256) of data instances linked to hospital encounters for asthma in the following year, respectively.

Table 2. Demographic characteristics of the asthmatic patients at Intermountain Healthcare during 2005-2016.

Characteristic	Data instances (*N*=315308), *n* (%)	Data instances linked to hospital encounters for asthma in the following year (*N*=11332), *n* (%)	Data instances linked to no hospital encounter for asthma in the following year (*N*=303976), *n* (%)
Age			
<6	37826 (12.00)	3118 (27.52)	34708 (11.42)
6 to <18	53162 (16.86)	2590 (22.86)	50572 (16.64)
18 to 65	177439 (56.27)	5003 (44.15)	172436 (56.73)
65+	46881 (14.87)	621 (5.48)	46260 (15.22)
Gender			
Male	127217 (40.35)	5169 (45.61)	122048 (40.15)
Female	188091 (59.65)	6163 (54.39)	181928 (59.85)
Race			
American Indian or Alaska native	2509 (0.80)	214 (1.89)	2295 (0.76)
Asian	2197 (0.70)	77 (0.68)	2120 (0.70)
Black or African American	5751 (1.82)	460 (4.06)	5291 (1.74)
Native Hawaiian or other Pacific islander	4288 (1.36)	411 (3.63)	3877 (1.28)
White	282626 (89.63)	9420 (83.13)	273206 (89.88)
Unknown or not reported	17937 (5.69)	750 (6.62)	17187 (5.65)
Ethnicity			
Hispanic	29293 (9.29)	2279 (20.11)	27014 (8.89)
Non-Hispanic	252599 (80.11)	8157 (71.98)	244442 (80.41)
Unknown or not reported	33416 (10.60)	896 (7.91)	32520 (10.70)
Insurance			
Private	206641 (65.54)	6192 (54.64)	200449 (65.94)
Public	80154 (25.42)	3238 (28.57)	76916 (25.30)
Self-paid or charity	28513 (9.04)	1902 (16.78)	26611 (8.75)
Duration of asthma in years			
≤3	234832 (74.48)	7666 (67.65)	227166 (74.73)
>3	80476 (25.52)	3666 (32.35)	76810 (25.27)
Asthma medication prescription			
Inhaled corticosteroid	78105 (24.77)	4539 (40.05)	73566 (24.20)
Inhaled steroid/rapid-onset long-acting beta2 agonist combination	44992 (14.27)	2196 (19.38)	42796 (14.08)
Leukotriene modifier	35507 (11.26)	2320 (20.47)	33187 (10.92)
Long-acting beta-2 agonist	1813 (0.58)	69 (0.61)	1744 (0.57)
Mast cell stabilizer	121 (0.04)	7 (0.06)	114 (0.04)
Short-acting, inhaled beta-2 agonist	129528 (41.08)	7545 (66.58)	121983 (40.13)
Systemic corticosteroid	136642 (43.34)	7324 (64.63)	129318 (42.54)
Comorbidity			
Allergic rhinitis	4715 (1.50)	181 (1.60)	4534 (1.49)
Anxiety or depression	56961 (18.07)	1716 (15.14)	55245 (18.17)
Bronchopulmonary dysplasia	429 (0.14)	35 (0.31)	394 (0.13)
Chronic obstructive pulmonary disease	12887 (4.09)	391 (3.45)	12496 (4.11)
Cystic fibrosis	458 (0.15)	11 (0.10)	447 (0.15)
Eczema	4927 (1.56)	443 (3.91)	4484 (1.48)
Gastroesophageal reflux	56196 (17.82)	1309 (11.55)	54887 (18.06)
Obesity	36291 (11.51)	1076 (9.50)	35215 (11.58)
Premature birth	5542 (1.76)	440 (3.88)	5102 (1.68)
Sinusitis	14756 (4.68)	592 (5.22)	14164 (4.66)
Sleep apnea	20892 (6.63)	471 (4.16)	20421 (6.72)
Smoking status			
Current smoker	35551 (11.28)	1811 (15.98)	33740 (11.10)
Former smoker	19304 (6.12)	569 (5.02)	18735 (6.16)

	Never smoker or unknown	260453 (82.60)	8952 (79.00)	251501 (82.74)

Table 3. Demographic characteristics of the asthmatic patients at Intermountain Healthcare in 2017.

Characteristic	Data instances (*N*=19256), *n* (%)	Data instances linked to hospital encounters for asthma in the following year (*N*=812), *n* (%)	Data instances linked to no hospital encounter for asthma in the following year (*N*=18444), *n* (%)
Age			
<6	1877 (9.75)	199 (24.51)	1678 (9.10)
6 to <18	3235 (16.80)	181 (22.29)	3054 (16.56)
18 to 65	10265 (53.31)	386 (47.54)	9879 (53.56)
65+	3879 (20.14)	46 (5.67)	3833 (20.78)
Gender			
Male	7816 (40.59)	373 (45.94)	7443 (40.35)
Female	11440 (59.41)	439 (54.06)	11001 (59.65)
Race			
American Indian or Alaska native	159 (0.83)	13 (1.60)	146 (0.79)
Asian	205 (1.06)	10 (1.23)	195 (1.06)
Black or African American	403 (2.09)	42 (5.17)	361 (1.96)
Native Hawaiian or other Pacific islander	346 (1.80)	47 (5.79)	299 (1.62)
White	17706 (91.95)	681 (83.87)	17025 (92.31)
Unknown or not reported	437 (2.27)	19 (2.34)	418 (2.27)
Ethnicity			
Hispanic	2212 (11.49)	192 (23.65)	2020 (10.95)
Non-Hispanic	16860 (87.56)	618 (76.11)	16242 (88.06)
Unknown or not reported	184 (0.96)	2 (0.25)	182 (0.99)
Insurance			
Private	12850 (66.73)	462 (56.90)	12388 (67.17)
Public	5128 (26.63)	208 (25.62)	4920 (26.68)
Self-paid or charity	1278 (6.64)	142 (17.49)	1136 (6.16)
Duration of asthma in years			
≤3	11133 (57.82)	423 (52.09)	10710 (58.07)
>3	8123 (42.18)	389 (47.91)	7734 (41.93)
Asthma medication prescription			
Inhaled corticosteroid	7241 (37.60)	424 (52.22)	6817 (36.96)
Inhaled steroid/rapid-onset long-acting beta2 agonist combination	4400 (22.85)	222 (27.34)	4178 (22.65)
Leukotriene modifier	3573 (18.56)	209 (25.74)	3364 (18.24)
Long-acting beta-2 agonist	52 (0.27)	5 (0.62)	47 (0.25)
Mast cell stabilizer	8 (0.04)	0 (0.00)	8 (0.04)
Short-acting, inhaled beta-2 agonist	13785 (71.59)	739 (91.01)	13046 (70.73)
Systemic corticosteroid	12020 (62.42)	693 (85.34)	11327 (61.41)
Comorbidity			
Allergic rhinitis	392 (2.04)	10 (1.23)	382 (2.07)
Anxiety or depression	3946 (20.49)	131 (16.13)	3815 (20.68)
Bronchopulmonary dysplasia	15 (0.08)	3 (0.37)	12 (0.07)
Chronic obstructive pulmonary disease	1056 (5.48)	23 (2.83)	1033 (5.60)
Cystic fibrosis	95 (0.49)	1 (0.12)	94 (0.51)
Eczema	307 (1.59)	34 (4.19)	273 (1.48)
Gastroesophageal reflux	3548 (18.43)	71 (8.74)	3477 (18.85)
Obesity	3505 (18.20)	116 (14.29)	3389 (18.37)
Premature birth	476 (2.47)	41 (5.05)	435 (2.36)
Sinusitis	780 (4.05)	34 (4.19)	746 (4.04)
Sleep apnea	3003 (15.60)	78 (9.61)	2925 (15.86)

Smoking status				
	Current smoker	2391 (12.42)	146 (17.98)	2245 (12.17)
	Former smoker	2326 (12.08)	83 (10.22)	2243 (12.16)
	Never smoker or unknown	14539 (75.50)	583 (71.80)	13956 (75.67)

Based on the χ^2 two-sample test, for both the 2005-2016 and 2017 data, the data instances linked to future hospital encounters for asthma and those linked to no future hospital encounter for asthma showed the same distribution for long-acting beta-2 agonist prescription (P=.67 for the 2005-2016 data and P=.11 for the 2017 data), mast cell stabilizer prescription (P=.29 for the 2005-2016 data and P=1.00 for the 2017 data), allergic rhinitis occurrence (P=.38 for the 2005-2016 data and P=.13 for the 2017 data), and cystic fibrosis occurrence (P=.21 for the 2005-2016 data and P=.20 for the 2017 data), and different distributions for gender (P<.001 for the 2005-2016 data and P=.002 for the 2017 data), race (P<.001), ethnicity (P<.001), insurance category (P<.001), inhaled corticosteroid prescription (P<.001), inhaled steroid/rapid-onset long-acting beta2 agonist combination prescription (P<.001 for the 2005-2016 data and P=.002 for the 2017 data), leukotriene modifier prescription (P<.001), short-acting, inhaled beta-2 agonist prescription (P<.001), systemic corticosteroid prescription (P<.001), anxiety or depression occurrence (P<.001 for the 2005-2016 data and P=.002 for the 2017 data), bronchopulmonary dysplasia occurrence (P<.001 for the 2005-2016 data and P=.02 for the 2017 data), chronic obstructive pulmonary disease occurrence (P<.001), eczema occurrence (P<.001), gastroesophageal reflux occurrence (P<.001), obesity occurrence (P<.001 for the 2005-2016 data and P=.004 for the 2017 data), premature birth occurrence (P<.001), sleep apnea occurrence (P<.001), and smoking status (P<.001). For the 2005-2016 data, different distributions were shown for sinusitis occurrence (P=.006). For the 2017 data, the same distribution was shown for sinusitis occurrence (P=.91). Based on the Cochran-Armitage trend test [44], for both the 2005-2016 and 2017 data, the data instances linked to future hospital encounters for asthma and those linked to no future hospital encounter for asthma showed different distributions for age (P<.001) and duration of asthma (P<.001).

Features and classification algorithm used

After finishing the search process to maximize the AUC, our automatic model selection method [42] chose the XGBoost classification algorithm [40] and the hyper-parameter values listed in the appendix. XGBoost is based on decision trees and can deal with missing feature values naturally. Since XGBoost only accepts numerical features as its inputs, each categorical feature was first converted into one or more binary features via one-hot encoding before being given to XGBoost. Our final model was constructed using XGBoost and the 142 features listed in descending order of their importance values in Table 2 in the appendix. Due to having no extra predictive power, the other features were automatically removed by XGBoost. As detailed in Section 10.13.1 of Hastie *et al.* [45], XGBoost automatically computed each feature's importance value as the mean of such values across all of the decision trees in the XGBoost model. In each tree, the feature's importance value was computed based on the performance improvement gained by the split at each internal node of the tree using the feature as the splitting variable, weighted by the number of data instances the node is responsible for.

Performance measures achieved

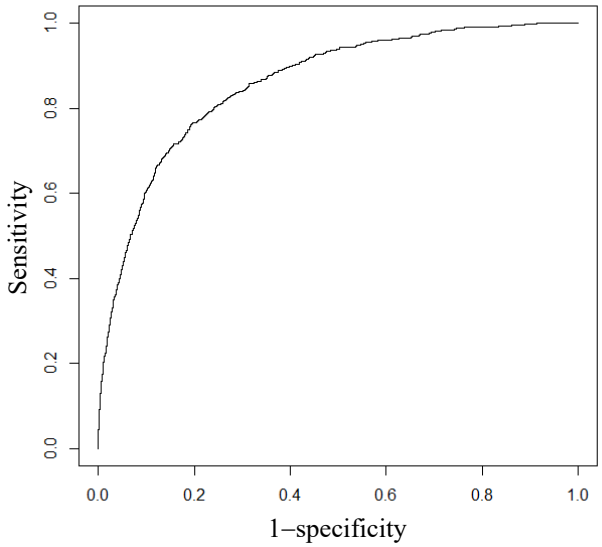

Figure 1. Our final model's receiver operating characteristic curve.

Our final model reached an AUC of 0.859 (95% CI: 0.846-0.871). Figure 1 shows our final model's receiver operating characteristic curve. Table 4 shows our final model's performance metrics when differing top percentages of asthmatic patients with the highest predicted risk were used as the cutoff threshold for conducting binary classifications. When this threshold was at 10%, our final model reached an accuracy of 90.31% (17391/19256; 95% CI: 89.86-90.70), a sensitivity of 53.69% (436/812; 95% CI: 50.12-57.18), a specificity of 91.93% (16955/18444; 95% CI: 91.54-92.31), a PPV of 22.65% (436/1925; 95% CI: 20.74-24.61), and an NPV of 97.83% (16955/17331; 95% CI: 97.60-98.04). Table 5 shows the corresponding confusion matrix of our final model.

Table 4. Our final model's performance metrics when differing top percentages of asthmatic patients with the highest predicted risk were used as the cutoff threshold for conducting binary classification.

Top percentage of asthmatic patients with the highest predicted risk (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
1	95.89	13.05	99.53	55.21	96.30
2	95.54	20.81	98.83	43.90	96.59
3	95.00	26.23	98.03	36.92	96.79
4	94.48	32.02	97.23	33.77	97.01
5	93.84	36.21	96.38	30.56	97.17
6	93.19	40.39	95.52	28.40	97.33
7	92.53	44.33	94.65	26.73	97.48
8	91.85	48.15	93.77	25.39	97.62
9	91.09	51.11	92.85	23.95	97.73
10	90.31	53.69	91.93	22.65	97.83
15	86.44	67.00	87.29	18.84	98.36
20	81.95	73.15	82.34	15.42	98.58
25	77.41	78.57	77.36	13.25	98.80

Table 5. Our final model's confusion matrix when the cutoff threshold for conducting binary classification was set at the top 10% of asthmatic patients with the highest predicted risk.

Class	Future hospital encounters for asthma	No future hospital encounter for asthma
Predicted future hospital encounters for asthma	436	1489
Predicted no future hospital encounter for asthma	376	16955

Recall that several features require more than one year of historical data to compute. If we exclude these features and use only those features computed on one year of historical data, the model's AUC degrades to 0.849.

Without excluding the features that require more than one year of historical data to compute, the model trained on both asthmatic adults' (age ≥ 18) and asthmatic children's (age < 18) data reached an AUC of 0.856 on asthmatic adults and an AUC of 0.830 on asthmatic children. In comparison, the model trained only on asthmatic adults' data reached an AUC of 0.855 on asthmatic adults. The model trained only on asthmatic children's data reached an AUC of 0.821 on asthmatic children.

If we used only the top 21 features listed in Table 2 in the appendix with an importance value ≥0.01 and excluded the other 121 features, the model's AUC degraded from 0.859 to 0.855 (95% CI: 0.842-0.867). When the cutoff threshold for conducting binary classification was set at the top 10% of asthmatic patients with the highest predicted risk, the model's accuracy degraded from 90.31% to 90.14% (17357/19256; 95% CI: 89.74-90.58), sensitivity degraded from 53.69% to 51.60% (419/812; 95% CI: 48.02-55.24), specificity degraded from 91.93% to 91.83% (16938/18444; 95% CI: 91.43-92.24), PPV degraded from 22.65% to 21.77% (419/1925; 95% CI: 20.03-23.68), and NPV degraded from 97.83% to 97.73% (16938/17331; 95% CI: 97.49-97.95).

Discussion

Principal results

We built a more accurate machine learning classification model to predict hospital encounters for asthma in the following year in asthmatic patients. Our final model achieved a higher AUC than what has been reported in the literature for this task [9-22]. After further refinement to improve its accuracy and to automatically explain its prediction results [46, 47], our final model could be integrated into an electronic medical record system to guide care management allocation for asthmatic patients. This could better allocate a scarce and expensive resource and help improve asthma outcomes.

Asthma in adults is different from asthma in children. Our final model reached a higher AUC on asthmatic adults than on asthmatic children. More work is needed to understand the reason for this difference. Also, more work is needed to improve the prediction accuracy on asthmatic children compared with asthmatic adults.

We considered 235 features in total, about 60% of which appeared in our final model. If a feature is unused by our final model, it does not necessarily mean this feature has no predictive power. Rather, it only shows that on our specific data set, this feature offers no extra predictive power beyond what the features used in our final model have. On a larger data set with more asthmatic patients, it is possible some of the excluded features will provide extra predictive power. This is particularly true with features whose non-trivial values occur on only a small portion of asthmatic patients, such as a co-morbidity with a low prevalence rate. When too few data instances take non-trivial values, the features' predictive power may not appear.

In Table 2 of the appendix, the two most important features, as well as several within the top 20, reflect overall instability of the patient's asthma. The instability could derive from physiologic characteristics of the patient's asthma, as reflected by the maximum blood eosinophil count, the maximum percentage of blood eosinophils, and the average respiratory rate. The instability could also result from treatment non-compliance, PCP changes, insurance changes, and socioeconomic issues for which data were unavailable.

Comparison with the prior work

Previously, researchers have developed multiple models to predict inpatient stays and ED visits in asthmatic patients [9-22]. Table 6 compares our final model with these models, which include all of the relevant ones mentioned in Loymans *et al*.'s recent systematic review [9]. None of these models obtained an AUC >0.81, whereas our final model's AUC is 0.859. In other words, compared with our final model, each of these models reached an AUC lower by 0.049 or more. Compared with prior model building, our model building assessed more candidate features with predictive power, adopted a more advanced classification algorithm, and used data from more asthmatic patients. All of these helped boost our final model's accuracy. Our principle of considering extensive candidate features to help enhance model accuracy is general and can be applied to other diseases and outcomes like healthcare cost [48].

Table 6. A comparison of our final model and multiple prior models for predicting inpatient stays and ED visits in asthmatic patients. "-" means the performance measure is not reported in the original paper describing the model.

Model	Prediction target	Classification algorithm	Number of features the model used	Number of data instances	AUC	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Our final model	Hospital encounters for asthma	XGBoost	142	334564	0.859	53.69	91.93	22.65	97.83
Loymans et al. [10]	Asthma exacerbation	Logistic regression	7	611	0.8	-	-	-	-

Schatz et al. [11]	Inpatient stay for asthma in children	Logistic regression	5	4197	0.781	43.9	89.8	5.6	99.1
Schatz et al. [11]	Inpatient stay for asthma in adults	Logistic regression	3	6904	0.712	44.9	87.0	3.9	99.3
Eisner et al. [12]	Inpatient stay for asthma	Logistic regression	1	2858	0.689	-	-	-	-
Eisner et al. [12]	ED visit for asthma	Logistic regression	3	2415	0.751	-	-	-	-
Sato et al. [13]	Severe asthma exacerbation	Classification and regression tree	3	78	0.625	-	-	-	-
Miller et al. [15]	Hospital encounters for asthma	Logistic regression	17	2821	0.81	-	-	-	-
Yurk et al. [17]	Hospital encounters or lost day for asthma	Logistic regression	11	4888	0.78	77	63	82	56
Lieu et al. [18]	Inpatient stay for asthma	Proportional-hazards regression	7	16520	0.79	-	-	-	-
Lieu et al. [18]	ED visit for asthma	Proportional-hazards regression	7	16520	0.69	-	-	-	-
Lieu et al. [19]	Hospital encounters for asthma	Classification and regression tree	4	7141	-	49.0	83.6	18.5	-
Schatz et al. [20]	Hospital encounters for asthma	Logistic regression	4	14893	0.614	25.4	92.0	22.0	93.2
Forno et al. [22]	Severe asthma exacerbation	Scoring	17	615	0.75	-	-	-	-

Except for that in Yurk et al. [17], all of the other prior models had a PPV ≤22% and a sensitivity ≤49%, which are lower than those achieved by our final model. Yurk et al.'s model [17] obtained better sensitivity and PPV primarily because the model used a different prediction target: hospital encounters or ≥1 day lost due to reduced activities or missed work for asthma. This prediction target occurs on over half of asthmatic patients, making it relatively easy to predict. If the prediction target were changed to hospital encounters for asthma, a rarer outcome that is harder to predict, we would expect the sensitivity and PPV reached by Yurk et al.'s model [17] to drop.

Considerations regarding potential clinical use

Despite being more accurate than the prior ones, our final model still reached a relatively low PPV of 22.65%. Yet, this does not prevent our final model from being clinically useful for several reasons.

(1) A PPV of 22.65% is reasonably good for identifying high-risk asthmatic patients as candidates for receiving relatively inexpensive preventive interventions. Four examples of such interventions are teaching the patient how to correctly use an asthma inhaler, teaching the patient how to correctly use a peak flow meter and giving it to the patient to use at home for self-monitoring, training the patient to keep an environmental trigger diary, and arranging for a nurse to make additional follow-up phone calls with the patient.

(2) The PPV depends highly on the outcome's prevalence rate [49]. A relatively rare outcome like future hospital encounters for asthma will occur in only a finite number of patients. Hence, most patients projected to have the outcome will inevitably turn out to not have the outcome, causing even a good predictive model to have a low PPV [49]. For such an outcome, sensitivity is more important than PPV for assessing the model's performance and potential clinical impact. As shown in Table 4, by setting the cutoff threshold for conducting binary classification at the top 10% of patients with the highest predicted risk, our final model has already captured 53.69% of the asthmatic patients who will incur future hospital encounters for asthma. If one is willing to increase the cutoff threshold to the top 25% of patients with the highest predicted risk, our final model would have captured 78.57% of the asthmatic patients who will incur future hospital encounters for asthma, even though the PPV is only 13.25%.

(3) Proprietary models with performance measures similar to those of the previously published models are being used at healthcare systems like Intermountain Healthcare, University of Washington Medicine, and Kaiser Permanente Northern California [18] for allocating preventive interventions. Our final model is an improvement over those models. Table 6 shows that compared with the previously published models, our final model reached a sensitivity higher by 4.69% or more.

If we could use our final model to find 4.69% more of the asthmatic patients who will incur future hospital encounters for asthma and enroll them in care management, we could improve outcomes and avoid up to 9,239 inpatient stays and 33,768 ED visits each year [1, 4-7]. Supporting the importance of relatively small improvements in the model's performance measures, Razavian *et al*. [50] showed that by reaching a gain of 0.05 in AUC (from 0.75 to 0.8) and a PPV of 15%, a large health insurance company like Independence Blue Cross would be willing to deploy a new predictive model to appropriately allocate preventive interventions.

Our final model used 142 features. Reducing features used in the model could ease its clinical deployment. For this, one could use the top few features with the highest importance values (e.g., ≥0.01) and exclude the others, if one is willing to accept a not-too-big degrade of model accuracy. Ideally, one should first assess the features' importance values on a data set from the target healthcare system before deciding which features should be kept for that system. A feature's importance value varies across different healthcare systems. A feature with a low importance value on the Intermountain Healthcare data set might have a decent importance value on a data set from another healthcare system. Like the case with many other complex machine learning models, an XGBoost model using a non-trivial number of features is difficult to interpret globally. As an interesting area for future work, we are in the process of investigating using the automatic explanation approach described in our prior papers [46, 47] to automatically explain our final XGBoost model's prediction results on individual asthmatic patients.

Our final model was built using the XGBoost classification algorithm [40]. For binary classification with two unbalanced classes, XGBoost uses a hyper-parameter scale_pos_weight to control the balance of the weights for the positive and negative classes [51]. One could set scale_pos_weight to the ratio of the number of negative data instances to the number of positive data instances [51], whereas the optimal value of scale_pos_weight often deviates from this value by a degree varying by the specific data set. In our case, to maximize the model's AUC, our automatic model selection method [42] did a search of possible hyper-parameter values and eventually set scale_pos_weight to a non-default value to balance the two classes of future hospital encounters for asthma or not [52]. This has the side effect of making the model's predicted probabilities of incurring future hospital encounters for asthma all very small and unaligned with the actual probabilities [52]. This side effect does not prevent us from selecting the top few percent of asthmatic patients with the highest predicted risk as candidates for receiving care management or other preventive interventions. To avoid this side effect, we could set scale_pos_weight to its default value one without balancing the two classes. But, that would degrade the model's AUC from 0.859 to 0.849 (95% CI: 0.836-0.862).

Limitations

This study has several limitations, all of which provide interesting areas for future work:

(1) We had no access to medication claim data. Consequently, we were unable to use as features the following major risk factors for hospital encounters for asthma in asthmatic patients: medication compliance reflected in refill frequency, the asthma medication ratio [53], the dose of inhaled corticosteroids [32], and the step number of the stepwise approach for managing asthma [32, 54]. We are in the process of obtaining an asthmatic patient data set from Kaiser Permanente Southern California including these attributes [55], so that we can investigate how much gain in prediction accuracy they can bring.

(2) Besides those considered in the study, other features could also help boost model accuracy. Our data set missed some of these features, such as pulmonary function test results. An example of pulmonary function test results is the forced expiratory volume in 1 second / forced vital capacity (FEV1/FVC) ratio, a known risk factor for hospital encounters for asthma in asthmatic patients. It would be interesting to find new predictive features from, but not limited to, the attributes available in our data set.

(3) Our study considered only structured data and non-deep learning machine learning classification algorithms. Adding features extracted from unstructured clinical notes and using deep learning may further improve model accuracy [47, 55].

(4) Our data set included no information on patients' healthcare use at non-Intermountain Healthcare facilities. As a result, we computed features using incomplete clinical and administrative data of the patients [56-59]. Also, instead of taking hospital encounters for asthma anywhere as the prediction target, we had to restrict it to hospital encounters for asthma at Intermountain Healthcare. It would be interesting to investigate how model accuracy would change if more complete clinical and administrative data of the patients are available [60].

(5) Our study used data from one healthcare system and did not assess our results' generalizability. After obtaining the asthmatic patient data set from Kaiser Permanente Southern California, we plan to evaluate our final model's performance on that data set, and explore the process of customizing models to features available in specific data sets as part of the approach to generalization.

Conclusions

Our final model improves the state-of-the-art for predicting hospital encounters for asthma in asthmatic patients. In particular, our final model reached an AUC of 0.859, which is higher than those previously reported in the literature for this task by ≥0.049. After further refinement, our final model could be integrated into an electronic medical record system to guide

allocation of scarce care management resources for asthmatic patients. This could help improve the value equation for asthma care by improving asthma outcomes while also decreasing resource use and cost.

Acknowledgments

We thank Farrant Sakaguchi, Adam B. Wilcox, Zachary C. Liao, Michael Schatz, Robert S. Zeiger, and Jeffrey Povilus for helpful discussions, and Farrant Sakaguchi for helping retrieve the Intermountain Healthcare data set. GL, BLS, FLN, MDJ, and SH were partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R01HL142503. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

GL was mainly responsible for the paper. He conceptualized and designed the study, performed literature review and data analysis, and wrote the paper. BLS, MDJ, and FLN provided feedback on various medical issues, contributed to conceptualizing the presentation, and revised the paper. SH took part in retrieving the Intermountain Healthcare data set and interpreting its detected peculiarities.

Conflicts of interest

None declared.

Abbreviations:

AUC: area under the receiver operating characteristic curve
BMI: body mass index
CPT: Current Procedural Terminology
ED: emergency department
FeNO: fractional exhaled nitrix oxide
FN: false negative
FP: false positive
HCPCS: Healthcare Common Procedure Coding System
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
IgA: immunoglobulin A
IgE: immunoglobulin E
NPV: negative predictive value
NSAID: nonsteroidal anti-inflammatory drug
PCP: primary care provider
PPV: positive predictive value
SpO_2: peripheral capillary oxygen saturation
TN: true negative
TP: true positive
XGBoost: extreme gradient boosting

Appendix

Table 1. The candidate features.

Category	Features
Patient demographics features	Age; gender; race; ethnicity (Hispanic or non-Hispanic); marital status (divorced, married, partnered, separated, single, or widowed); language; and religion.
Features reflecting properties of the area related to the five-digit zip code of the patient's home address	The area's population size; black population percentage; Hispanic population percentage; white population percentage; estimated number of households; average house value; average household income; estimated average number of people per household; average elevation; and the 2003 rural-urban continuum code, which is a number between 1 (most urban) and 9 (most rural) reflecting rurality [61]. Except for the last two, all of these features were derived from 2010 census data.
Features reflecting properties of the census	The block group's number of employed people 16 and older; percentage of employed people 16 and older in a white-collar occupation; percentage of households with >1 person per room;

block group where the patient resides	percentage of households that are owner-occupied; percentage of single-parent households with dependents <18 years old; percentage of occupied housing units without complete plumbing; percentage of households without a phone; percentage of households without a motor vehicle; income disparity measure = log(the number of households with median income < US $15,000 / the number of households with median income > US $75,000); number of civilian labor force 16 and older; median family income; median home value; median monthly mortgage payment; median monthly rent payment; percentage of families below 150% of the federal poverty level; percentage of families below the federal poverty level; percentage of population 25 and older with < 9 years of education; percentage of population 25 and older with a high school diploma or higher education; combined population of each of the census blocks within the block group that qualifies as rural under the 2013 US Census; number of families; number of households; number of occupied households; size of the population 25 and older; population size; number of single-parent households; percentage of the civilian labor force 16 and older that is unemployed; combined population of each of the census blocks within the block group that qualifies as urban under the 2013 US census; national health literacy score developed by the University of North Carolina at Chapel Hill [62]; Singh's area deprivation index measuring the neighborhood's socioeconomic deprivation [63]; indicator of whether the urban population is larger than the rural population; and rural/urban status (rural, urban, or mixed urban/rural). Except for the last four, all of these features were based on the US Census 2013 American Community Survey five-year rolling averages. The last three features were based on 2013 US census data.
Laboratory test-related features	The number of laboratory tests; the number of laboratory tests with abnormal results; the number of days since having the last laboratory test; the maximum blood eosinophil count; the maximum percentage of blood eosinophils; the maximum total serum immunoglobulin E (IgE) level; whether the maximum total serum IgE level is abnormally high; and whether an IgE test was done.
Vital sign features	The maximum diastolic blood pressure; the average diastolic blood pressure; the maximum systolic blood pressure; the average systolic blood pressure; the maximum heart rate; the average heart rate; the maximum respiratory rate; the average respiratory rate; the maximum temperature; the average temperature; the minimum peripheral capillary oxygen saturation (SpO_2); the average SpO_2; the maximum body mass index (BMI); the change of BMI in percentage defined as (the last recorded BMI / the first recorded BMI - 1) × 100%; and the change of weight in percentage defined as (the last recorded weight / the first recorded weight - 1) × 100%.
Diagnosis-related features computed from ICD-9 and ICD-10 diagnosis codes only	The duration of asthma defined as the number of years for which the patient had asthma; the number of ICD-9 and ICD-10 diagnosis codes; chronic obstructive pulmonary disease; the duration of chronic obstructive pulmonary disease defined as the number of years for which the patient had this disease; ischemic heart disease; allergic rhinitis; gastroesophageal reflux; esophagitis; anxiety or depression; eczema; sleep apnea; obesity; gastrostomy tube; upper respiratory tract infection; Alzheimer's or Parkinson's disease; bronchiolitis; bronchopulmonary dysplasia; cystic fibrosis; decreased tone; increased tone; pneumonia; premature birth; vocal cord dysfunction; immunoglobulin A (IgA) deficiency; psoriasis; anaphylaxis; vasculitis; cirrhosis; inflammatory bowel disease; gastrointestinal bleeding; gastrointestinal obstruction; breathing abnormality like dyspnea; substance use; mental disorder; pregnancy; vitamin D deficiency; folate deficiency; myocardial infarction; congestive heart failure; peripheral vascular disease; cerebrovascular disease; dementia; rheumatic disease; peptic ulcer disease; mild liver disease; diabetes without chronic complication; diabetes with chronic complication; hemiplegia or paraplegia; renal disease; malignancy; moderate or severe liver disease; metastatic solid tumor; and acquired immunodeficiency syndrome/human immunodeficiency virus.
Diagnosis-related features computed simultaneously from ICD-9 and ICD-10 diagnosis codes, as well as Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) procedure codes	Cataract; and sinusitis.

Diagnosis-related feature computed simultaneously from ICD-9 and ICD-10 diagnosis codes, as well as ICD-9 and ICD-10 procedure codes	Tracheostomy.
Diagnosis-related feature computed simultaneously from ICD-9 and ICD-10 diagnosis codes, as well as clinical assessment results	The patient's smoking status (current smoker, former smoker, or never smoker or unknown).
Medication-related features	The number of medication orders; the total number of medications in all of the medication orders; the total number of distinct medications in all of the medication orders; the total number of refills allowed in all of the medication orders; the total number of units ordered in all of the medication orders; the number of asthma medication orders; the total number of medications in all of the asthma medication orders; the total number of distinct medications in all of the asthma medication orders; the total number of refills allowed in all of the asthma medication orders; the total number of units of asthma medications ordered; the total number of short-acting beta-2 agonists ordered; the total number of units of short-acting beta-2 agonists ordered; the total number of refills allowed in all of the short-acting beta-2 agonist orders; the total number of systemic corticosteroids ordered; the total number of units of systemic corticosteroids ordered; the total number of refills allowed in all of the systemic corticosteroid orders; the total number of asthma reliever medications ordered that are neither systemic corticosteroids nor short-acting beta-2 agonists; the total number of units of asthma reliever medications ordered that are neither systemic corticosteroids nor short-acting beta-2 agonists; the total number of inhaled corticosteroids ordered; the total number of units of inhaled corticosteroids ordered; the total number of refills allowed in all of the inhaled corticosteroid orders; the total number of mast cell stabilizers ordered; the total number of units of mast cell stabilizers ordered; the total number of refills allowed in all of the mast cell stabilizer orders; the total number of nonsteroidal anti-inflammatory drugs (NSAIDs) ordered; the total number of units of NSAIDs ordered; the total number of refills allowed in all of the NSAID orders; the total number of antihistamines ordered; the total number of units of antihistamines ordered; the total number of refills allowed in all of the antihistamine orders; the total number of allergen immunotherapy medications ordered; the total number of nasal steroid sprays ordered; the total number of units of nasal steroid sprays ordered; the total number of refills allowed in all of the nasal steroid spray orders; the total number of beta blockers ordered; the total number of units of beta blockers ordered; the total number of refills allowed in all of the beta blocker orders; the total number of statins ordered; the total number of units of statins ordered; the total number of refills allowed in all of the statin orders; whether spacer was used; and whether nebulizer was used.
Insurance-related features	The primary payer's insurance category (Medicaid, Medicare, Intermountain Healthcare's own health insurance plan SelectHealth, other private insurance, or self-paid or charity) at the patient's last visit; the number of insurances of the patient at the last visit; and the number of distinct primary payers across all of the patient's visits.
Visit type-related features for the patient	The number of outpatient visits; the number of outpatient visits with a primary diagnosis of asthma; the number of outpatient visits to the patient's PCP; the number of outpatient visits to specialists; the number of outpatient visits to allergists and immunologists; the number of ED visits; the length of stay of the last ED visit; the average length of stay of an ED visit; the number of inpatient stays; the total length of all of the inpatient stays; the average length of an inpatient stay; the number of admissions to intensive care; the length of the last intensive care unit stay; the average length of an intensive care unit stay; the last visit's admission type (emergency, urgent, elective, or trauma); the most emergent one among all of the visits' admission types; and the number of major visits for asthma.
Features related to appointment scheduling and visit status	The number of no shows; the number of cancelled appointments; the number of visits that were referred; the day of the week at the last ED visit's admission time; the admit hour of the last ED visit; the discharge disposition location (home, left against medical advice, or other non-home location) of the last visit; the time between making the request and the actual visit of the last visit

	reflecting the request's urgency; the shortest time between making the request and the actual visit among all of the visits; the number of days since the last inpatient stay; whether the last inpatient stay was through the ED; the number of days since the last outpatient visit; the number of days since the last ED visit; the number of times the patient left against medical advice; and the acuity level (resuscitation, emergent, urgent, semi-urgent, or non-urgent) of the last ED visit.
Features reflecting care continuity of the patient	The number of distinct EDs the patient visited; the number of distinct providers seen in outpatient visits; the number of distinct PCPs of the patient; the number of distinct medication prescribers; the number of distinct asthma medication prescribers; whether the patient was homeless; whether the patient had no phone number; and the number of distinct addresses the patient had, reflecting the number of times the patient moved.
Procedure-related features	The number of ICD-9 and ICD-10 procedure codes; the number of CPT/HCPCS procedure codes; the number of CPT/HCPCS procedure codes for influenza vaccination; the number of HCPCS procedure codes for home oxygen therapy; the number of CPT procedure codes for pulmonary function tests; the number of CPT procedure codes for the fractional exhaled nitric oxide (FeNO) test; and mechanical ventilation shown by ICD-9 and ICD-10 procedure codes.
Radiology-related feature	The number of chest X-ray exams.
Allergy features	Whether the patient had any drug or material allergy; whether the patient had any environmental allergy; whether the patient had any food allergy; and the number of allergies of the patient.
Clinical assessment-related feature	The number of times the patient was assessed to be confused.
Provider features	We considered several features of the patient's current PCP defined as the patient's PCP known at the patient's last clinic visit. These features include: whether the patient and the PCP are of the same gender; the PCP's age; whether the PCP is a preferred provider of Intermountain Healthcare's health insurance plan SelectHealth; the level of affiliation that the PCP has with Intermountain Healthcare (independent practitioner, employed by an Intermountain Healthcare hospital, employed by the Intermountain Medical Group managing Intermountain Healthcare's clinics, or non-credentialed provider); the PCP's primary specialty; the PCP's primary profession type (Doctor of Medicine, Doctor of Osteopathic Medicine, nurse practitioner, family nurse practitioner, advanced practice registered nurse, physician assistant, or other); the number of asthmatic patients of the PCP; and the proportion who incurred hospital encounters for asthma in the index year out of all asthmatic patients of the PCP in the year before.
Facility features	The ellipsoid great circle distance between the patient's home and the closest ED, which was computed based on the longitude and latitude coordinates of the ED location and the five-digit zip code of the patient's home address; and the ellipsoid great circle distance between the patient's home and the patient's current PCP's office, which was computed based on the longitude and latitude coordinates of the PCP's office location and the five-digit zip code of the patient's home address.

Table 2. The features adopted in our final model and their importance values.

Rank	Feature	Importance based on the feature's fractional contribution to the model
1	The number of major visits for asthma	0.1413
2	The total number of units of systemic corticosteroids ordered	0.1241
3	The number of days since the last ED visit	0.0787
4	Age	0.0586
5	The last visit's admission type = elective	0.0515
6	Duration of asthma	0.0482
7	The number of ED visits	0.0363
8	The total number of units of short-acting beta-2 agonists ordered	0.0327
9	The total number of short-acting beta-2 agonists ordered	0.0251
10	The total number of systemic corticosteroids ordered	0.0204
11	The maximum blood eosinophil count	0.0178
12	The maximum percentage of blood eosinophils	0.0177
13	The number of ICD-9 and ICD-10 procedure codes	0.0173

14	The number of distinct asthma medication prescribers	0.0167
15	The average respiratory rate	0.0139
16	The average heart rate	0.0129
17	The total number of units ordered in all of the medication orders	0.0128
18	The proportion who incurred hospital encounters for asthma in the index year out of all asthmatic patients of the PCP in the year before	0.0123
19	Ethnicity	0.0118
20	Whether nebulizer was used	0.0110
21	The time between making the request and the actual visit of the last visit	0.0109
22	The number of asthma medication orders	0.0077
23	The number of ICD-9 and ICD-10 diagnosis codes	0.0076
24	The total number of distinct medications in all of the medication orders	0.0075
25	The total number of units of asthma medications ordered	0.0072
26	The block group's national health literacy score	0.0068
27	The total number of distinct medications in all of the asthma medication orders	0.0061
28	The block group's median family income	0.0059
29	Marital status = married	0.0059
30	The number of outpatient visits	0.0057
31	The percentage of families below 150% of the federal poverty level in the block group	0.0049
32	The total number of medications in all of the medication orders	0.0049
33	The maximum BMI	0.0048
34	The shortest time between making the request and the actual visit among all of the visits	0.0046
35	The number of laboratory tests with abnormal results	0.0042
36	The number of distinct providers seen in outpatient visits	0.0038
37	The average diastolic blood pressure	0.0035
38	The ellipsoid great circle distance between the patient's home and the closest ED	0.0034
39	The change of BMI in percentage	0.0034
40	The total number of units of inhaled corticosteroids ordered	0.0034
41	Singh's area deprivation index of the block group	0.0034
42	The number of insurances of the patient at the last visit	0.0033
43	The area's black population percentage	0.0032
44	Race = white	0.0028
45	The average length of an inpatient stay	0.0028
46	The percentage of population 25 and older with a high school diploma or higher education in the block group	0.0028
47	The block group's income disparity measure	0.0028
48	The area's white population percentage	0.0027
49	The percentage of employed people 16 and older in the block group who are in a white-collar occupation	0.0027
50	The area's average house value	0.0025
51	The number of allergies of the patient	0.0024
52	The number of families in the block group	0.0024
53	The number of distinct medication prescribers	0.0024
54	The change of weight in percentage	0.0022
55	The ellipsoid great circle distance between the patient's home and the patient's current PCP's office	0.0022
56	The average SpO_2	0.0022
57	The area's Hispanic population percentage	0.0022
58	Race = Asian	0.0022
59	The number of days since the last outpatient visit	0.0020
60	The total number of refills allowed in all of the medication orders	0.0020
61	The block group's median monthly rent payment	0.0019
62	The number of laboratory tests	0.0019
63	The admit hour of the last ED visit	0.0019

64	Gender	0.0018
65	The combined population of each of the census blocks within the block group that qualifies as urban under the 2013 US census	0.0018
66	The percentage of single-parent households with dependents <18 years old in the block group	0.0018
67	Bronchiolitis	0.0018
68	The maximum temperature	0.0017
69	The block group's median monthly mortgage payment	0.0017
70	The total length of all of the inpatient stays	0.0017
71	The number of asthmatic patients of the PCP	0.0017
72	Religion = Protestant	0.0016
73	The percentage of the civilian labor force 16 and older in the block group that is unemployed	0.0016
74	The percentage of households in the block group that are owner-occupied	0.0015
75	The number of civilian labor force 16 and older in the block group	0.0015
76	Whether the patient had any food allergy	0.0015
77	The PCP's age	0.0015
78	The minimum SpO_2	0.0015
79	The estimated average number of people per household in the area	0.0014
80	Religion = Catholic	0.0013
81	The number of households in the block group	0.0013
82	The average systolic blood pressure	0.0013
83	The average temperature	0.0012
84	The area's population size	0.0012
85	The total number of units of asthma reliever medications ordered that are neither systemic corticosteroids nor short-acting beta-2 agonists	0.0011
86	The number of chest X-ray exams	0.0011
87	The percentage of households in the block group without a motor vehicle	0.0011
88	The number of medication orders	0.0011
89	The area's average household income	0.0011
90	The size of the population 25 and older in the block group	0.0011
91	The area's average elevation	0.0011
92	The percentage of households in the block group with >1 person per room	0.0011
93	The primary payer's insurance category at the patient's last visit = other private insurance	0.0010
94	Smoking status = current smoker	0.0010
95	The number of no shows	0.0010
96	Whether the last inpatient stay was through the ED	0.0009
97	Whether the patient had any drug or material allergy	0.0009
98	The number of CPT procedure codes for pulmonary function tests	0.0009
99	The maximum respiratory rate	0.0009
100	The number of CPT/HCPCS procedure codes	0.0009
101	The acuity level of the last ED visit	0.0008
102	The number of days since having the last laboratory test	0.0008
103	The median home value in the block group	0.0008
104	The number of occupied households in the block group	0.0008
105	Religion = Christian	0.0008
106	The maximum heart rate	0.0008
107	The primary payer's insurance category at the patient's last visit = SelectHealth	0.0007
108	The percentage of population 25 and older in the block group with < 9 years of education	0.0007
109	The PCP's primary specialty = family medicine	0.0007
110	The number of employed people 16 and older in the block group	0.0007
111	The area's 2003 rural-urban continuum code	0.0006
112	The percentage of families in the block group that are below the federal poverty level	0.0006
113	The percentage of households in the block group without a phone	0.0006

114	The length of stay of the last ED visit	0.0006
115	Diabetes without chronic complication	0.0006
116	The number of days since the last inpatient stay	0.0006
117	The day of the week at the last ED visit's admission time	0.0006
118	The maximum diastolic blood pressure	0.0005
119	The total number of refills allowed in all of the short-acting beta-2 agonist orders	0.0005
120	Religion = Baptist	0.0005
121	Smoking status = former smoker	0.0005
122	The number of cancelled appointments	0.0005
123	The estimated number of households in the area	0.0004
124	The PCP's primary profession type = Doctor of Osteopathic Medicine	0.0004
125	Whether the patient had any environmental allergy	0.0004
126	The total number of refills allowed in all of the inhaled corticosteroid orders	0.0004
127	The maximum systolic blood pressure	0.0004
128	The total number of units of NSAIDs ordered	0.0004
129	Among the admission types of all of the visits of the patient, the one with the highest priority = urgent	0.0004
130	Chronic obstructive pulmonary disease	0.0004
131	Language = Spanish	0.0004
132	Obesity	0.0004
133	Marital status = single	0.0004
134	Upper respiratory tract infection	0.0004
135	The number of outpatient visits to the patient's PCP	0.0003
136	The length of the last intensive care unit stay	0.0003
137	The number of visits that were referred	0.0003
138	The total number of refills allowed in all of the nasal steroid spray orders	0.0002
139	Breathing abnormality like dyspnea	0.0002
140	The block group's rural/urban status	0.0002
141	The total number of antihistamines ordered	0.0002
142	The duration of chronic obstructive pulmonary disease	0.0002

The 39 native machine learning classification algorithms in Weka: Bayes net, naive Bayes, naive Bayes multinomial, Gaussian process, linear regression, logistic regression, single-layer perceptron, stochastic gradient descent, support vector machine, simple linear regression, simple logistic regression, voted perceptron, *k*-nearest neighbor, *K*-star, decision table, RIPPER, M5 rules, 1-R, PART, 0-R, decision stump, C4.5 decision tree, logistic model tree, M5 tree, random forest, random tree, REP tree, locally weighted learning, AdaBoost M1, additive regression, attribute selected, bagging, classification via regression, LogitBoost, multiclass classifier, random committee, random subspace, voting, and stacking.

The hyper-parameter values of the XGBoost classification algorithm used in the final predictive model: alpha=1, Booster=gbtree, colsample_bytree=1, eta=0.3, eval_metric=auc, gamma=0, lambda=0, max.depth=4, min_child_weight=5, nrounds=100, objective=binary:logistic, scale_pos_weight=0.02, and subsample=1.

References

1. Moorman JE, Akinbami LJ, Bailey CM, Zahran HS, King ME, Johnson CA, Liu X. National surveillance of asthma: United States, 2001-2010. Vital Health Stat 3 2012;(35):1-58. PMID:24252609
2. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc 2018;15(3):348-56. PMID:29323930
3. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood). 2004;Suppl Web Exclusives:W4-427-36. PMID:15451964
4. Caloyeras JP, Liu H, Exum E, Broderick M, Mattke S. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. Health Aff (Millwood) 2014;33(1):124-31. PMID:24395944
5. Greineder DK, Loane KC, Parks P. A randomized controlled trial of a pediatric asthma outreach program. J Allergy Clin Immunol 1999;103(3 Pt 1):436-40. PMID:10069877

6. Kelly CS, Morrow AL, Shults J, Nakas N, Strope GL, Adelman RD. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in Medicaid. Pediatrics 2000;105(5):1029-35. PMID:10790458
7. Axelrod RC, Zimbro KS, Chetney RR, Sabol J, Ainsworth VJ. A disease management program utilizing life coaches for children with asthma. J Clin Outcomes Manag 2001;8(6):38-42.
8. Axelrod RC, Vogel D. Predictive modeling in health plans. Disease Management & Health Outcomes 2003;11(12):779-87. doi:10.2165/00115677-200311120-00003
9. Loymans RJB, Debray TPA, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Schermer TRJ, Assendelft WJJ, Timp M, Chung KF, Sousa AR, Sont JK, Sterk PJ, Reddel HK, Ter Riet G. Exacerbations in adults with asthma: a systematic review and external validation of prediction models. J Allergy Clin Immunol Pract 2018;6(6):1942-52.e15. PMID:29454163
10. Loymans RJ, Honkoop PJ, Termeer EH, Snoeck-Stroband JB, Assendelft WJ, Schermer TR, Chung KF, Sousa AR, Sterk PJ, Reddel HK, Sont JK, Ter Riet G. Identifying patients at risk for severe exacerbations of asthma: development and external validation of a multivariable prediction model. Thorax 2016;71(9):838-46. PMID:27044486
11. Schatz M, Cook EF, Joshua A, Petitti D. Risk factors for asthma hospitalizations in a managed care organization: development of a clinical prediction rule. Am J Manag Care 2003;9(8):538-47. PMID:12921231
12. Eisner MD, Yegin A, Trzaskoma B. Severity of asthma score predicts clinical outcomes in patients with moderate to severe persistent asthma. Chest 2012;141(1):58-65. PMID:21885725
13. Sato R, Tomita K, Sano H, Ichihashi H, Yamagata S, Sano A, Yamagata T, Miyara T, Iwanaga T, Muraki M, Tohda Y. The strategy for predicting future exacerbation of asthma using a combination of the Asthma Control Test and lung function test. J Asthma 2009;46(7):677-82. PMID:19728204
14. Osborne ML, Pedula KL, O'Hollaren M, Ettinger KM, Stibolt T, Buist AS, Vollmer WM. Assessing future need for acute care in adult asthmatics: the Profile of Asthma Risk Study: a prospective health maintenance organization-based study. Chest 2007;132(4):1151-61. PMID:17573515
15. Miller MK, Lee JH, Blanc PD, Pasta DJ, Gujrathi S, Barron H, Wenzel SE, Weiss ST; TENOR Study Group. TENOR risk score predicts healthcare in adults with severe or difficult-to-treat asthma. Eur Respir J 2006;28(6):1145-55. PMID:16870656
16. Peters D, Chen C, Markson LE, Allen-Ramey FC, Vollmer WM. Using an asthma control questionnaire and administrative data to predict health-care utilization. Chest 2006;129(4):918-24. PMID:16608939
17. Yurk RA, Diette GB, Skinner EA, Dominici F, Clark RD, Steinwachs DM, Wu AW. Predicting patient-reported asthma outcomes for adults in managed care. Am J Manag Care 2004;10(5):321-8. PMID:15152702
18. Lieu TA, Quesenberry CP, Sorel ME, Mendoza GR, Leong AB. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998;157(4 Pt 1):1173-80. PMID:9563736
19. Lieu TA, Capra AM, Quesenberry CP, Mendoza GR, Mazar M. Computer-based models to identify high-risk adults with asthma: is the glass half empty or half full? J Asthma 1999;36(4):359-70. PMID:10386500
20. Schatz M, Nakahiro R, Jones CH, Roth RM, Joshua A, Petitti D. Asthma population management: development and validation of a practical 3-level risk stratification scheme. Am J Manag Care 2004;10(1):25-32. PMID:14738184
21. Grana J, Preston S, McDermott PD, Hanchak NA. The use of administrative data to risk-stratify asthmatic patients. Am J Med Qual 1997;12(2):113-9. PMID:9161058
22. Forno E, Fuhlbrigge A, Soto-Quirós ME, Avila L, Raby BA, Brehm J, Sylvia JM, Weiss ST, Celedón JC. Risk factors and predictive clinical scores for asthma exacerbations in childhood. Chest 2010;138(5):1156-65. PMID:20472862
23. Desai JR, Wu P, Nichols GA, Lieu TA, O'Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care 2012;50 Suppl:S30-5. PMID:22692256
24. Wakefield DB, Cloutier MM. Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma. Pediatr Pulmonol 2006;41(10):962-71. PMID:16871628
25. Luo G, Stone BL, Sakaguchi F, Sheng X, Murtaugh MA. Using computational approaches to improve risk-stratified patient management: rationale and methods. JMIR Res Protoc 2015;4(4):e128. PMID:26503357
26. Luo G, Sward K. A roadmap for optimizing asthma care management via computational approaches. JMIR Med Inform 2017;5(3):e32. PMID:28951380
27. Puranik S, Forno E, Bush A, Celedón JC. Predicting severe asthma exacerbations in children. Am J Respir Crit Care Med 2017;195(7):854-9. PMID:27710010
28. Buelo A, McLean S, Julious S, Flores-Kim J, Bush A, Henderson J, Paton JY, Sheikh A, Shields M, Pinnock H; ARC Group. At-risk children with asthma (ARC): a systematic review. Thorax 2018;73(9):813-24. PMID:29871982
29. Greenberg S. Asthma exacerbations: predisposing factors and prediction rules. Curr Opin Allergy Clin Immunol 2013;13(3):225-36. PMID:23635528
30. Fleming L. Asthma exacerbation prediction: recent insights. Curr Opin Allergy Clin Immunol 2018;18(2):117-23. PMID:29406359

31. Ledford DK, Lockey RF. Asthma and comorbidities. Curr Opin Allergy Clin Immunol 2013;13(1):78-86. PMID:23222157
32. Blakey JD, Price DB, Pizzichini E, Popov TA, Dimitrov BD, Postma DS, Josephs LK, Kaplan A, Papi A, Kerkhof M, Hillyer EV, Chisholm A, Thomas M. Identifying risk of future asthma attacks using UK medical record data: a respiratory effectiveness group initiative. J Allergy Clin Immunol Pract 2017;5(4):1015-24.e8. PMID:28017629
33. Das LT, Abramson EL, Stone AE, Kondrich JE, Kern LM, Grinspan ZM. Predicting frequent emergency department visits among children with asthma using EHR data. Pediatr Pulmonol 2017;52(7):880-90. PMID:28557381
34. Purdey S, Huntley A. Predicting and preventing avoidable hospital admissions: a review. J R Coll Physicians Edinb 2013;43(4):340-4. PMID:24350320
35. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, Saunders LD, Beck CA, Feasby TE, Ghali WA. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43(11):1130-9. PMID:16224307
36. Wallace E, Stuart E, Vaughan N, Bennett K, Fahey T, Smith SM. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Med Care 2014;52(8):751-65. PMID:25023919
37. The Guinness World Records homepage. 2019. https://www.guinnessworldrecords.com.
38. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York, NY: Springer; 2009. ISBN:038777243X
39. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann; 2016. ISBN:0128042915
40. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 p. 785-94. doi:10.1145/2939672.2939785
41. XGBoost JVM package. 2019. https://xgboost.readthedocs.io/en/latest/jvm/index.html.
42. Zeng X, Luo G. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Health Inf Sci Syst 2017;5(1):2. PMID:29038732
43. Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013 p. 847-55. doi:10.1145/2487575.2487629
44. Agresti A. Categorical Data Analysis, 3rd ed. Hoboken, NJ: Wiley; 2012. ISBN:9780470463635
45. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer; 2016. ISBN:0387848576
46. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst 2016;4:2. PMID:26958341
47. Luo G. A roadmap for semi-automatically extracting predictive and clinically meaningful temporal features from medical data for predictive modeling. Glob Transit 2019;1:61-82. PMID:31032483
48. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(1):198-208. PMID:27189013
49. Ranganathan P, Aggarwal R. Common pitfalls in statistical analysis: Understanding the properties of diagnostic tests - Part 1. Perspect Clin Res 2018;9(1):40-43.
50. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 2015;3(4):277-87. PMID:27441408
51. XGBoost parameters. 2019. https://xgboost.readthedocs.io/en/latest/parameter.html.
52. Notes on parameter tuning. 2019. https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html.
53. Andrews AL, Simpson AN, Basco WT Jr, Teufel RJ 2nd. Asthma medication ratio predicts emergency department visits and hospitalizations in children with asthma. Medicare Medicaid Res Rev 2013;3(4):E1-10. PMID:24834366
54. National Asthma Education and Prevention Program. Expert panel report 3: guidelines for the diagnosis and management of asthma. 2007. http://www.nhlbi.nih.gov/files/docs/guidelines/asthgdln.pdf.
55. Luo G, Stone BL, Koebnick C, He S, Au DH, Sheng X, Murtaugh MA, Sward KA, Schatz M, Zeiger RS, Davidson GH, Nkoy FL. Using temporal features to provide data-driven clinical early warnings for chronic obstructive pulmonary disease and asthma care management: protocol for a secondary analysis. JMIR Res Protoc 2019;8(6):e13783. PMID:31199308
56. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. Arch Intern Med 2010;170(22):1989-95. PMID:21149756
57. Finnell JT, Overhage JM, Grannis S. All health care is not local: an evaluation of the distribution of emergency department care delivered in Indiana. AMIA Annu Symp Proc 2011;2011:409-16. PMID:22195094
58. Luo G, Tarczy-Hornoch P, Wilcox AB, Lee ES. Identifying patients who are likely to receive most of their care from a specific health care system: demonstration via secondary analysis. JMIR Med Inform 2018;6(4):e12241. PMID:30401670
59. Kern LM, Grinspan Z, Shapiro JS, Kaushal R. Patients' use of multiple hospitals in a major US city: implications for population management. Popul Health Manag 2017;20(2):99-102. PMID:27268133

60. Samuels-Kalow ME, Faridi MK, Espinola JA, Klig JE, Camargo CA Jr. Comparing statewide and single-center data to predict high-frequency emergency department utilization among patients with asthma exacerbation. Acad Emerg Med 2018;25(6):657-67. PMID:29105238
61. Populations studies center. Data sharing for demographic research knowledge base. 2017. https://dsdr-kb.psc.isr.umich.edu/answer/1102.
62. The Health Literacy Data Map homepage. 2019. http://healthliteracymap.unc.edu.
63. Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. Am J Public Health 2003;93(7):1137-43. PMID:12835199