

ORIGINAL ARTICLE: ASTHMA

Predicting frequent emergency department visits among children with asthma using EHR data

Lala T. Das MS¹  | Erika L. Abramson MD, MS^{1,2,3} | Anne E. Stone MD^{2,3} |
Janienne E. Kondrich MD^{2,3} | Lisa M. Kern MD, MPH^{1,3,4} |
Zachary M. Grinspan MD, MS^{1,2,3}

¹ Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, New York

² Department of Pediatrics, Weill Cornell Medicine, New York, New York

³ New York Presbyterian Hospital, New York, New York

⁴ Department of Medicine, Weill Cornell Medicine, New York, New York

Correspondence

Zachary M. Grinspan, MD, MS, Department of Healthcare Policy & Research, Room LA 220, 402 East 67th Street, New York, NY 10065.
Email: zag9005@med.cornell.edu

Abstract

Objective: For children with asthma, emergency department (ED) visits are common, expensive, and often avoidable. Though several factors are associated with ED use (demographics, comorbidities, insurance, medications), its predictability using electronic health record (EHR) data is understudied.

Methods: We used a retrospective cohort study design and EHR data from one center to examine the relationship of patient factors in 1 year (2013) and the likelihood of frequent ED use (≥ 2 visits) in the following year (2014), using bivariate and multivariable statistics. We applied and compared several machine-learning algorithms to predict frequent ED use, then selected a model based on accuracy, parsimony, and interpretability.

Results: We identified 2691 children. In bivariate analyses, future frequent ED use was associated with demographics, co-morbidities, insurance status, medication history, and use of healthcare resources. Machine learning algorithms had very good AUC (area under the curve) values [0.66–0.87], though fair PPV (positive predictive value) [48–70%] and poor sensitivity [16–27%]. Our final multivariable logistic regression model contained two variables: insurance status and prior ED use. For publicly insured patients, the odds of frequent ED use were 3.1 [2.2–4.5] times that of privately insured patients. Publicly insured patients with 4+ ED visits and privately insured patients with 6+ ED visits in a year had $\geq 50\%$ probability of frequent ED use the following year. The model had an AUC of 0.86, PPV of 56%, and sensitivity of 23%.

Conclusion: Among children with asthma, prior frequent ED use and insurance status strongly predict future ED use.

KEYWORDS

asthma and early wheeze, biostatistics, electronic health records, evidence-based medicine and outcomes, predictive modeling

1 | INTRODUCTION

Nearly half of the seven million children with asthma in the United States (US) visit the emergency department (ED) each year for an acute exacerbation of their disease.¹ These visits are disruptive to children and their families, and expensive to the US healthcare system. Patients with poorly controlled severe asthma cost nearly \$5000 more per patient per year compared to average pediatric asthma costs.² Many of these visits may be avoidable since asthma is an ambulatory care

sensitive condition,³ that is, high quality outpatient care may reduce the frequency of ED visits.

Thus, a child with asthma who frequently visits the ED is likely to have limited access to care and/or poor disease control.⁴ Children at high risk for *future* frequent ED use are a vulnerable population who may benefit from improved outpatient care intervention. For example, well-implemented care management programs can reduce hospital use, lower costs, and improve clinical outcomes for children with chronic diseases.^{5,6} Care managers work directly with patients to

coordinate care, answer questions, provide information about diseases, reconcile medications, and help families navigate the medical system.⁷ However, it is unclear whom to target for such interventions.

There are several candidate predictors that are associated with frequent ED use in children with asthma, such as healthcare utilization history,⁸ demographics,⁹ co-morbidities,¹⁰ insurance status,¹¹ medication history,¹² and disease severity.⁹ Guidelines from the National Asthma Education and Prevention Program (NAEPP) and the Global Initiative for Asthma (GINA) summarize evidence for how several of these risk factors contribute to high resource utilization (such as frequent ED use, hospitalizations, ICU admissions) and in turn, rising healthcare costs.^{10,12} However, it is not clear how accurately these concepts can be used to predict frequent ED visits using data from electronic health record (EHR) systems.

In this study, we explored the predictability of frequent ED use among children with asthma using data from an EHR from one medical center. Our approach was twofold. First, we developed and evaluated several machine-learning algorithms to predict which children will use the ED two or more times in the next year. Second, we selected and described the algorithm that was the most parsimonious, interpretable, and accurate.

2 | MATERIALS AND METHODS

2.1 | Study design

We performed a retrospective cohort study at one center using outpatient EHR data from 2013 to predict frequent ED use in 2014 among children with asthma. The Weill Cornell Medical Center Institutional Review Board approved the study.

2.2 | Potential predictors

Through review of the literature and consultation with experts in health services research (ZG, LK, EA), pediatric pulmonary medicine (AS), pediatric emergency medicine (JK), and general pediatrics (ZG, AS, JK, EA), we generated a list of candidate variables potentially associated with frequent ED use. These roughly fell into four categories: (i) demographics; (ii) clinical factors and comorbidities; (iii) medication history; and (iv) prior healthcare utilization history (Table 1).^{1,3,4,8-20}

2.3 | Inclusion criteria

We included all individuals 18 years of age or younger who were patients in our center's outpatient general pediatric and pediatric pulmonary clinics and who had either a prescription for a short acting beta agonist²¹ and/or an ICD-9 code of 493.x (asthma)²² in 2013.

2.4 | Predictor variables

Data for each patient included (i) demographic variables (age, sex, race and ethnicity, home zip code aggregated to geographic regions, type of insurance, and an identified primary care physician in the EHR (yes/no)); (ii) prior use of health services (number of ED, inpatient, and outpatient

visits); (iii) clinical factors (Body Mass Index (BMI), IgE blood test, sensitization to common aeroallergens, and the presence of comorbid conditions); and (iv) medication prescription orders in the EHR. We aggregated ICD-9 codes into several comorbidities (appendix B), as well as an overall measure of medical complexity in children (Pediatric Medical Complexity Algorithm; PMCA).¹⁹

2.5 | Outcome

The commonly used threshold for "frequent ED use" of four or more visits²³⁻²⁵ was rare in our data (1.5% of patients). Thus, we defined frequent ED use as two or more (2+) all-cause ED visits in 2014. We focused on all ED visits rather than visits related to asthma. Our goal was to identify vulnerable individuals for enrollment into an intervention to improve outpatient care. We reasoned that frequent ED use, irrespective of cause, was a sufficiently compelling marker of vulnerability that would appeal to clinicians, administrators, and patients alike.

2.6 | Bivariate analysis

We investigated associations between predictor variables in 2013 and frequent all-cause ED visits in 2014. We compared continuous variables with the Wilcoxon test (due to skewed distributions), and categorical variables with the chi-square test. For categorical variables, we provide odds ratios and 95% confidence intervals as measures of effect size.

2.7 | Predictive modeling strategy

It was not clear a priori which algorithm would best predict frequent ED use. We selected a variety of algorithms in order to explore the overall predictability of ED use, to generate multiple candidate models for review, and to overcome potential limitations of regression, such as missing data, large numbers of variables, non-linear relationships, and unusual distributions commonly seen in healthcare data. These algorithms included (i) logistic regression with the best subsets algorithm for 1, 2, and 3 variables; (ii) LASSO (least absolute shrinkage and selection operator) logistic regression; (iii) CART (classification and regression tree); (iv) Random Forests; and (v) SVM (support vector machines). Additional information about these techniques appears in appendix A.

2.8 | Evaluation of models

We randomly selected individuals into a training set and test set of equal size. Next, we evaluated the predictive ability of models on the test set using the following criteria: (i) area under the receiver operating curve (AUC); (ii) calibration (appendix A); (iii) classification error; (iv) sensitivity; and (v) positive predictive value (PPV).²⁶ We qualified AUC, sensitivity, and PPV as follows: <0.55 poor, 0.56-0.69 fair, 0.7-0.79 good, 0.8-0.89 very good, and 0.9-1 excellent. We qualified calibration as follows: 0-4% excellent, 5-9% very good, 10-14% good, 15-20% fair, and >20% poor.²⁶

TABLE 1 Bivariate relationship between selected patient characteristics in year 1 (2013) and frequency of ED use in year 2 (2014), among 2691 children with asthma

Patient characteristics in year 1	0 or 1 visits in year 2 (N = 2507)	2 or More visits in year 2 (N = 184)	P-value	Odds ratio (95% CI)
Age				
Age	6 (3-11)	6 (3-11)	ns	-
Sex				
Female	996 (40%)	87 (47%)	0.05	1.36 (1.01-1.84)
Race				
White	940 (37%)	30 (16%)	<0.001	Ref
Black	323 (13%)	36 (20%)		3.49 (2.12-5.79)
Asian	151 (6%)	3 (2%)		0.62 (0.14-1.78)
American Indian/Alaskan	0 (0%)	5 (3%)		15.6 (4.64-47.1)
Other	696 (28%)	93 (50%)		4.17 (2.76-6.47)
Declined	354 (14%)	17 (9%)		1.50 (0.80-2.73)
Unknown	43 (2%)	0 (0%)		1.45 (0.23-5.04)
Ethnicity				
Not hispanic	1461 (58%)	80 (43%)	<0.001	Ref
Hispanic	520 (21%)	69 (38%)		2.42 (1.72-3.39)
Declined	421 (17%)	19 (10%)		0.82 (0.48-1.34)
Unknown	105 (4%)	16 (9%)		2.78 (1.52-4.81)
Borough of New York				
Manhattan	1149 (46%)	76 (41%)	<0.001	Ref
Bronx	303 (12%)	33 (18%)		1.64 (1.06-2.50)
Brooklyn	327 (13%)	19 (10%)		0.87 (0.51-1.44)
Queens	336 (13%)	48 (26%)		2.16 (1.46-3.15)
Staten island	24 (1%)	0 (0%)		0.00 (0.00-2.26)
Other	368 (15%)	8 (5%)		3.28 (0.14-0.65)
Insurance type				
Private	1749 (70%)	59 (32%)	<0.001	Ref
Public	635 (25%)	122 (66%)		5.69 (4.13-7.92)
Unknown	123 (5%)	3 (2%)		0.72 (0.17-1.99)
PCP identified in EHR				
Has PCP	1980 (79%)	130 (71%)	<0.01	Ref
Does not have PCP	16 (1%)	0 (0%)		0.00 (0.00-3.54)
Unknown	511 (20%)	54 (29%)		1.60 (1.14-2.23)
Encounters year 1				
ED Visits	0 [0 - 0]	1 [0.75 - 3]	<0.001	
Inpatient visits	0 [0 - 0]	0 [0 - 1]	<0.001	
Outpatient visits	3 [1 - 6]	6 [3 - 11]	<0.001	
Other visits	0 [0 - 0]	0 [0 - 0]	ns	
PMCA				
No chronic disease	1306 (52%)	67 (36%)	<0.001	Ref
Chronic disease	795 (32%)	72 (39%)		1.76 (1.25-2.49)
Medically complex	277 (11%)	42 (23%)		2.95 (1.95-4.42)
Other	129 (5%)	3 (2%)		0.47 (0.02-2.23)
BMI				
Normal	2080 (83%)	158 (86%)	<0.01	Ref

(Continues)

TABLE 1 (Continued)

Patient characteristics in year 1	0 or 1 visits in year 2 (N = 2507)	2 or More visits in year 2 (N = 184)	P-value	Odds ratio (95% CI)
Overweight	123 (5%)	8 (5%)		0.86 (0.38-1.67)
Obese	72 (3%)	11 (6%)		2.01 (0.99-3.71)
Unknown	232 (9%)	7 (3%)		0.40 (0.17-0.80)
IgE test				
Normal	47 (2%)	3 (2%)	ns	Ref
Abnormal	54 (2%)	6 (3%)		1.74 (0.43-8.60)
Not Sent	2406 (96%)	175 (95%)		1.14 (0.41-4.72)
Psychiatric comorbidities				
Anxiety	1 (0.03%)	1 (0.5%)	ns	13.69 (0.54-347)
Depression	12 (0.5%)	2 (1%)	ns	2.28 (0.35-8.46)
Medical comorbidities				
Allergic rhinitis	415 (16%)	39 (21%)	ns	1.35 (0.93-1.96)
Bronchiolitis	213 (8.5%)	15 (8%)	ns	0.95 (0.55-1.65)
Bronchitis	17 (1%)	0 (0%)	ns	0.00 (0.00-3.31)
Bronchopulmonary dysplasia	69 (3%)	7 (4%)	ns	1.39 (0.63-3.08)
Chronic bronchitis	4 (0.2%)	0 (0%)	ns	0.00 (0.00-20.8)
Cystic fibrosis	0 (0%)	0 (0%)	ns	0.00 (0.00-0.00)
Decreased tone	10 (0.4%)	2 (1%)	ns	2.74 (0.60-12.6)
Eczema	156 (6%)	18 (10%)	0.05	1.63 (0.97-2.72)
Increased tone	23 (1%)	5 (3%)	<0.05	3.06 (1.13-8.03)
Pneumonia	132 (5%)	17 (9%)	<0.05	1.83 (1.07-3.11)
Premature birth	35 (1%)	3 (2%)	ns	1.17 (0.35-3.84)
Vocal cord dysfunction	11 (0.4%)	3 (2%)	<0.05	3.76 (1.04-13.6)
Technology dependence				
Gastrostomy tube	40 (1.5%)	9 (5%)	<0.01	3.17 (1.51-6.64)
Tracheostomy	38 (1.5%)	7 (4%)	<0.05	2.57 (1.13-5.83)
Allergies				
Animal	253 (12%)	16 (9%)	ns	0.84 (0.49-1.44)
Chemical	19 (0.8%)	3 (2%)	ns	2.17 (0.63-7.40)
Cockroach	47 (2%)	8 (4%)	<0.05	2.37 (1.10-5.11)
Drug	294 (12%)	27 (15%)	ns	1.29 (0.84-1.98)
Dust	140 (6%)	14 (8%)	ns	1.39 (0.78-2.46)
Food	506 (20%)	42 (23%)	ns	1.17 (0.81-1.67)
Fungus	5 (0.2%)	0 (0%)	ns	0.00 (0.00-14.9)
Herbal	12 (0.5%)	0 (0%)	ns	0.00 (0.00-4.92)
Latex	13 (0.5%)	2 (1%)	ns	2.10 (0.47-9.41)
Mold	1574 (62%)	102 (55%)	<0.05	0.74 (0.54-0.99)
Other	111 (4.4%)	9 (5%)	ns	1.11 (0.55-2.22)
Pollen	115 (5%)	11 (6%)	ns	1.32 (0.69-2.50)
Seasonal	20 (0.7%)	2 (1%)	ns	1.36 (0.31-5.89)
Trees	1663 (66%)	109 (59%)	0.06	0.73 (0.54-1.00)
Medications prescribed				
Asthma medication device	565 (23%)	54 (29%)	<0.05	1.43 (1.02-1.98)
Inhaled drug combination	76 (3%)	8 (4%)	ns	1.45 (0.69-3.06)

(Continues)

TABLE 1 (Continued)

Patient characteristics in year 1	0 or 1 visits in year 2 (N = 2507)	2 or More visits in year 2 (N = 184)	P-value	Odds ratio (95% CI)
Inhaled rescue	2274 (91%)	177 (96%)	<0.05	2.59 (1.20-5.58)
Inhaled steroids	1037 (41%)	82 (45%)	ns	1.14 (0.84-1.54)
Leukotrienes	294 (12%)	22 (12%)	ns	1.02 (0.64-1.62)
Cromolyn	4 (0.2%)	2 (1%)	<0.05	6.87 (1.25-37.7)
Oral steroids	528 (21%)	43 (23%)	ns	1.14 (0.80-1.63)
Other medications	1848 (74%)	157 (85%)	<0.001	2.07 (1.36-3.20)

ns, not significant. All values are either Median [Inter-quartile Range] or N (% of patients within the category of ED use in Year 2). P-values and Odds Ratios (95%CI) determined by Pearson correlation for continuous variables (using ED visits in year 2 as a continuous variable) and chi-square test for categorical variables.

2.9 | Model selection

We selected a final model based on parsimony (ie, small number of variables), accuracy, and interpretability. We operationalized *parsimony* to mean four or fewer variables, *interpretability* to mean that our research team clearly understood the meaning of each included variable, and *accuracy* to mean that the AUC was within 0.02 of the best performing model.

2.10 | Threshold selection

A predictive model can be tuned to emphasize sensitivity or PPV by adjusting the threshold used to identify individuals for intervention. We created a table to illustrate the potential tradeoffs of using different thresholds. For each threshold, we report the number of individuals that would be enrolled, the number correctly predicted to have 2+ ED visits in year 2, the PPV, and the sensitivity. We also report the mean and total number of ED visits by the selected cohort.

2.11 | Statistical software

We used the R software (version 3.1.1)²⁷ with additional packages detailed in appendix A.

3 | RESULTS

Description of cohort

We identified 2691 patients. The median age was 6 [Interquartile Interval: 3-11]. Sixty percent were male. A plurality (36%) was White, 13% were Black or African American, 6% Asian, and 1% American Indian or Alaska National. A large group (29%) fell into the category "Other Combinations not Described." Nearly half (46%) lived in Manhattan, 40% in other boroughs of New York City, and the remaining 14% lived outside New York City. Nearly a quarter (22%) had Hispanic, Latin, or Spanish origin. Two thirds (67%) had private insurance, 28% were publicly insured, and the remaining 5% were unknown.

TABLE 2 Evaluation of predictive modeling techniques on a test set of 1346 patients

Technique	# of variables	AUC	Calibration (%)	Classification error (%)	Sensitivity (%)	PPV (%)	Variables
Logistic Regression	1	0.84	13	5.9	27	59	(1) ED visits in Year 1
Best subsets	2	0.86	13	5.9	23	56	(1) ED visits in year 1 (2) Type of insurance
	3	0.86	14	6.1	23	56	(1) ED visits in year 1, (2) Type of insurance, (3) Cromolyn use
LASSO	2	0.87	20	5.8	16	70	(1) ED visits in year 1, (2) Type of insurance
Random forests	all (52)	0.83	4.1	6.4	19	52	All
CART	2	0.66	15	6.6	24	48	(1) ED visits in year 1, (2) Type of insurance
Support vector machines	all (52)	0.81	14	6.0	19	62	All

3.1 | Prevalence of frequent ED users

There were 184 (6.8%) frequent ED users (2+ visits) during Year 2 (2014) among the 2691 individuals in the full dataset.

3.2 | Bivariate associations between predictors and frequent ED use

Several potential predictor variables were significantly associated with frequent ED use in Year 2 (2014) (Table 1). Girls had a 1.36 [95% confidence interval 1.01-1.84] times higher odds to be frequent ED users compared to boys. Race and ethnicity were also associated with frequent ED use. Compared with Whites, Blacks had higher odds of frequent ED use (3.5 [2.1-5.8]), as did those with "Other" race (4.2 [2.8-6.5]). Hispanics had higher odds (2.4 [1.7-3.4]) as did those with "Unknown" ethnicity (2.8 [1.5-4.8]). All five children with American Indian/Alaskan race were frequent ED users. Living in areas of New York City with lower median incomes was associated with higher odds of frequent ED use (Bronx [1.6 [1.06-2.5]] and Queens [2.2 [1.5-3.2]], compared with Manhattan). Individuals with public insurance had a 5.7 [4.1-7.9] times higher odds of frequent ED use than individuals with commercial insurance. Other factors associated with ED use were medical complexity, obesity, and multiple health care visits (ED, outpatient, or inpatient) in the baseline year; a history of increased tone (representative of neuromuscular disease), eczema, pneumonia, vocal cord dysfunction, technology dependence (gastrostomy tube or tracheostomy), and allergies to cockroaches. Among specific allergies

TABLE 3 Logistic regression model for frequent ED Use (2+ Visits in 2014)

Baseline probability, private insurance, zero ED visits in 2013 (95%CI)	2.2 (1.7-3)%
Factor	Odds ratio (95%CI)
Private insurance	Ref
Public insurance	3.11 (2.17-4.45)
Unknown insurance	0.42 (0.11-1.62)
Each ED visit in 2013	2.22 (1.97-2.51)

reviewed, sensitivities to mold, or tree pollen was less common among frequent ED users. Finally, frequent ED users were more likely to have received prescriptions for asthma medication devices such as spacers and nebulizers, inhaled rescue drugs, mast cell stabilizers, and medications for other diseases (Table 1).

3.3 | Predictability of frequent ED use

When we applied the predictive algorithms, the AUC ranged from 0.81-0.87 for most algorithms, indicating very good predictability. (The performance of CART was fair [AUC 0.66]). The PPVs were fair to poor (48-70%). Sensitivity was uniformly poor (16-27%). Calibration was fair to good (13-20%), with the exception of Random Forests, which had excellent calibration (4.1%). Classification error ranged from 5.8-6.6%. (Table 2)

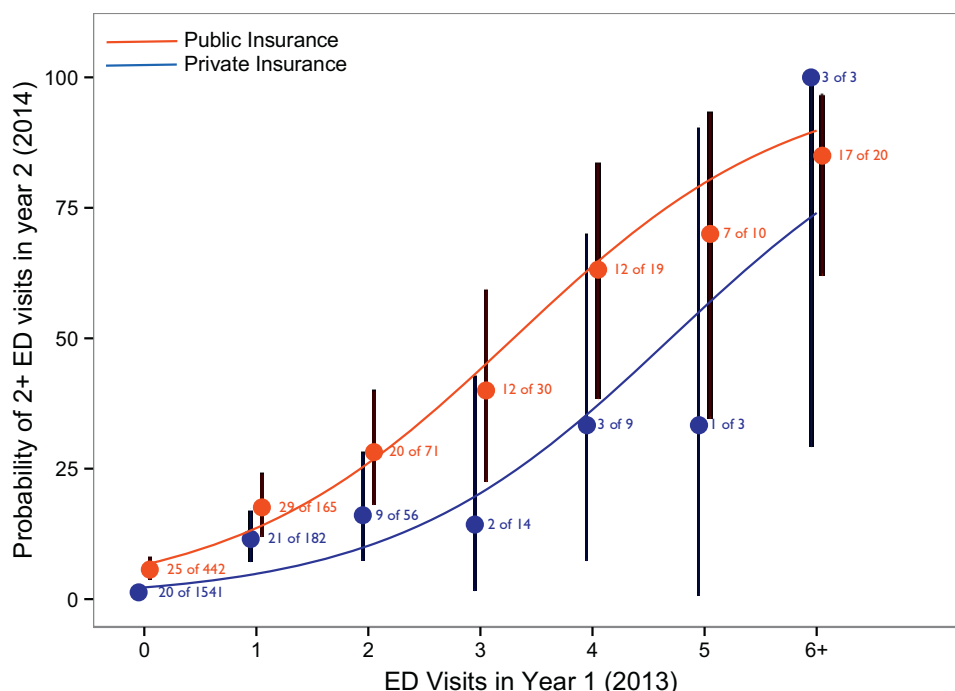


FIGURE 1 Relationship between number of ED visits in year 1 and probability of frequent ED use in year 2, stratified by insurance type. The x-axis denotes the number of ED visits in baseline Year 1. The y-axis denotes the corresponding proportion of patients (%) who were frequent ED users in the following year. Red data is for children with public insurance, blue for children with private insurance. Each point is annotated "a of b," a is the number of individuals who visited the ED two or more times in year 2 (2014), b is the number of individuals who visited the ED the given number of times in year 1 (2013). The solid curve is the best fit from the logistic regression model in Table 3. Error bars are the 95% confidence interval of the binomial distribution of the observed proportion

TABLE 4 Threshold selection table for care management enrollment recommendations

# ED visits in year 1 (score threshold)	# to enroll	"Hits" (pts with 2+ ED visits in year 2)	PPV (%)	Sensitivity (%)	Total # of ED visits in year 2	Mean # of ED visits per enrollee in year 2
Public insurance						
6	20	17	85	14	89	4.5
5	30	24	80	20	109	3.6
4	49	36	73	30	185	3.8
3	79	48	61	39	281	3.6
2	150	68	45	56	391	2.6
1	315	97	31	80	527	1.7
0	757	122	16	100	1026	1.4
Private insurance						
6	3	3	100	5	32	10.7
5	6	4	67	7	37	6.2
4	15	7	47	12	41	2.7
3	29	9	31	15	71	2.4
2	85	18	21	31	157	1.8
1	267	39	15	66	291	1.1
0	1808	59	3	100	1906	1.1

3.4 | Selection of final predictors

Two predictors consistently appeared in the algorithms: (i) Number of ED visits in Year 1; and (ii) Type of Insurance (Table 2). Our final model included only these two variables, satisfying our parsimony criterion. Our research team indicated clear understanding of these concepts, satisfying our interpretability criterion. A generalized linear model with a logit link (ie, a logistic regression model) with these two variables had high AUC (0.86), good calibration (13%), fair PPV (56%), though poor sensitivity (23%). The classification error was 5.9%, which was among the most accurate values. The AUC was within 0.01 of the best performing model (0.87), satisfying our criteria of accuracy.

According to the final model, the probability that a child with private insurance and no ED visits in 2013 will visit the ED 2 or more times in 2014 is 2.2% [1.7-3]. The odds of frequent ED use are 3.1 [2.2-4.5] higher if the child has public insurance, and 2.2 [2-2.5] times higher for each ED visit in 2013 (Table 3).

3.5 | Current versus future ED visits, stratified by insurance

Visual inspection of the data suggested a nearly linear relationship between ED use in year 1 (2013) and the probability of two or more ED visits in year 2 (2014), with different slopes depending on the insurance type (Fig. 1). Among the publicly insured, individuals with 4+ ED visits in 2013 had 50% or higher probability of frequent ED use (2+ visits) in 2014 (Fig. 1). Seventeen out of twenty patients with 4+ ED visits in 2013 had 2+ visits in 2014. However, among the privately insured, the threshold was higher: individuals with 6+ ED visits in 2013 had 50% or higher probability of 2+ visits in 2014 (Fig. 1). All three

children with 6+ ED visits in 2013 were also frequent ED users (2+ visits) in 2014.

3.6 | Threshold selection table

A threshold selection table (Table 4) illustrates the expected pattern of future ED use in groups of different sizes. For example, a care management team with the capacity for 80 publicly insured patients would enroll children with three or more ED visits in the prior year (ie, a threshold of 3). Without intervention, the enrolled cohort of 79 individuals would be expected to visit the ED 281 times (mean of 3.6 visits per enrollee). Of the 79, 48 would be expected to be frequent ED users (ie, a PPV of 61%). However, the group of 79 would only include 39 of 122 expected frequent ED users (ie, a sensitivity of 39%).

4 | DISCUSSION

4.1 | Summary

We found several clinical, demographic, and healthcare utilization factors that were significantly associated with future frequent ED use. We developed a parsimonious, interpretable, and accurate predictive model to identify children at high risk for future frequent ED use, based on prior ED visit history and insurance status. This simple, two-variable model had a high AUC (0.86), good calibration (13%), fair PPV (56%), though poor sensitivity (23%). Its performance was similar to more sophisticated multivariable models and machine learning techniques. Publicly insured patients with 4+ ED visits and privately insured patients with 6+ ED visits in a year had $\geq 50\%$ probability of frequent ED use in the following year.

4.2 | Comparison to published literature

Our bivariate analysis found associations between frequent ED use and several potential predictor variables, including race, insurance status, geography, insurance type, medical complexity, and history of frequent healthcare visits (ED, inpatient, and outpatient visits). Many of these results are consistent with past findings, summarized in the literature cited in the NAEPP and GINA guidelines.^{10,12} A 2001 California study found that African American and Latino children are more likely to visit the ED for asthma than White non-Hispanic children.⁹ Insurance status is, to some degree, a marker of socioeconomic status, and has been linked to increases in asthma-related ED visits,^{9,16} as well as poor access to pediatric subspecialists.²⁸ In New York City, geographic areas (Bronx in particular) are associated with higher rates of environmental exposure to household aeroallergens, which can exacerbate asthma.²⁹ Medical complexity is associated with increased use of health services, particularly in the ED.³⁰ Finally, in general, frequent ED use is associated with frequent use other health services, both in a general³¹ as well as in an asthma population.⁸

Interestingly, the percentage of frequent ED users in our study cohort (6.8%) was lower than that of other published studies. Published rates include 32% in a group of Latino children³² and 17% in a group of Medicaid beneficiaries.³³ It is unclear if our low rate is due to factors related to our population (ie, the racial, ethnic, or insurance makeup) or health system (ie, access to primary and specialty care), or if we have underreported ED use due to high levels of fragmented care in New York City (ie, people seek care in multiple health systems).^{34,35}

4.3 | Simplicity

Our final model had only two variables, consistent with our previous work demonstrating that simple predictive models can be just as accurate as more complex ones.²⁶ It is unclear why there was little additional predictive information from variables apart from insurance status and prior ED visits. It is possible that many of these variables were statistically collinear, or in the causal pathway between insurance status, prior ED use, and future ED use.

4.4 | Electronic health records as a data source

Our work demonstrates that EHRs can be a valuable source of data for pediatric asthma research. Many health services research studies use discharge data from statewide or national databases.³⁶ Such publicly available databases have notable strengths: relatively low data collection costs compared to medical record abstraction or surveys, information on uninsured patients not available from third party payers, and greater reliability than self-reported medical expenditures. However, there are inconsistencies in how providers, government agencies, and payers report specific data elements. Additionally, there may be hospital-specific errors in data reporting that can lead to data quality problems.

EHRs have several advantages over administrative databases. In addition to billing codes, EHRs contain detailed clinical information such as test results, severity of illness, behavioral risk factors, and free text within clinical notes.³⁷ In our study, many significant variables that

were extracted from the EHR are not typically reported in administrative data such as BMI, and allergic sensitization. Furthermore, EHR data is more readily translatable into action, as the data can be used for clinical decision support.

4.5 | Implications for clinical care and care management

Our results have implications for clinical practice. We described an easily implemented decision rule that care management programs can use to enroll patients for enhanced outpatient services. Although many of the bivariate associations do not appear in the final model, they provide deeper insight into the kinds of interventions that might be effective for frequent ED users. For example, children with asthma who are dependent on technologies (a comorbidity) such as gastrostomy or tracheostomy tubes may require more involved clinical interventions, such as incorporation of pulmonology consultation into a medical complexity clinic. Other factors, such as access to care, may be amenable to health services interventions such as care management or increased social work staffing. There is likely a spectrum of reasons why selected children frequently use the ED⁷—thus it is important that care managers tailor plans for each enrolled individual.

4.6 | Limitations

Several limitations merit discussion.

First, our data only captures ED visits to one medical center, which may have caused undercounting of ED visits and limited generalizability. People in New York City frequently use multiple centers for care,³⁸ and thus we likely underestimated overall ED use in our population.^{38,39} In the future, gathering health utilization data from regional health information exchanges may provide a more comprehensive picture of each child's ED visits. Additionally, the generalizability of these predictions is unclear. Since our source of data is one medical center, it is unknown how well our model would work at other centers.

Second, we were unable to operationalize certain concepts that are likely important in asthma care. Concepts like asthma severity or insecure housing typically appear in free text rather than in structured fields within an EHR system. Successful operationalization of these concepts by using natural language processing might improve predictive accuracies. In addition, we were unable to accurately calculate the ratio of controller to total asthma medicine usage, which has been associated with ED use.¹² Future work that merges EHR data with pharmacy data may allow assessment of this concept.

Third, though our model identifies a cohort of individuals at risk for frequent ED use, it is unclear which of these individuals specifically would benefit from intervention. Because asthma is an ambulatory care sensitive condition, it is reasonable to expect that improved outpatient care would reduce ED use. However, identification of high risk individuals does not necessarily identify which individuals will benefit by such interventions.⁴⁰

Finally, the definition of asthma we used to identify patients (a prescription for a short acting beta agonist and/or an ICD-9 code of 493.x) has limitations. It may include false positives (ie children with

neuromuscular disease who use albuterol to improve airway clearance or young children with wheezing for reasons such as bronchiolitis) as well as false negatives, due to undercoding of asthma.²² Inclusion of children who received an Albuterol prescription may be confounded by children to whom the medication was incorrectly prescribed. Finally, even though the number of children with other chronic respiratory diseases that can present with wheezing, such as cystic fibrosis, neuromuscular disease, and chronic bronchitis, was small, we may have introduced a degree of misclassification by not excluding them from the final analysis.

5 | CONCLUSION

Children with asthma who frequently visit the ED are likely to continue doing so, though the likelihood depends on insurance type.

ACKNOWLEDGMENTS

We would like to thank Dr. Jessica Ancker and Dr. Steve Johnson for their valuable advice, and Baria Hafeez for her assistance with the manuscript. Additionally, we would like to thank David Woolf and Craig Suchodolski from the Epic Data Core team at Weill Cornell Medicine for data extraction from the EHR, and Dr. Lyubov Tmanova for help with library services.

CONFLICT OF INTEREST

Authors have no potential conflict of interest to disclose.

REFERENCES

1. Akinbami LJ, Moorman JE, Bailey C, et al. Trends in Asthma Prevalence, Health Care Use, and Mortality in the United States, 2001–2010. NCHS Data Brief. 2012 [accessed 2016 December 6]; 94. Available online at: <https://www.cdc.gov/nchs/data/databriefs/db94.pdf>
2. Patel K, Farmer S, George M, McStay F, McClellan M. Pediatric asthma: an opportunity in payment reform and public health. Health Affairs Blog. 2014 [accessed 2016 December 6]. Available online at: <https://healthaffairs.org/blog/2014/09/18/pediatric-asthma-an-opportunity-in-payment-reform-and-public-health/>
3. Trachtenberg AJ, Dik N, Chateau D, Katz A. Inequities in ambulatory care and the relationship between socioeconomic status and respiratory hospitalizations: a population-based study of a canadian city. *Ann Fam Med*. 2014;12:402–407.
4. LaCalle E, Rabin E. Frequent users of emergency departments: the myths, the data, and the policy implications. *Ann Emerg Med*. 2010;56:42–48.
5. McCarthy D, Cohen A, Johnson MB. Gaining ground care management programs to reduce hospital admissions and readmissions among chronically ill and vulnerable patients. The Commonwealth Fund, Case Study Series Synthesis. 2013 [accessed 2016 November 24];5. Available online at: https://www.commonwealthfund.org/~media/files/publications/case-study/2013/jan/1658_mccarthy_care_transitions_synthesis_v2.pdf
6. Sweeney L, Halpert A, Waranoff J. Patient-centered management of complex patients can reduce costs without shortening life. *Am J Manag Care*. 2007;13:84–92.
7. Enard KR, Ganelin DM. Reducing preventable emergency department utilization and costs by using community health workers as patient navigators. *J Healthc Manag*. 2013;58:412–427.
8. Tolomeo C, Savrin C, Heinzer M, Bazyz-Asaad A. Predictors of asthma-related pediatric emergency department visits and hospitalizations. *J Asthma*. 2009;46:829–834.
9. Wright K. Disparities and predictors of emergency department use among California's African American, Latino, and White children, aged 1–11 years, with asthma. *Ethn Dis*. 2009;19:71–77.
10. National Asthma Education and Prevention Program. Expert Panel Report 3: Guidelines for the diagnosis and management of asthma. 2007 [accessed 2017 February 24]. Available online at: <https://www.nhlbi.nih.gov/files/docs/guidelines/asthgdln.pdf>
11. Ferris TG, Crain EF, Oken E, Wang L, Clark S, Camargo CA. Insurance and quality of care for children with acute asthma. *Ambul Pediatr*. 2001;1:267–274.
12. Global Initiative for Asthma. Global strategy for asthma management and prevention. 2017; [accessed 2017 February 25]. Available online at: <https://ginasthma.org/2017-gina-report-global-strategy-for-asthma-management-and-prevention/>
13. Association AL. Asthma & Children Fact Sheet. 2014; [accessed 2015 August 1]. Available online at: <https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/asthma/learn-about-asthma/asthma-children-facts-sheet.html>
14. Heffernan PM. 2012. Socioeconomic status, asthma, and emergency department use in Ontario, Canada [thesis]. [St. John's (Canada)]: Memorial University.
15. Hygiene NYCDohaM. The New York City Community Air Survey. Results from Year One Monitoring 2008–2009. 2011 [accessed 2015 August 1]. Available online at: <https://www.nyc.gov/html/doh/downloads/pdf/eode/comm-air-survey-report.pdf>
16. Ortega AN, Belanger KD, Paltiel AD, Horwitz SM, Bracken MB, Leaderer BP. Use of health services by insurance status among children with asthma. *Med Care*. 2001;39:1065–1074.
17. Prevention CfDca. Asthma in the US. 2011 [accessed 2015 August 1]. Available online at: <https://www.cdc.gov/vitalsigns/asthma>
18. Prevention CfDca. Asthma 2015. [accessed 2015 August 3]. Available online at: <https://www.cdc.gov/nchs/fastats/asthma.htm>
19. Simon TD, Cawthon ML, Stanford S, et al. COE4CCN medical complexity working group. Pediatric medical complexity algorithm: a new method to stratify children by medical complexity. *Pediatrics*. 2014;133:e1647–e1654.
20. Szeffler SJ, Zeiger RS, Haselkorn T, et al. Economic burden of impairment in children with severe or difficult-to-treat asthma. *Ann Allergy Asthma Immunol*. 2011;107:110–119. e11.
21. Prosser R, Carleton B, Smith A. The comorbidity burden of the treated asthma patient population in British Columbia. *Chronic Dis Can*. 2010;30:46–55.
22. Juhn Y, Kung A, Voigt R, Johnson S. Characterisation of children's asthma status by ICD-9 code and criteria-based medical record review. *Prim Care Respir J*. 2011;20:79–83.
23. Byrne M, Murphy AW, Plunkett PK, McGee HM, Murray A, Bury G. Frequent attenders to an emergency department: a study of primary health care use, medical profile, and psychosocial characteristics. *Ann Emerg Med*. 2003;41:309–318.
24. Hunt KA, Weber EJ, Showstack JA, Colby DC, Callahan ML. Characteristics of frequent users of emergency departments. *Ann Emerg Med*. 2006;48:1–8.
25. Ledoux Y, Minner P. Occasional and frequent repeaters in a psychiatric emergency room. *Soc Psychiatry Psychiatr Epidemiol*. 2006;41:115–121.
26. Grinspan ZM, Shapiro JS, Abramson EL, Hooker G, Kaushal R, Kern LM. Predicting frequent ED use by people with epilepsy with health information exchange data. *Neurology*. 2015;85:1031–1038.

27. Locker TE, Baston S, Mason SM, Nicholl J. Defining frequent use of an urban emergency department. *Emerg Med J*. 2007;24:398–401.
28. Bisgaier J, Rhodes KV. Auditing access to specialty care for children with public insurance. *N Engl J Med*. 2011;364:2324–2333.
29. Warman K, Silver EJ, Wood PR. Modifiable risk factors for asthma morbidity in Bronx versus other inner-city children. *J Asthma*. 2009;46:995–1000.
30. O'Mahony L, O'Mahony DS, Simon TD, Neff J, Klein EJ, Quan L. Medical complexity and pediatric emergency department and inpatient utilization. *Pediatrics*. 2013;131:e559–e565.
31. Billings J, Raven MC. Dispelling an urban legend: frequent emergency department users have substantial burden of disease. *Health Aff (Millwood)*. 2013;32:2099–2108.
32. Canino G, Garro A, Alvarez MM, et al. Factors associated with disparities in emergency department use among Latino children with asthma. *Ann Allergy Asthma Immunol*. 2012;108:266–270.
33. Fredrickson DD, Molgaard CA, Dismuke SE, Schukman JS, Walling A. Understanding frequent emergency room use by Medicaid-insured children with asthma: a combined quantitative and qualitative study. *J Am Board Fam Pract*. 2004;17:96–100.
34. Grinspan ZM, Abramson EL, Banerjee S, Kern LM, Kaushal R, Shapiro JS. People with epilepsy who use multiple hospitals; prevalence and associated factors assessed via a health information exchange. *Epilepsia*. 2014;55:734–745.
35. 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*. 2016;20:99–102.
36. Ohno-Machado L. Realizing the full potential of electronic health records: the role of natural language processing. *J Am Med Inform Assoc*. 2011;18:539.
37. Crockett DK. Why Predictive Modeling in Healthcare Requires A Data Warehouse. Health Catalyst—White Paper. 2014; [accessed 2016 November 21]. Available online at: <https://www.healthcatalyst.com/predictive-modeling-healthcare-requirements/>
38. Shapiro JS, Johnson SA, Angiolillo J, Fleischman W, Onyile A, Kuperman G. Health information exchange improves identification of frequent emergency department users. *Health Aff (Millwood)*. 2013;32:2193–2198.
39. Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. *Arch Intern Med*. 2010;170:1989–1995.
40. Lewis GH. Impactability models: identifying the subgroup of high-risk patients most amenable to hospital-avoidance programs. *Milbank Q*. 2010;88:240–255.
41. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
42. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York, NY: Springer; 2009.
43. Eubanks AL, Aguirre B, Bourgeois JA. Severe acute hyperammonemia after brief exposure to valproate. *Psychosomatics*. 2008;49:82–83.
44. Liaw A, Wiener M. Classification and regression by randomforest. *R News*. 2002;2:18–22.
45. Alexander MK, Bourgeois A, Goodman LR, Higgs M. An acute/critical care nurse practitioner program. *Nursing Manag*. 1996;27:28B–28C.
46. Warnes GR Bolker B, Gorjanc G, et al. Gdata: various R programming tools for data manipulation. R package version 2.17.0. <http://CRAN.R-project.org/package=gdata>. 2015.
47. Millborrow S. Rpart.plot: plot rpart Models. R package version 1.5.2. 2015.
48. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCr: visualizing classifier performance in R. *bioinformatics*. 2005;21:3940–3941.
49. Kuhn M. Caret: classification and regression training. R package version 6.0-47.
50. Wickham H. *ggplot2: elegant graphics for data analysis*. New York, NY: Springer; 2009.
51. Dowle M, Short T, Lianoglou S. Data.table: extension of data.frame. R package version 1.9.4. 2014.
52. Lumley T. Leaps: regression subset selection. R package version 2.9. 2009.
53. Locker TE, Baston S, Mason SM, Nicholl J. Defining frequent use of an urban emergency department. *EMJ*. 2007;24:398–401.
54. Robin X, Turck N, Hainard A, et al. pROC: an open-source packager for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
55. Aragon TJ. Epitools: epidemiology tools. R package version 0.5-8.

How to cite this article: Das LT, Abramson EL, Stone AE, Kondrich JE, Kern LM, Grinspan ZM. Predicting frequent emergency department visits among children with asthma using EHR data. *Pediatric Pulmonology*. 2017;52:880–890. <https://doi.org/10.1002/ppul.23735>

APPENDIX A

Logistic Regression: Best Subsets. We created 1, 2, and 3-variable models using the best subsets algorithm and logistic regression. This technique identifies the best-fitting regression model by examining all combinations of predictor variables that are specified. This technique is computationally feasible for large numbers of variables, when the total the number of allowed predictor variables is restricted to a smaller number.⁴¹

Regularized Logistic Regression: Lasso. The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that involves penalizing the absolute size of regression coefficients as they grow too large. By constraining the sum of the absolute values of the estimates, some of the parameter estimates become zero, thereby limiting the number of variables in the final model. The LASSO requires a parameter λ , which we determined via cross-validation within the training set.^{41,42}

Decision Trees: We used two decision tree algorithms. CART (Classification and Regression Tree) algorithm constructs and prunes a decision tree automatically.⁴³ Since our outcome variable is binary, we used the classification method in the algorithm. Random Forests builds a series of trees using random subsets of the training data.⁴⁴ When a new example is introduced, each tree of the Random Forest “votes” on the outcome. The outcome with the majority vote is the final prediction.

Geometric: Support vector machines is a supervised learning model based on geometric principles. The algorithm represents every example, belonging to one of two categories, as a point in space such that examples of either category are separated by a clear gap that is as wide as possible. When a new example is introduced the predicted outcome is determined based on which side of the gap it appears.⁴⁵

Definition of Calibration: We defined “calibration” as a measurement of the error between predicted probability and observed proportion, calculated as follows:

1. Split the individuals into 5 groups, based on evenly distributed bins of predicted probability (ie, 0-20%, 20-40%, etc).
2. Calculate the mean absolute difference between the predicted probability and the observed proportion of frequent ED use, across these five groups.

Statistical Packages. We used the following packages in R: “gdata,”⁴⁶ “randomForest,”⁴⁴ “rpart,”⁴³ “rpart.plot,”⁴⁷ “glmnet,”⁴¹ “ROCR,”⁴⁸ “caret,”⁴⁹ “e1071,”⁴⁵ “ggplot2,”⁵⁰ “data table,”⁵¹ “leaps,”⁵² “stats,”⁵³ “pROC,”⁵⁴ and “epictools.”⁵⁵

APPENDIX B

List of comorbidities and corresponding ICD-9 codes

Condition	ICD-9 code
Psychiatric comorbidities	
Anxiety	300
Depression	296.2, 296.3, 296.5, 300.4, 309.0, 311
Medical comorbidities	
Allergic rhinitis	477.9
Bronchiolitis	466.1
Bronchitis	466.0, 490
Broncho-pulmonary dysplasia	496, 518.89, 770.7, 748.6
Chronic bronchitis ^a	491.x
Cystic fibrosis	277
Decreased Tone ^b	779.89, 781.3, 358.x, 359.x, 335.x
Eczema	692.9
Increased tone ^b	342.x, 343.x
Pneumonia	480-488
Premature birth	765.0, 765.1, 765.21-765.27
Vocal cord dysfunction	478.3, 478.5
IgA Deficiency	279.01
Gastroesophageal reflux disease	530.81
Technology dependence	
Gastrostomy tube	536.40-536.42, 536.49, V44.1, V55.1
Tracheostomy	312, 312.1, 312.9, V44.0, V55.1

^aInstead of bronchiolitis obliterans, we created a Chronic Bronchitis category.

^bWe created two categories for neuromuscular disease—increased tone and decreased tone.