Preventive Medicine Reports 10 (2018) 55–61

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

Optimizing community-level surveillance data for pediatric asthma management

Wande O. Benka-Coker[a,*], Sara L. Gale[b], Sylvia J. Brandt[c], John R. Balmes[d,e], Sheryl Magzamen[a]

[a] Department of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, CO, USA
[b] Division of Epidemiology, School of Public Health, University of California, Berkeley, CA, USA
[c] Department of Resource Economics, University of Massachusetts, Amherst, MA, USA
[d] Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA, USA
[e] Division of Occupational and Environmental Medicine, University of California, San Francisco, CA, USA

ARTICLE INFO

Keywords:
Asthma
Classification
Risk stratification
Statistical data analysis
Disease management

ABSTRACT

Community-level approaches for pediatric asthma management rely on locally collected information derived primarily from two sources: claims records and school-based surveys. We combined claims and school-based surveillance data, and examined the asthma-related risk patterns among adolescent students.

Symptom data collected from school-based asthma surveys conducted in Oakland, CA were used for case identification and determination of severity levels for students (high and low). Survey data were matched to Medicaid claims data for all asthma-related health care encounters for the year prior to the survey. We then employed recursive partitioning to develop classification trees that identified patterns of demographics and healthcare utilization associated with severity.

A total of 561 students had complete matched data; 86.1% were classified as high-severity, and 13.9% as low-severity asthma. The classification tree consisted of eight subsets: three indicating high severity and five indicating low severity. The risk subsets highlighted varying combinations of non-specific demographic and socioeconomic predictors of asthma prevalence, morbidity and severity. For example, the subset with the highest class-prior probability (92.1%) predicted high-severity asthma and consisted of students without prescribed rescue medication, but with at least one in-clinic nebulizer treatment. The predictive accuracy of the tree-based model was approximately 66.7%, with an estimated 91.1% of high-severity cases and 42.3% of low-severity cases correctly predicted.

Our analysis draws on the strengths of two complementary datasets to provide community-level information on children with asthma, and demonstrates the utility of recursive partitioning methods to explore a combination of features that convey asthma severity.

1. Introduction

Despite recent data showing stabilization in asthma prevalence (Akinbami et al., 2016), childhood asthma morbidity and mortality remain high, particularly in urban communities (Keet et al., 2015). Due to a variety of reasons, including socioeconomic disparities and access to healthcare, asthma diagnosis and assessment of asthma severity are problematic in low-income and non-White populations (Akinbami et al., 2016; Mitchell et al., 2016; Akinbami et al., 2009). The resulting poor asthma control in these groups is characterized largely by increased hospitalizations, emergency department visits and medical costs, health outcomes considered to be avoidable with appropriate management (Vital, 2011; NHLBI, 2007; Barnett and Nurmagambetov, 2011; Gupta et al., 2006).

In conjunction with proper clinical management, community-level surveillance approaches have been suggested as an appropriate strategy to reduce asthma-related morbidity. Community-level data can provide geographically resolved information on asthma prevalence and asthma-related morbidity. These data can add to the general understanding of challenges and solutions for local asthma management, while being detailed enough to decipher unique community patterns of determinants of disease morbidity, and inform asthma control and management efforts (Asher et al., 1995; Busi et al., 2012; Magzamen et al., 2005).

Presently, community-level asthma prevalence data are limited to two commonly described sources: administrative and healthcare claims records, and school-based surveys. Surveillance data based on

* Corresponding author at: Department of Environmental and Radiological Health Sciences, Colorado State University, 1681 Campus Delivery, Fort Collins, CO 80523, USA.
E-mail address: wande.benka-coker@colostate.edu (W.O. Benka-Coker).

https://doi.org/10.1016/j.pmedr.2018.02.004
Received 17 July 2017; Received in revised form 3 January 2018; Accepted 5 February 2018
Available online 08 February 2018

administrative data are generally available and accessible to researchers; these data are particularly useful when the outcomes of interest are objective metrics of healthcare utilization (Morris et al., 1997; Reeves et al., 2006; Roberts et al., 2006; Labrèche et al., 2008; Walsh-Kelly et al., 2008; Dombkowski et al., 2012; Smith et al., 2005), and by extension, important risk factors for asthma severity and management (NHLBI, 2007; Reeves et al., 2006; Roberts et al., 2006). However, claims data may overrepresent the most severe and/or suboptimally managed cases of asthma, rather than reflect the total burden of disease in a community (Piecoro et al., 2001; Dombkowski et al., 2005). Additionally, measures of severity such as a history of symptom type and frequency are largely not found in the claims databases. Results of the objective metrics of asthma severity, such as pulmonary function testing, also tend to be absent from these sources. Consequently, administrative and claims data may not serve as an optimal stand-alone surveillance system for capture of the community burden of asthma. The challenge of use of these data is developing supplementary surveillance measures to augment the highlighted gaps.

School settings represent an alternative ingress point for health surveillance data due to extensive access to members of the target population (Quinn et al., 2006; Redline et al., 2004). School-based surveillance is an efficient way to capture information on asthma-related morbidity in communities with high pediatric asthma prevalence (Bruzzese et al., 2009). School-based surveys may be limited by the self-reported nature of the data and the frequent lack of objective measures of asthma case status (Davis et al., 2008). Further, some of the measures of asthma symptoms and severity vary temporally (Davis et al., 2008). Data (particularly symptom type and frequency) available through surveys can be coupled with claims data to provide a more comprehensive understanding of the landscape of asthma-related morbidity in a community.

We examine the asthma-related healthcare utilization patterns among adolescent students who are clients of a Medicaid managed care program and have completed a school-based asthma questionnaire. This linkage provides a comprehensive dataset with adequate community-level prevalence and severity information in this population of children with asthma. To identify risk factors (demographic and healthcare utilization) that predict asthma severity, we use recursive partitioning analysis to define key pediatric asthma severity subgroups within this population.

2. Methods

2.1. Study population

As part of the CDC-funded Controlling Asthma in American Cities Project, *Oakland Kicks Asthma*™ (OKA) (co-sponsored by the American Lung Association of California; the University of California, Berkeley; Children's Hospital Oakland; and the Oakland Unified School District (OUSD)) conducted school-based asthma surveillance for adolescent students enrolled in OUSD middle schools. From 2003 to 2008, asthma surveillance was conducted in all OUSD middle schools (n = 20) and three high schools at the start of each school year. Methodology and implementation of the asthma surveillance in the OUSD has previously been described (Magzamen et al., 2005). Briefly, a self-administered, 14-question survey based on the International Study of Asthma and Allergy in Childhood (ISAAC) questionnaire, a standardized asthma questionnaire used to describe the prevalence and severity of asthma (Asher et al., 1995), was provided to students during class. The modified ISAAC survey was designed to be short, easy to complete, and provide information not available from routine administrative health forms. Prior to administration of the case-identification survey, parents were sent a letter that described the project; parents were given the option to opt-out of the survey. At the time of survey administration, students were able to decline to participate. All activities conducted under OKA were approved by the OUSD, the Committee for the Protection of Human Subjects at the University of California, Berkeley, and the Institutional Review Board at Colorado State University.

Interpretation of the survey results was based primarily on the National Asthma Education and Prevention Program Expert Panel III Guidelines for the Diagnosis and Treatment of Asthma (NHLBI, 2007). Students who reported a physician diagnosis of asthma as well as a constellation of symptoms associated with asthma-related morbidity were classified as current asthma, and assigned into severity categories (high- and low-severity). The high-severity students reported either a broader range of symptoms, or higher symptom frequency compared to low-severity students; all students who reported an ED visit were classified as high-severity (algorithm available upon request from the authors). All students identified as current asthma were eligible for education and management interventions conducted by OKA. The study survey and classification algorithm are available upon request from authors.

2.2. Survey and medical claims data linkage

OUSD students identified as current asthma were matched to medical claims data provided by the Alameda Alliance for Health (AAH), the not-for-profit county Medicaid umbrella organization that manages and provides health care services for low-income families in Alameda County, CA (Fig. 1). The project established a memorandum of understanding with AAH for data sharing; students were matched by name, date of birth and current address, before de-identifying the data for analysis. Only AAH data for students who had a primary diagnosis code for asthma (ICD-9 code of 493.xx), resided in the city of Oakland, and who had complete records for primary language, race/ethnicity, and residence were used for data analysis.

Healthcare utilization measures were selected based on previously described methods (Brandt et al., 2010). Briefly, billing codes for asthma-relevant health encounters over the year prior to completion of the OKA survey were selected as potential covariates to explain survey-based classification group. Tables A1 and A2 show asthma-related billing codes collected for this analysis, as well as demographic indicators from the AAH database that were included as potential explanatory factors.

For final analysis we only included students with complete covariate set in their records, completed surveys and with no eligibility gaps in health coverage.

2.3. Statistical methods

Asthma severity profiles were created based on the characteristics of individuals in our linked dataset using the Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) tree algorithm (http://www.stat.wisc.edu/~loh/guide.html) (Loh, 2011; Loh, 2009), and asthma severity categories as the outcome. GUIDE is based on recursive partitioning (classification and regression trees), a group of non-parametric, exploratory techniques that capture variation of a single response variable by repeatedly splitting data into homogenous groups based on a set of explanatory variables (Breiman et al., 1984; Strobl et al., 2009; Lemon et al., 2003; De'ath and Fabricius, 2000). Recursive partitioning builds a classification rule to predict class membership (e.g., high- or low-severity) on the basis of its associated covariates (Zhang et al., 2001). Among other advantages over traditional parametric methods (Strobl et al., 2009; Lemon et al., 2003; Afonso et al., 2012; Kuchibhatla and Fillenbaum, 2002; Speybroeck et al., 2004; Kitsantas et al., 2006), recursive partitioning allows for flexibility regarding distributional assumptions and is well suited to data analyses with limited a priori knowledge of variable relationships (Pagán et al., 2009).

Methodology for the GUIDE algorithm is described in detail in Loh (2009). Briefly, GUIDE uses a two-step splitting approach to identify variables that best predict the outcome based on chi-square significance

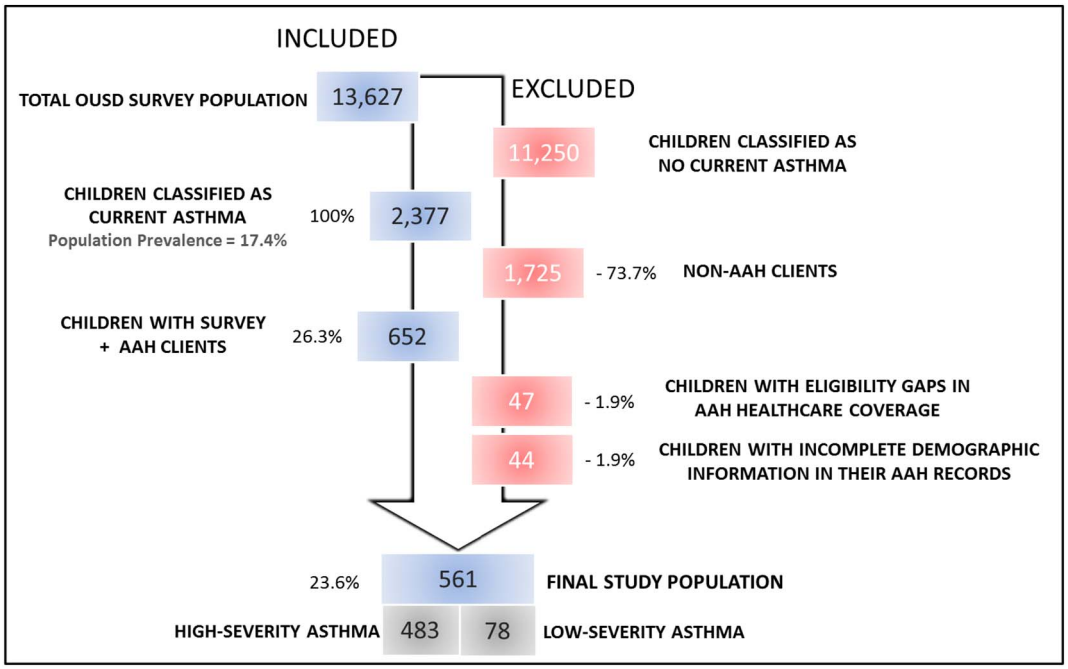

Fig. 1. Flow chart showing study participants selection from the Oakland Unified School District. The study connects the surveyed students with asthma to a health insurance claims database. Abbreviations: OUSD, Oakland Unified School District; AAH, Alameda Alliance for Health.

tests. Each independent variable was first tested for association with the outcome, significance probability values (similar to a p-value) were computed, and the variable with the lowest significant probability value reserved. Subsequently, a comprehensive search was performed on the set of all remaining significant variables and the reserved variable was selected to split the node at each level; this process ran until significant differences between groups were exhausted, or user-defined stopping rules were met (e.g., minimum number of observations in each group). The GUIDE algorithm was also able to identify local pairwise interactions among the predictor variables independent of a priori knowledge. The final tree was pruned (determining the most predictive tree model least vulnerable to the noise in the data) based on a 10-fold cross-validation criterion to minimize the unbiased estimate of misclassification cost. Misclassification cost is a measure of error that results from wrongly classifying an individual's outcome based on their exposure profile. The GUIDE algorithm uses this cross-validation technique to estimate the misclassification cost of each subtree it creates, and chooses an optimal tree with the lowest estimated cost (Loh, 2011; Yao et al., 2009). To reduce the effect of split-domination (i.e., large difference among the class observation numbers may cause GUIDE to predict the same class at all the splitting intervals) in our resulting trees, we apportioned equal priors to each class in our analysis. Therefore, class prior probabilities represent the probability of being in the high-severity group given that the prior probability that a student belongs to either of the two groups is equal. Additional model-specific pseudocode for analytic choices (including split type, classification model choice, pruning and cross-validation methods) is included in the Supplementary material (Fig. A1, Table A3).

3. Results

Over the five-year study period, 13,627 students completed the OUSD survey. A total of 2377 (17.4%) students were classified as current asthma, with 2033 (85.5%) of these students classified as high-severity asthma and 344 (14.5%) as low-severity asthma. We identified 652 (27.4%) OUSD students with asthma who completed the surveillance questionnaires and were AAH patients. Approximately 7% of these matched students (n = 47) were excluded from the study population due to eligibility gaps in healthcare coverage. In addition, we excluded 44 students who had incomplete demographic information in their AAH records resulting in 561 students with complete claims records (Fig. 1). This final study population included 86.1% high-severity asthma students (n = 483) and 13.9% low-severity asthma students (n = 78). The proportion of students classified as high and low severity in the matched data set did not differ significantly from the severity classifications of the study population (p = 0.7).

Demographic and health care utilization information for the study population are presented by outcome categories in Table 1. The majority of students reported ethnicity as non-Hispanic Black (63.5%), spoke English at home (72.9%), and were registered with the MediCal® health care group (90.7%). Compared to the low-severity asthma group, students classified as high-severity asthma were significantly more likely to be non-Hispanic Black (p = 0.008), more likely to be on controller medication (p = 0.015), but less likely to be on prescribed rescue medications (p = 0.002).

3.1. Summary of final GUIDE tree

A classification tree that predicts severity class generated by the GUIDE algorithm from the set of predictor/candidate variables is presented in Fig. 2. Tree interpretation is based on node pathways starting at the top of the tree (root node) and evaluating the prediction (classification) rules at each branch along the pathway until a terminal node is reached (Khan et al., 2015). The splitting variables at each level are shown labeled beside each node; upon satisfaction of the preset condition for the variable beside the labeled node, a defined occurrence goes to the left of the tree structure if the condition is met (e.g., in the topmost root node if the student is on prescribed rescue medication [Rescue Meds = Yes]), and to the right of the tree if unmet (student is *not* on prescribed rescue medication [Rescue Meds ≠ Yes]). Each terminal node represents predicted high or low-severity classes of asthma as indicated below the nodes; sample sizes for the severity groups are included beside each terminal node. Each low-severity asthma observation was treated as equivalent to 6.2 (483/78) high-severity asthma observations to prevent split-domination by the much larger high-severity asthma outcome class. The optimal split of our

Table 1
Study population demographic and healthcare utilization data by outcome categories.

Variable	Total	Low-severity	High-severity	χ^2 p-value
N (%)	561	78	483	
Sex				
Female	281 (50.1)	35 (44.9)	246 (50.9)	0.32
Age				
≤11 years	274 (48.8)	39 (50.0)	235 (48.7)	0.96
12 years	158 (28.2)	21 (26.9)	137 (28.3)	
≥13 years	129 (23.0)	18 (23.1)	111 (23.0)	
Race/ethnicity				
Non-Hispanic Black	356 (63.4)	39 (50.0)	317 (65.6)	**0.01**
Non-Hispanic White	89 (20.7)	13 (16.7)	76 (15.7)	
Hispanic	116 (15.9)	26 (33.3)	90 (18.6)	
Primary language group				
Asian	79 (14.1)	18 (23.0)	61 (12.6)	0.07
English	409 (72.9)	53 (68.0)	409 (73.7)	
Spanish	67 (11.9)	7 (9.0)	67 (12.4)	
Other	6 (1.1)	0 (0.0)	6 (1.3)	
Health insurance group				
FCP	3 (0.5)	0 (0.0)	3 (0.6)	0.08
HFP	48 (8.6)	7 (9.0)	41 (8.5)	
HKP	1 (0.2)	1 (1.3)	0 (0.0)	
MCAL	509 (90.7)	70 (89.7)	439 (90.9)	
Health insurance plan				
COMM	3 (0.5)	0 (0.0)	3 (0.6)	0.12
HFP	48 (8.6)	7 (9.0)	41 (8.5)	
HKP	1 (0.2)	1 (1.2)	0 (0.0)	
MCAL	467 (83.2)	66 (84.6)	401 (83.0)	
MCF	42 (7.5)	4 (5.1)	38 (7.9)	
Any allergy diagnosis				
Yes	266 (47.4)	32 (41.0)	234 (48.5)	0.22
Inpatient visit				
None	531 (94.6)	75 (96.2)	456 (94.4)	0.53
≥1	30 (5.4)	3 (3.8)	27 (5.6)	
ER visit				
None	416 (74.2)	64 (82.1)	352 (72.9)	**0.05**
1	76 (13.5)	11 (14.1)	65 (13.4)	
> 1	69 (12.3)	3 (3.8)	66 (13.7)	
Outpatient visit				
0	117 (20.8)	19 (24.4)	98 (20.3)	0.31
1	127 (22.6)	15 (19.2)	112 (23.2)	
2	88 (15.7)	15 (19.2)	73 (15.1)	
3	57 (10.2)	11 (14.1)	46 (9.5)	
> 3 times	172 (30.7)	18 (23.1)	154 (31.9)	
Outpatient treatment				
0	438 (78.1)	70 (89.7)	368 (76.2)	**0.02**
1	77 (13.7)	7 (9.0)	70 (14.5)	
> 1	46 (8.2)	1 (1.3)	45 (9.3)	
Visited a specialist				
None	544 (97.0)	73 (93.6)	471 (97.5)	0.06
≥1	17 (3.0)	5 (6.4)	12 (2.5)	
HEDIS defined controller meds				
0	358 (63.8)	61 (78.2)	297 (61.5)	**0.02**
1	104 (18.5)	10 (12.8)	94 (19.5)	
2	99 (17.7)	7 (9.0)	92 (19.0)	
HEDIS defined rescue meds				
Yes	249 (44.4)	47 (60.3)	202 (41.8)	< **0.01**
Prednisone prescription				
0	517 (92.2)	76 (97.4)	441 (91.3)	0.06
≥1	44 (7.8)	2 (2.6)	42 (8.7)	
Nebulizer treatment				
0	457 (81.5)	71 (91.0)	386 (79.9)	0.06
1	62 (11.0)	5 (6.4)	62 (11.8)	
> 1	2 (7.5)	2 (2.6)	42 (8.3)	
Influenza/pneumonia vaccine				
0	396 (70.59)	55 (70.51)	341 (70.60)	0.99
≥1	165 (29.41)	23 (29.49)	142 (29.40)	
Home asthma equipment[a]				
0	523 (93.2)	76 (97.4)	447 (92.5)	0.11
≥1	38 (6.8)	2 (2.6)	36 (7.5)	
Pulmonary function testing				
0	364 (64.9)	55 (70.5)	309 (64.0)	0.52
1	97 (17.3)	11 (14.1)	86 (17.8)	
2	44 (7.8)	7 (9.0)	37 (7.7)	
> 2	56 (10.0)	5 (6.4)	51 (10.5)	
Comorbidity				
Yes	162 (28.9)	22 (28.2)	140 (29.0)	0.89

Statistically significant ($p < 0.05$) results are bold.
Abbreviations. HEDIS: Healthcare Effectiveness Data and Information Set; FCP: Family Care Program; HFP: Healthy Families Program; HKP: Healthy Kids Program; MCAL: Medi-Cal Program; COMM: Community Health Program; MCF: Children's First Medical group.
[a] Home nebulizers for aerosolized bronchodilator administration was the primary home asthma equipment.

entire sample was on the variable "prescribed rescue medications" (Fig. 2). The group on the right (students with no prescribed rescue medication) contained a larger proportion of children with high-severity asthma (281/483 = 58.2%) than the one on the left (students with prescribed rescue medication; 41.8% high-severity asthma).

The eight predicted risk subsets for the asthma severity groups are presented in Table 2. For example, if a student *was* on prescribed rescue medications and *was not* Hispanic or White Non-Hispanic, the likelihood to be in the high-severity asthma group was 50.1% (Subset 4). Conversely, if a student *was* on prescribed rescue medications, was living in a home where the primary language group (spoken at home) *was* Spanish, and *was* Hispanic or White non-Hispanic, the likelihood to be in the low-severity asthma group was 56.3% (Subset 1). Terminal Subsets 4, 5 and 8 all predicted high-severity asthma, and terminal Subsets 1, 2, 3, 6 and 7 predicted low-severity asthma. The probability of being in the high-severity group was highest in Subset 5 (class probability: 92.1%), and lowest in Subset 2 (class probability: 19.0%).

Table 2 and the tree highlight the importance, in combination, of demographic risk variables (ethnicity and primary language spoken at home) and healthcare utilization variable (use of rescue medication, frequency of nebulizer use and outpatient visits) groupings in predicting asthma severity class. However, the presence of a particular risk variable in a subset was not consistently related to the classification of the outcome. The value of the covariate, in addition, was also an important determinant of subgroup segmentation. For example both subsets 4 and 5 predicted high-severity asthma, but while subset 4 consisted of non-Hispanic Black students on prescribed rescue medication, subset 5 consisted of students currently *not* on prescribed rescue medication, but with at least one time in-clinic nebulizer treatment in the past 6 months. Another element highlighted by the tree is the inconsistent relationship with severity and outpatient visits. Among students currently not on prescribed rescue medication with no in-clinic nebulizer treatment in the past six months and whose primary language spoken at home was Spanish, English or an Asian language, two-three outpatient visits in the past 6 months predicted low-severity asthma (Subset 7). However, having less than two, or more than three outpatient visits among a group with similar variable characteristics predicted high-severity asthma (Subset 8).

We show the overall predictive accuracy of our tree-based model using the resubstituted estimate of the mean misclassification cost in Table 3. The 10-fold cross-validated estimate of misclassification error in the final model was 0.33 (i.e., there was about a 33% chance of misclassifying a student's asthma severity group).

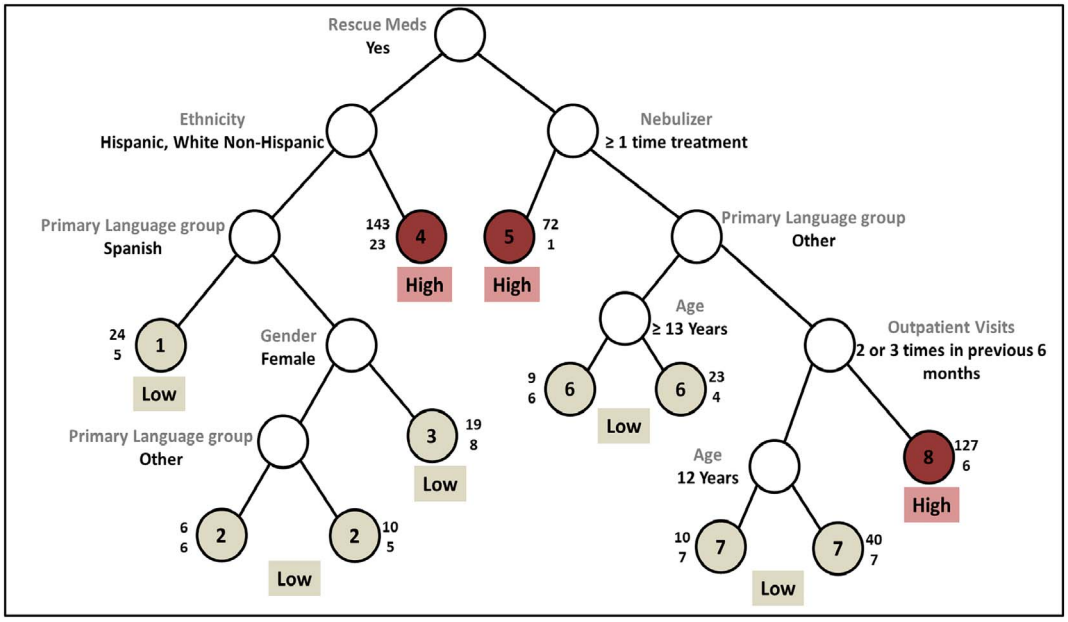

Fig. 2. Classification tree for predicting SEVERITY using univariate, kernel discrimination node models, equal priors and unit misclassification costs. At each split, an observation goes to the left branch if and only if the condition is satisfied. For splits on categorical variables, values not present in the training sample go to the right. Predicted classes (based on estimated misclassification cost) are printed below terminal nodes; sample sizes for SEVERITY = High and Low, respectively, beside nodes. To reduce the effect of split-domination, equal priors were assigned to each outcome class; each 'low-severity' asthma observation was treated as equivalent to 6.2 "high-severity" asthma observations.

4. Discussion

Community-level pediatric and adolescent asthma surveillance has become increasingly important given demographic, location and healthcare utilization differences in asthma risk factor distribution, prevalence and severity spectrum (Akinbami et al., 2009; Vital, 2011; Asher et al., 1995; Roberts et al., 2006; Davis et al., 2008). Given that one of the primary objective of community-based surveillance approaches is to identify asthma severity risk groups for adequate planning and management, it is key that these programs are based on comprehensive data (Reeves et al., 2006; Borowsky et al., 2013; Magzamen et al., 2014). Our study presents the use of data linkage and a predictive population segmentation technique to create asthma severity risk profiles based on community-level data.

Linkage of survey and administrative data appears to be an efficient way to obtain reliable health profile data (Hure et al., 2015; Gresham et al., 2015; Kreuter et al., 2010; Young et al., 2001; Sakshaug and Kreuter, 2012; Cullen et al., 2006). Our analysis draws on the strengths of two distinct yet complementary datasets frequently used in asthma epidemiology: a community-based survey and medical claims data. The AAH claims data provided important information on socioeconomic descriptors and healthcare utilization measures such as outpatient visits, in-clinic treatment and prescribed medications over the last calendar year. These claims data, however, lacked detailed information on symptom and severity variables such as presence and frequency of

Table 3
Classification Matrix for Establishing Accuracy of Model.

Predicted class	True class	
	High	Low
High	440 (91.1%)	45 (57.7%)
Low	43 (8.9%)	33 (42.3%)
Total	483 (100.0%)	78 (100.0%)

Resubstitution est. of mean misclassification cost = 0.33.

Table 2
Risk characteristics of classification tree terminal subsets.

Terminal subset	Risk characteristics of subjects in subset	Class prior srobabilities[a,b]	Predicted class
1	Rescue Meds = Yes; Ethnicity = Hispanic or White Non-Hispanic; Primary Language group (spoken at home) = Spanish	0.44	Low
2	Rescue Meds = Yes; Ethnicity = Not Hispanic or White Non-Hispanic; Primary Language group (spoken at home) = Not Spanish; Gender = Male; Primary Language group (spoken at home) = Other or English/Asian	Other (0.14) English/Asian (0.24)	Low
3	Rescue Meds = Yes; Ethnicity = Hispanic or White Non-Hispanic; Primary Language group (spoken at home) = Not Spanish; Gender = Male	0.28	Low
4	Rescue Meds = Yes; Ethnicity = Not Hispanic or White Non-Hispanic	0.50	High
5	Rescue Meds = No; Nebulizer = ≥ 1 time treatment	0.92	High
6	Rescue Meds = No; Nebulizer = No treatment; Primary Language group (spoken at home) = Other; Age = ≥ 13 years/ < 13 years old	≥ 13 years (0.20) < 13 years (0.48)	Low
7	Rescue Meds = No; Nebulizer = No treatment; Primary Language group (spoken at home) = Spanish, English or Asian; Outpatient Visits = 2 or 3 times in previous 6 months; Age = 12 years/ ≠ 12 years old	12 years (0.19) ≠ 12 years (0.48)	Low
8	Rescue Meds = No; Nebulizer = No treatment; Primary Language group (spoken at home) = Spanish, English or Asian; Outpatient Visits ≤ 2 times in previous 6 months *OR* > 3 times in previous 6 months	0.77	High

[a] To reduce the effect of split-domination, we apportioned equal priors to each class in our analysis with each "low-severity" asthma observation treated as equivalent to 6.2 'high-severity' asthma observations.

[b] The probability of being in high-severity group (a priori set cutpoint: 0.50).

symptoms like wheezing, coughing and nocturnal dyspnea. Our ISAAC-based asthma questionnaire obtained this supplementary information, which is essential for assessing symptoms, response to treatment, and ultimately severity categories.

Recursive partitioning or tree-based methods have been implemented in asthma surveillance and management studies (Gorelick et al., 2008; Barton et al., 2005; Lieu et al., 1998; Peters, 2006). In a case-control study to investigate the effects of psychosocial risk factors on asthma mortality, Barton et al. (2005) used the Chi-Square Automatic Interaction Detection (CHAID) recursive partitioning algorithm to show that family and psychosocial problems were associated with an increased risk of mortality for patients who did not have an asthma action plan (Barton et al., 2005). Using logistic regression and then recursive partitioning, Gorelick et al. (2008) found that clinical score and number of albuterol treatments given in the ED accurately distinguished between patients who could be discharged from the ED without subsequent relapse, and patients who had a need for further inpatient care (Gorelick et al., 2008).

Through application of recursive partitioning to our linked dataset, we identified mutually exclusive subgroups (for high and low asthma severity) of children that were relatively homogenous with respect to important risk variables. Classification trees produced by GUIDE and other tree-based algorithms tend to be intuitively more appealing than statistical or arithmetic scores, especially in settings where decisions are made based on practical or logical delineation of determinants (Marshall, 2001). We demonstrate that recursive partitioning methods can also play an important part in efficient analysis of available surveillance data because this method does not have restrictions by sample size, number of predictor variables or limited a priori knowledge of variable relationships, as highlighted in previous literature (Strobl et al., 2009; Lemon et al., 2003; Pagán et al., 2009). Compared to other recursive partitioning techniques (for example, CART, C4.5 and CHAID), the GUIDE algorithm used in our study has the added advantage of building more parsimonious and precise tree structures by reducing selection bias observed with other recursive partitioning methods (i.e., every predictor variable has an equal chance to be selected to be a split variable if the outcome is truly dependent on it) (Loh, 2011; Loh, 2009). GUIDE also offers more robust options for tree pruning (using cross-validation or test-sample pruning) and fitting nontrivial models (kernel and nearest-neighbor models) to the nodes (Loh, 2009; Khan et al., 2015).

Our results highlight some important groups, which based on our severity criteria, provide suitable guidance for monitoring and intervention in this asthma population. For example, Subset 5 describes a homogenous high-severity asthma risk (92.1% probability) subset of students who are not taking prescribed rescue medication but have needed in-clinic nebulizer treatment at least once in the past 6 months. This finding can inform asthma intervention programs among similar children in the population, as well as help improve subsequent asthma management and surveillance methods (e.g., streamlining survey questions to include nebulizer/rescue medication attitude and practices). Another notable feature among the high-severity risk subsets is the combined group characteristic of terminal Subset 8. This subset highlights the risk associated with children currently not on controller or maintenance medication, but who either had limited (< two visits) or many (> three visits) outpatient clinic visits for asthma. These features may be due to the influence of low socioeconomic status and components of poor asthma management, which can result in increased risk of adverse asthma-related outcomes (Akinbami et al., 2009; Zahran and Bailey, 2013; Akinbami et al., 2011). Other subsets provide additional information on risk profiles in this community. For example, Subset 1 describes a homogenous, low-severity asthma risk group and contains Hispanic and White Non-Hispanic male students currently on rescue medication, whose primary language group is not Spanish. This information may be useful in implementing a targeted management solution specific to groups at low risk for asthma complications.

Our recursive partitioning technique resulted in highly sensitive but non-specific predictions. In the context of making asthma management decisions, this means the model may serve as useful first-line form of testing when predicting severe cases of pediatric asthma for which the goal is to minimize the false negative rate at the expense of increasing the false positive rate.

There are several limitations to this study. First, non-coverage and eligibility issues reduced our effective sample size. Approximately 75% of the students who participated in the surveillance were not AAH members. An additional 47 students who were AAH members had eligibility gaps and were not included in the analysis. It is difficult to attribute representativeness to our limited study population regardless of features of our recursive partitioning algorithm. Second, the use of billing data may lead to a potential misclassification of the outcome usually as a result of inconsistent definitions and classifications of asthma (Dombkowski et al., 2005). With regard to the previously mentioned coverage and eligibility gaps for several students in this analysis, absence of healthcare utilization records may indicate changes in income, geography, or family structure (with families moving in and out of healthcare organizations), and not necessarily the absence of an encounter for a given period. Further, directly linking prescription records to actual frequency or correct use of the medication is not without flaws. By linking such administrative data with surveillance data, we attempted to reduce this bias. Third, the self-reported nature of our survey data may have also resulted in outcome misclassification via either recall bias or misreporting, although linkage with administrative data likely minimized this information bias. Lacking pulmonary function testing, our data did not allow us to distinguish between children with clinically severe asthma and students with poorly controlled or mismanaged asthma from the survey, we believe that the “high-severity” classification potentially reflects both sets of children. Finally, one of the main criticisms levied against tree-based algorithms is their inability to produce individual exposure-effect estimates (Lemon et al., 2003; Kuchibhatla and Fillenbaum, 2002; Kitsantas et al., 2006). We built our trees to predict and highlight characteristics of asthma severity groups for more focused asthma management approaches, and not tailored for estimating direct effects of risk factors on asthma severity. Our study population was diverse but not necessarily representative of the general US pediatric asthma population; healthcare utilization patterns of patients (in particular) may differ from other populations. We have not had the opportunity to validate our algorithm with additional data; future studies should aim to validate our linkage and data segmentation methodology in other population settings.

We illustrate the complementary relation between community-level asthma survey data and medical claims data, and highlight potential approaches for community-based asthma surveillance to leverage data linkage. It is important to consider the creation of a varied portfolio of management plans conceivably based on predictive population segmentation techniques. The relative ease (access and implementation) of our technical approach for population segmentation results in beneficial predictive design that can provide a template for clinicians and local public health officials to optimize asthma patient management plans.

Primary source of funding

Dr. Benka-Coker and Dr. Magzamen were supported by a grant from the National Institute for Environmental Health Sciences (K22ES023815, PI: Magzamen). Data collection for this project was supported by CDC Cooperative Agreements U59/CCU923264-01 (PI: Tager) and CDC 50/CCU922409-02 (PI: Balmes).

Conflict of interest

At the time this paper was written and submitted, there were no conflicts of interest for all named authors.

Acknowledgments

The authors are grateful to Dr. Ira Tager, Elizabeth Edwards of Alameda Alliance for Health, and the students and families of the Oakland Unified School District for their contributions to this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2018.02.004.

References

Afonso, A.M., Ebell, M.H., Gonzales, R., et al., 2012. The use of classification and regression trees to predict the likelihood of seasonal influenza. Fam. Pract. 29, 671–677.
Akinbami, L.J., Moorman, J.E., Garbe, P.L., et al., 2009. Status of childhood asthma in the United States, 1980–2007. Pediatrics 123, S131–45.
Akinbami, L.J., Moorman, J.E., Liu, X., 2011. Asthma prevalence, health care use, and mortality: United States, 2005–2009. Natl. Health Stat. Rep. 1–14.
Akinbami, L.J., Simon, A.E., Rossen, L.M., 2016. Changing trends in asthma prevalence among children. Pediatrics 137, 1–7.
Asher, M.I., Keil, U., Anderson, H.R., et al., 1995. International study of asthma and allergies in childhood (ISAAC): rationale and methods. Eur. Respir. J. 8, 483–491.
Barnett, S.B.L., Nurmagambetov, T.A., 2011. Costs of asthma in the United States: 2002–2007. J. Allergy Clin. Immunol. 127, 145–152.
Barton, C. a, McKenzie, D.P., Walters, E.H., et al., 2005. Interactions between psychosocial problems and management of asthma: who is at risk of dying? J. Asthma 42, 249–256.
Borowsky, B., Little, A., Cataletto, M., 2013. Determining the relative burden of childhood asthma at the local level by surveying school nurses. Pediatr. Allergy Immunol. Pulmonol. 26, 76–80.
Brandt, S., Gale, S., Tager, I.B., 2010. Estimated effect of asthma case management using propensity score methods. Am. J. Manag. Care 16, 257–264.
Breiman, L., Friedman, J.H., Olshen, R.A., et al., 1984. Classification and regression. Trees 19.
Bruzzese, J.-M., Evans, D., Kattan, M., 2009. School-based asthma programs. J. Allergy Clin. Immunol. 124, 195–200.
Busi, L.E., Sly, P.D., Restuccia, S., et al., 2012. Validation of a school-based written questionnaire for asthma case identification in Argentina. Pediatr. Pulmonol. 47, 1–7.
Cullen, M.R., Vegso, S., Cantley, L., et al., 2006. Use of medical insurance claims data for occupational health research. J. Occup. Environ. Med. 48, 1054–1061.
Davis, A., Savage Brown, A., Edelstein, J., et al., 2008. Identification and education of adolescents with asthma in an urban school district: results from a large-scale asthma intervention. J. Urban Health 85, 361–374.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.
Dombkowski, K.J., Wasilevich, E.A., Lyon-Callo, S.K., 2005. Pediatric asthma surveillance using Medicaid claims. Public Health Rep. 120, 515–524.
Dombkowski, K.J., Lamarand, K., Dong, S., et al., 2012. Using Medicaid claims to identify children with asthma. J. Public Health Manag. Pract. 18, 196–203.
Gorelick, M., Scribano, P.V., Stevens, M.W., et al., 2008. Predicting need for hospitalization in acute pediatric asthma. Pediatr. Emerg. Care 24, 735–744.
Gresham, E., Forder, P., Chojenta, C.L., et al., 2015. Agreement between self-reported perinatal outcomes and administrative data in New South Wales, Australia. BMC Pregnancy Childbirth 15, 161.
Gupta, R.S., Carrión-Carire, V., Weiss, K.B., 2006. The widening black/white gap in asthma hospitalizations and mortality. J. Allergy Clin. Immunol. 117, 351–358.
Hure, A.J., Chojenta, C.L., Powers, J.R., et al., 2015. Validity and reliability of stillbirth data using linked self-reported and administrative datasets. J. Epidemiol. 25, 30–37.
Keet, C.A., McCormack, M.C., Pollack, C.E., et al., 2015. Neighborhood poverty, urban residence, race/ethnicity, and asthma: rethinking the inner-city asthma epidemic. J. Allergy Clin. Immunol. 135, 655–662.
Khan, G., Bill, A.R., Noyce, D.A., 2015. Exploring the feasibility of classification trees versus ordinal discrete choice models for analyzing crash severity. Transp. Res. Part C Emerg. Technol. 50, 86–96.
Kitsantas, P., Hollander, M., Li, L., 2006. Using classification trees to assess low birth weight outcomes. Artif. Intell. Med. 38, 275–289.
Kreuter, F., Muller, G., Trappmann, M., 2010. Nonresponse and measurement error in employment research: making use of administrative data. Public Opin. Q. 74, 880–906.
Kuchibhatla, M., Fillenbaum, G.G., 2002. Assessing risk factors for mortality in elderly White and African American people: implications of alternative analyses. Gerontologist 42, 826–834.
Labrèche, F., Kosatsky, T., Przybysz, R., 2008. Childhood asthma surveillance using administrative data: consistency between medical billing and hospital discharge diagnoses. Can. Respir. J. 15, 188–192.
Lemon, S.C., Roy, J., M a, Clark, et al., 2003. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann. Behav. Med. 26, 172–181.
Lieu, T.A., Quesenberry, C.P., Sorel, M.E., et al., 1998. Computer-based models to identify high-risk children with asthma. Am. J. Respir. Crit. Care Med. 157, 1173–1180.
Loh, W.-Y., 2009. Improving the precision of classification trees. Ann. Appl. Stat. 3, 1710–1737.
Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1, 14–23.
Magzamen, S., Mortimer, K.M., Davis, A., et al., 2005. School-based asthma surveillance: a comparison of student and parental report. Pediatr. Allergy Immunol. 16, 669–678.
Magzamen, S., Brandt, S.J., Tager, I.B., 2014. Examining household asthma management behavior through a microeconomic framework. Health Educ. Behav. 41, 651–662.
Marshall, R.J., 2001. The use of classification and regression trees in clinical epidemiology. J. Clin. Epidemiol. 54, 603–609.
Mitchell, S.J., Bilderback, A.L., Okelo, S.O., 2016. Racial disparities in asthma morbidity among pediatric patients seeking asthma specialist care. Acad. Pediatr. 16, 64–67.
MMWR, 2011. Vital signs: asthma prevalence, disease characteristics, and self-management education: United States, 2001–2009. Morb. Mortal. Wkly Rep. 60, 547–552.
Morris, R.D., Naumova, E.N., Goldring, J., Hersch, M., Munasinghe, R.L., Anderson, H., 1997. Childhood asthma surveillance using computerized billing records: a pilot study. Public Health Rep. 112 (6), 506–512.
NHLBI, 2007. Expert panel report 3: guidelines for the diagnosis and management of asthma full report 2007. Child. Aust. 120, S94–138.
Pagán, J.A., Pratt, W.R., Sun, J., 2009. Which physicians have access to electronic prescribing and which ones end up using it? Health Policy (New York) 89, 288–294.
Peters, D., 2006. Using an asthma control questionnaire and administrative data to predict health-care utilization. Chest J. 129, 918.
Piecoro, L.T., Potoski, M., Talbert, J.C., et al., 2001. Asthma prevalence, cost, and adherence with expert guidelines on the utilization of health care services and costs in a state Medicaid population. Health Serv. Res. 36, 357–371.
Quinn, K., Shalowitz, M.U., Berry, C.A., et al., 2006. Racial and ethnic disparities in diagnosed and possible undiagnosed asthma among public-school children in Chicago. Am. J. Public Health 96, 1599–1603.
Redline, S., Gruchalla, R.S., Wolf, R.L., et al., 2004. Development and validation of school-based asthma and allergy screening questionnaires in a 4-city study. Ann Allergy Asthma Immunol 93, 36–48.
Reeves, M.J., Lyon-Callo, S., Brown, M.D., et al., 2006. Using billing data to describe patterns in asthma-related emergency department visits in children. Pediatrics 117, S106–17.
Roberts, E.M., English, P.B., Van den Eeden, S.K., et al., 2006. Progress in pediatric asthma surveillance I: the application of health care use data in Alameda County, California. Prev. Chronic Dis. 3, A91.
Sakshaug, J.W., Kreuter, F., 2012. Assessing the magnitude of non-consent biases in linked survey and administrative data. Surv. Res. Methods 6, 113–122.
Smith, M., Rascati, K., Barner, J., 2005. A descriptive analysis of asthma-related medical services and prescription utilization among recipients in a Medicaid program. J. Asthma 42, 447–453.
Speybroeck, N., Berkvens, D., Mfoukou-Ntsakala, A., et al., 2004. Classification trees versus multinomial models in the analysis of urban farming systems in Central Africa. Agric. Syst. 80, 133–149.
Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348.
Walsh-Kelly, C.M., Kelly, K.J., Drendel, A.L., et al., 2008. Emergency department revisits for pediatric acute asthma exacerbations - association of factors identified in an emergency department asthma tracking system. Pediatr. Emerg. Care 24, 505–510.
Yao, L., Zhong, W., Zhang, Z., et al., 2009. Classification tree for detection of single-nucleotide polymorphism (SNP)-by-SNP interactions related to heart disease: Framingham Heart Study. BMC Proc. 3, S83.
Young, A.F., Dobson, A.J., Byles, J.E., 2001. Health services research using linked records: who consents and what is the gain? Aust. N. Z. J. Public Health 25, 417–420.
Zahran, H.S., Bailey, C., 2013. Factors associated with asthma prevalence among racial and ethnic groups—United States, 2009–2010 behavioral risk factor surveillance system. J. Asthma 50, 583–589.
Zhang, H., Yu, C.-Y., Singer, B., et al., 2001. Recursive partitioning for tumor classification with gene expression microarray data. Proc. Natl. Acad. Sci. 98, 6730–6735.