The Report Committee for Courtney Marie Chancellor
Certifies that this is the approved version of the following report:

Predicting Emergency Department events due to Asthma: Results from the BRFSS Asthma Call Back Survey 2006-2009

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Lauren A. Meyers

James Scott

Predicting Emergency Department events due to Asthma: Results from the BRFSS Asthma Call Back Survey 2006-2009

by

Courtney Marie Chancellor B.S. B.A.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Computational Science, Engineering and Mathematics

The University of Texas at Austin

May 2012

Abstract

Predicting Emergency Department events due to Asthma: Results from the BRFSS Asthma Call Back Survey 2006-2009

Courtney Marie Chancellor MSCSEM
The University of Texas at Austin, 2012

Supervisor: Lauren A. Meyers

The identification of asthma patients most at risk of experiencing an emergency department event is an important step toward lessening public health burdens in the United States. In this report, the CDC BRFSS Asthma Call Back Survey Data from 2006 to 2009 is explored for potential factors for a predictive model. A metric for classifying the control level of asthma patients is constructed and applied. The data is then used to construct a predictive model for ED events with the rpart algorithm.

Table of Contents

1 Introduction

Asthma is a chronic respiratory disease characterized by reversable impairment of breathing due to inflammation and narrowing of the airways. In the United States, roughly 24.6 million people (17.5 million adults and 7.1 million children) are affected by asthma[1]. Severity ranges from mild, occational symptoms to severe, persistant symptoms; yet, an individual with mild asthma can experience severe episodes with serious health consequences. In 2009, approximately 1.75 million emergency department visits and over 3,000 deaths occured in the United States due to asthma. Although what causes and contributes to the development of this disease is not well understood, extensive work has been done to explore the irritants and comorbidities of those diagnosed. Attack triggers such as tobacco smoke, allergens, stress and excersize and self-management strategies are regularly taught to asthma patients: it is thought that, with proper self-managment and treatment compliance, the majority of adverse health outcomes such as hospitalization can be dramatically reduced[8].

Health care providers have made greater efforts to emphasize proper self-management as an important component of asthma control, in conjunction with medical treatment and care. In 1999, the CDC initialized the National Asthma Control Program (NACP)to this end. The NACP funds states, citites schools and non-governmenal programs to improve asthma management through evidence-based practices and survaillance. It would be a great benefit to personal physicians to know which patients are most at risk so that they may provide more appropriate care. More so, it would benefit larger agents, from hospitals to state governments, to predict likely asthma incidents based on the demographics of a particular population. Currently, there exists a large body of research dedicated to predicting overnight hospitalization given inpatient data collected within the ER. At this point, however, an asthma event has already occurred and it is too late for low cost intervention. Therefore, we seek a more general model which utilizes basic, easily obtained information such as demographic and self-reported behaviors.

2 Definition and Analysis of the Data

2.1 The BRFSS Asthma Call Back Survey

The BRFSS, or Behavioral Risk Factor Surveillance System, is an ongoing, random-digit-dialed telephone survey of adults (aged 18 and older) in the United States and its territories. With a variety of questions on behaviors, demographics and disease states, the yearly BRFSS data is publicly available up to the year 1984. In order to gather prevalence data on children, an adult may answer as a proxy for a single, random child in the household. As of 2005, the ACBS, or Asthma Call Back Survey, has been conducted as a follow up to the BRFSS; if an affirmative response was given to "Have you ever been told by a doctor, nurse or other health professional that you have asthma?", the patient was asked to participate in a call back. The ACBS contains detailed questions pertaining to asthma symptoms, self-management and healthcare. Such a dataset fits the desired spirit of our modeling in that there are a variety of questions specific to asthma and self-managment stragegies, yet all data is self-reported and non clinical by nature.

Territories and states not contained in the mainland of the United States (Hawaii, Alaska, Puerto Rico, Guam and the Virgin Islands) were not included in the analysis. Only patients who reported having current asthma symptoms were considered. Since it is beleived that asthma trends differ substantially in adults and children, only adults (≥ 18) were considered for this analysis. The BRFSS contains several hundred questions; we must somehow narrow the scope of our analysis to a subset of attributes. Variables were selected by hand under the guidance of existing literature and biological plausibility. Care was taken to include known confounders and comorbidities such as obesity, depression and COPD[9]. In addition, some questions were combined to create calculated variables such as exposure to environmental triggers, interventive methods against dust mite eposure and degree of self-management education. For a full list of the chosen variables and their explanations, see the Appendix.

2.2 Level of Asthma Control

In our model, we wish to incorporate some measure of seriousness in a patient's asthma. In literature, a clear distinction is made between the concepts of asthma severity and asthma control. It is possible for a patient to have severe asthma– that is, to be at risk of extreme airway restriction– yet have well controlled symptoms in which the patient has little risk of serious health consequences due to excellent preventative measures or management stragegies. The National Lung, Heart and Blood institute, through the National Asthma Education and Prevention Program, has released guidelines for categorizing degrees of control in asthma patients[5]. By adapting these guidelines to the available data set, we were thereby able to construct a tool for patient categorization to be incorporated into the model[4]. For a list of BRFSS variables used, see Appendix A. Characteristics used to categorize asthma patients as "Well Controlled", "Not Well Controlled" or "Poorly Controlled" are outlined in Table 2. Points were awarded to each patient based on responces to the selected variables and patients then classified by total 'control score'.

	Well Controlled	Not Well Controlled	Poorly Controlled
Symptoms	$\leq$ 8 per month	$>$ 8 per month	$\geq$ 20 per month or throughtout day
Nighttime Symptoms	$\leq$ 2 per month	$\geq$ 2 per month	$\geq$ 15 per month
Short-acting Beta Antagonist Use	$<$ 2 per week	$\geq$ 2per week	multiple times per day
Interference with Normal Activity	Not at all	A little to A Moderate Amount	A lot
Corticosteroid Use	None	Yes	Yes
Asthma attacks	No episode within 3 months	1-2 episodes in past 3 months	3 or more episodes in past 3 months

Table 1: Characteristics of Asthma Control Levels, as outlined by the EPR3 Panal and avaliable in the BRFSS data. Not all characteristics descriped in the report could be used due to the limits of the avaliable data set. Of the factors used, units of the EPR3 guidelines were scaled to the units of the BRFSS data, such as reporting symptoms within a month rather than a week.

2.3 Exploring Relationships of Predicitve Factors with Asthma Control

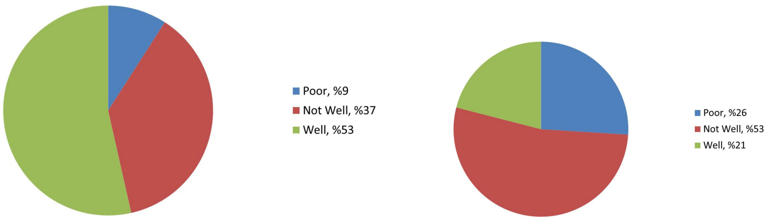

Figure 1: Composition of the Population by Control Levels: On the right is the decomposition of the entire population of records from 2006 to 2009, clearly demonstrating that the majority of patients with asthma can be classified as "Well Controlled". On the left is the decomposition of those who reported having an ED event within the last year. It is clear that both "Well Controlled" and "Poorly Controlled" patients are under and over represented in this category.

Once we have categorized patients by their respective levels of control, we can begin to explore potential correlations and relationships within the data. Some of this work is investigative in that we wish to discover potential variables for our predictive model, but primarily serves as validation: if our method of control categorization is valid, those relationships which have been thoroughly established in literature as being significant should be present. Of the BRFSS ACBS data from the years 2006 to 2009, the majority of patients were classified as either 'Well Controlled' or Not 'Well Controlled' (See Figure 1). However, of those two experienced an ED episode, the vast majority were of the categories 'Not Well Controlled' and 'Poorly Controlled'. There is a known relationship between asthma severity and demographics[8][1]. Low socioeconomic status, for example, can limit either the ability or the tendency to access to health care. We hope to see these same traits reflected in the distributions of asthma control levels. In fact, we do find the desired trends. As an illustration, in Figure 2, determinants of socioeconomic status such as education, income and smoking status are examined. As level of income increased, so did the proportion of well controlled to poorly controlled patients. Education level did not exhibit this same linear behavior: those with High School degrees were not significantly different than those who had attended but not completed college or technical schooling. Large differences were found between those with and those without college degrees.

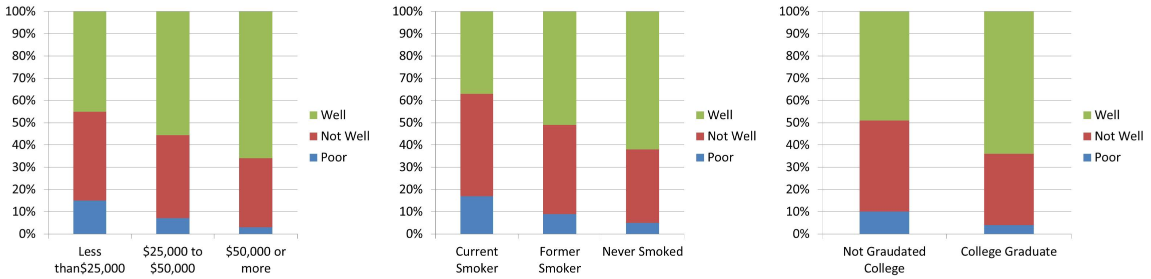

Figure 2: Socioeconomic Status and Asthma Control Levels: levels of socioeconomic status indicators such as income and education were compared within the full data set. Increasing income resulted in porportional changes in the composition of asthma control levels, as did smoking status. Education did not have this same linear effect, but rather, seemed to jump once a patient reported completion of a college or technical school.

In further investigation of socioeconomic factors and health care utilization, we wished to see if there was a relationship between financial hardship and a poor level of asthma control. This measure is not necessarily captured through reported income; we want to know whether access to health care specifically was impacted by financial circumstances. As seen in Figure 3, those with poorly controlled asthma were four times as likely to have reported not being able to pay for medication or physician visit in the past year than those with well controlled asthma. They were also nearly two times as likely to report having insurance problems within the past year. This is, of course, a more complicated issue when elligability for government programs are taken into accont.

Preventative measures can range from medication compliance, consistant contact with a primary physican, efforts to control environmental triggers, or degree of self-management education. It is possible that those with severe asthma would participate more in preventative care than those with mild or moderate asthma. However, patients who do take preventative measures should exhibit better control levels. Indeed, this seems to be the case as illustrated in Figure 4. For example, those patients who reported having a health care provider that they considered their primary physician were more than five times more likely to have well controlled asthma than poorly controlled asthma. In addition, those who reported making efforts to control environmental factors such as dust mites, including the use of mattress covers, pillow covers and washing sheets in hot water, were six times as likely to have well controlled asthma than poorly controlled asthma.

At this point, we wish to discover more about the population which does experience an ED

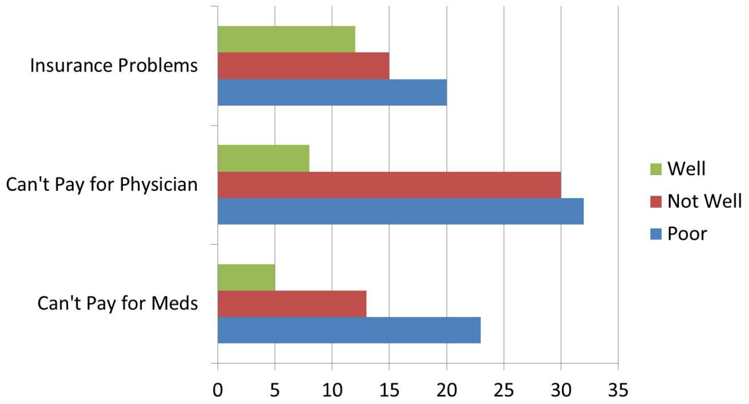

Figure 3: Indicators of Cost as a Barrier to Heath Care: Each bar indicates the percent of patients who responded afirmatively to “Has there been a time when you wanted to see a doctor for your asthma but could not due to cost?”, “Has there been a time when you did not have access to medication for your asthma due to cost?”, and “Has there been a time within the last year that you did not have health insurance?” within the respective asthma categories.

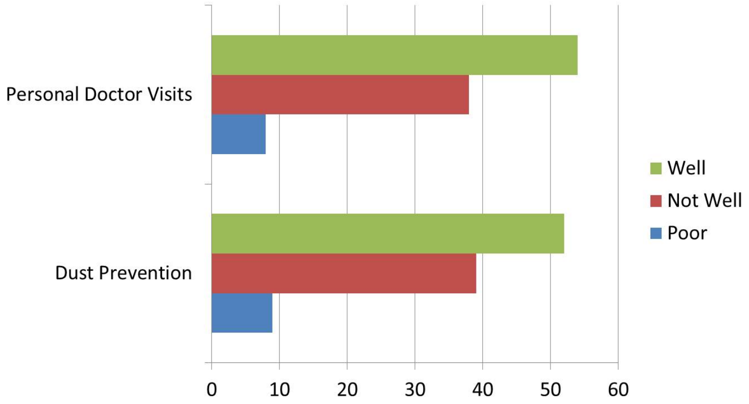

Figure 4: Preventative Measures: A patient is much more likely to have “Well Controlled” asthma if they have taken precentative measures. For example, those who took precaution against dust by using specialized mattress and pillow covers, washing sheets regularly in hot water or using an aircleaner were five times more likely to have well controlled asthma than poorly controlled asthma. Likewise, if a patient reported that they had an individual whom they thought of as their personal physician or health care provider.

episode, the target of our predictive model. Those factors which characterize the at-risk population will certainly be important in constructing the best possible model. However, we also wish to know more about the control levels of those who tend to be hospitalized. For example, women carry greater prevalence burdens than men, and tend to have more severe asthma. Of those who participated in the ACBS, men and women had very similar distributions of asthma control which mirrored the population distributions. Women made up the majoirty of cases in the ACBS data, a characteristic which may be attributed to the tendency of women to be more likely to respond to survey questionaires. However, of the women, 14% reported having an ED event verses only 9% of men. Of those who did report having an ED event, the majority were women, had lower levels of income and lower educational levels as represented in Figure 5.

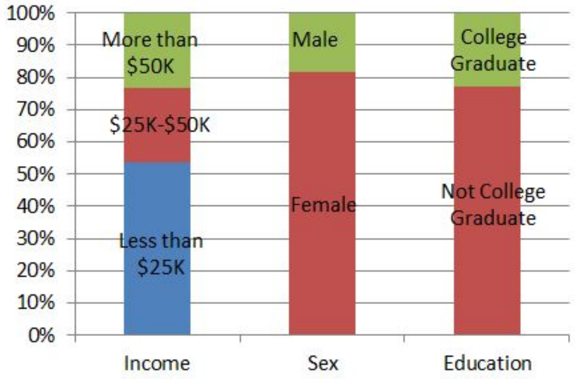

Figure 5: Demographics of patients who experienced an ED episode. Sex, education and income remained relevant within possible confounders such as race or metropolitan code. Here, the composition of these factors is displayed for those who reported ED events such as hospitalization or ER visits.

3 Developing a Predictive Model for ED Events

3.1 The RPART Algorithm

With a better understanding of the data and the relationships between variables, we are ready to construct a predictive model through recursive partitioning and regression trees (or 'rpart') as implemented by R. See Appendix B for a complete list of the definitions of variables used within this paper. A two-stage procedure, rpart first builds a tree by recursively determining what variable best splits the data until the nodes, A, reach a minimum size or until an additional split results in no improvement. The full tree is then searched for some optimal sub-tree which minimizes risk through cross-validation. In rpart, we define the splitting criterion through first choosing a measure of impurity,

$$I(A) = \sum_{i=1}^{C} \phi(p_{iA})$$

where p_{iA} is the proportion of cases who belong to class i in node A and ϕ is a function such that $\phi \geq 0$, for any $p\epsilon(0,1)$, $\phi(p) = \phi(1-p)$ and $\phi(0) = \phi(1) < \phi(p)$. The most commonly used impurity functions are the Gini index $\phi(p) = p(1-p)$ and the Information Gain $\phi(p) = -p\log(p)$. These measures are very similar and often select the same splits, making the appropriate choice of ϕ unclear.[7] However, some literature[6] suggested that the Gini index may have considerable faults in its incorporation of loss; therefore, the information gain was used instead.

Arguably one of the greatest strengths of the rpart algorithm is its method for handling missing data. While no cases whose target data is missing are used, we do not want to completely discard cases with only partially complete survey data. If case is missing the data on which the node is split, we wish to allow it to nevertheless progress down the tree. In such instances, we use what is called a surrogate split. A surrogate is a variable split on the data for which the most cases would be classified in the same way as the original split. No surrogate which does worse than "go with the majority" is used. If a case is missing data for the first best surrogate, the second best is used, and so forth.

Once the tree has been fully grown, we search all of the contained sub trees, considering the effect of changing internal nodes to terminal node. Each terminal node must be assigned to a class; this choice is based on the risk of a node, $R(A)$ Here, as with many applications, we wish to distinguish different kinds of errors, for classification, false positives and false negatives. Since intervention measures in the case of asthma are cheap and have little to no negative impact on the patient, it is far more costly to overlook a patient who does visit the ER than to intervene with patients who were not at risk. Therefore a false negative is more costly than a false positive, which we quantify through the loss function, L. The risk of a node is defined as

$$R(A) = \sum_{i=1}^{C} p_{iA} L(i, c(A))$$

The cost of the tree as a whole is, naturally a summation of the risks of its terminal nodes, weighted by the probability of the node.

3.2 Choosing a Fitted Tree

Without any further considerations, we would always select the largest possible tree since a fully saturated tree with n nodes would always classify the cases correctly. However, this is not very informative as a model. Instead, we penalize large trees by introducing a new parameter α, the cost of adding another variable to the model. This is known as the complexity parameter. The cost of the tree is defined as

$$R_\alpha(T) = R(T) + \alpha |T| R(T_0)$$

with $|T|$ being the size of the tree and T_0 being the single node tree(no splits). It is proven that for any value of α, there exists a unique, smallest subtree that minimizes the tree cost complexity.[2] However, that does not mean that each value of α results in a unique subtree. In fact, we expect to find intervals of α for which a single subtree is favored. When the rpart algorithm is implemented, subtrees for various intervals of α are presented, each optimal by cross validation.

All that is left is to determine the best choice of the complexity parameter. We can do this by cross-validation, however, we may also use the 1-SE rule and avoid this complicated implementation. According to Breiman et al[2], any risk within one standard error (deviation) of the minimum is equivalent to the minimum. Therefore, we take the smallest tree whose error is lower than the sum of the smallest cross validation and corresponding standard error. Pruning to this sub-tree, we have reached the final model and can quantify its error with the testing data set.

The underlying structure of a tree constructed with rpart is determined by the distributions of factors in the training data: splits recursively determined on entropy will not always capture what would otherwise be considered statistically significant characteristics. However, the pruned subtree is determined via risk calculation of the terminal nodes, a measure which takes into account weighted penalties for false negatives and false positives as directed by the loss matrix. Before we can examine any one particular tree, we must fix the off-diagonal entries of the loss matrix, $L(i, j)$. To this end, we must decide what the measure of goodness should be in comparing the resulting errors, specificities and sensitivities of potential trees. To use any one of these features alone push the tree selection toward a single node in which all individuals are classified as experiencing no ED episode (100% sensitivity) or experiencing an ED episode (100% specificity). Since these factors are in some sense balanced in the receiver operating characteristic curve, or ROC, the area under the ROC curve, or AUC, was used. Exploring the space of appropriate entry combinations resulted in maximal AUCs of 0.75 at equivalent false positive to false negative penalty ratios of 4.6:1.

3.3 The Optimal Tree

With the aforementioned inputs, the resulting tree is incredibly simple and intuitive(Figure 6).

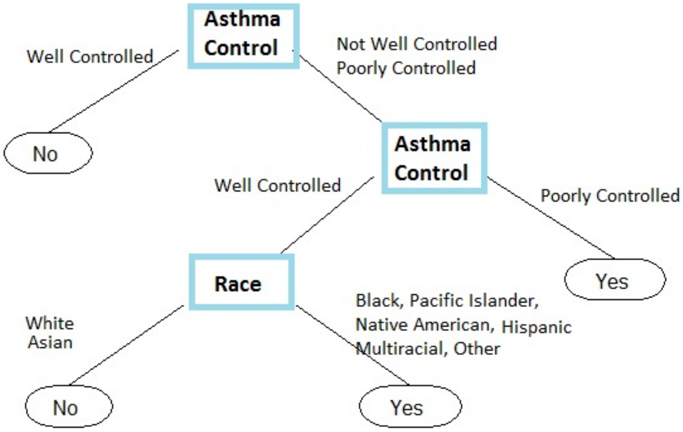

Figure 6: Optimal Tree with FP:FN Ratio 4.6:1: To follow the tree, we begin at the top node at which risk is equal to poulation risk. With each split we are guaranteed better separation of the target data. Patients follow the tree based on their individual data until they arrive at a terminal node where they take on the class of the node. For the optimal tree, the data is split on control level and race.

complexity parameter	0.01	AUC	0.75
error residual	0.260	F score	0.35
specificity	0.761	sensitivity	0.543

Table 2: Static Performace Measures of the Optimal Tree: Here we have listed the actual (that is, for a single confusion matrix) values most quoted in literature.

Patients are first broken into the the categories of Well Controlled, Not Well Controlled, and Poorly Controlled. Note that all splits in the rpart algorithm are 2-way but do not necessarily bifurcate into have and have-nots. For numerical data a best cutoff point is found, and for categorical data all possible combinations are considered. Therefore, for the optimal tree, we see that the greatest change in entropy is made in dividing "Well" from the "Not Well" and "Poor" control levels. The very next split is also a division based on control levels. For those with poorly controlled asthma are further divided by race, with Asian and White patients being passed to the non-ED episode category. Our initial examination of the chosen variables and asthma control suggested that many exhibited strong correlations. Since the rpart algorithm favors small trees, it is not surprising that asthma control might serve as the best proxy for other underlying factors. By examining the full summary of the tree object, we do, in fact, get notions of these effects. As well as knowing the actual split chosen by the algorithm, we also get to know the top five splits which were not chosen and their relative performance. Surprisingly, we do not see income, education or indicators of financial harship. The differences in these factors have already been accounted for by dividing on control levels.

Now that we have constructed the model, we woud like a measure of how well it performs. There is no simple answer to this question. After all, this is entirely dependent on the goals of the model. Since intervention in this circumstance is assumed to be low-cost and puts the patient at minimal or no additional risk, we value sensitivity and the minimization of false negatives. However, we do not want to compeltely compromise overall error or specificity. There must be a compromise, which is a subjective choice. The most common measures of performance were included in this analysis, including static measures (Table 2) and demonstrations of variability (Figure 7).

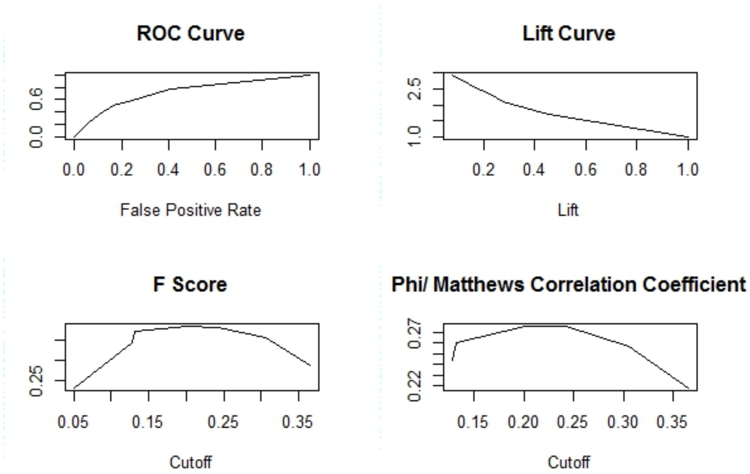

Figure 7: Dynamic Performance Measures of the Optimal Tree: One can imagine that an inverse relationship exists for such measures such as sensitivity and specificity. To visualize the actual trade off for a single model, we look at the above measures. The most commonly noted and most informative are the ROC and F Score, which demonstrate the relationship between sensitivity and specificity, and precision and recall, respectively.

3.4 Comparison to the Basic Logistic Regression

To be able to speak objectively about our methods results, we must have something to compare them to. Logistic regression is favored in the literature for the construction of predictive models in the public health sphere. The output of a logistic model is inherently tied to the concept that health outcomes are probibalistic rather than deterministic. In addition, regression allows for an analysis of the relative importances of each variable to the assigned probibalities.

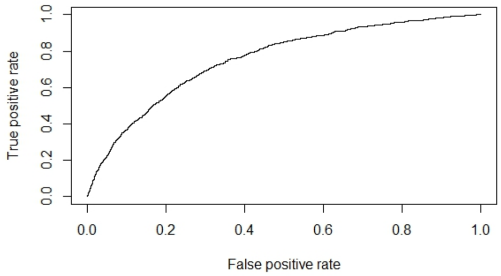

Figure 8: ROC curve of Logisitic Regression: Here we see that the tradeoff between specificity and sensitivity is non-ideal. That is, there is no obvious optimal trade off at which errors are minimized.

accuracy	.86	AUC	0.75
error residual	0.11	F score	0.11
specificity	0.99	sensitivity	0.06

Table 3: Static Performace Measures of Logistic Regression: To compare logistic regression to the optimal tree, we note the same single confusion matrix values as before. While specificity increased, all other measures perform worse than the rpart predicitve model.

For logistic regression, all factors utilized in the rpart model were used. Each variable was subjected to a null hypothesis, p value test with cutoff $p = 0.05$ in order to be included in the final model. In addition, once the logistic model was fitted, terms below a certain threshold were removed in order to avoid overfitting. The resulting fit had a much higher specificity than the rpart predicitve model since type I and type II errors were equally weighted. Else, all performance measures indicated the logisit regression performed at a level equivalent to or below that of the classification tree. It is notable that at the same sensitivity of the tree model, the logistic model has much worse specificity. While it is certainly true that the logisitic model was constructed naively, it does indicate that the method we have implemented in this report has isolated some higher degree of signal.

3.5 Forced Divisions

In the construction of the optimal tree, the data was first divided by level of asthma control with little further branching necessary. It would be informative to know what predicts an ED event within these categories. When we run the rpart algorithm with a forced initial split on asthma control,

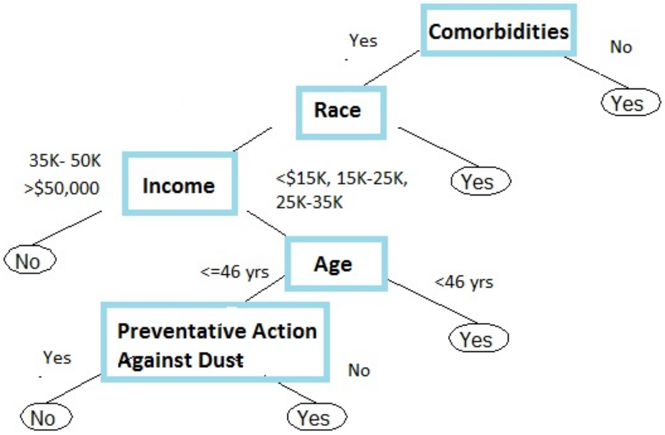

Figure 9: Continued Branches for Poorly Controlled Asthma Patients: Of the asthma control levels, only in the poorly controlled cases of asthma was enough signal gained from extensive branching to counteract the favoring of a small tree. The initial split on comorbidity is not surprising: patients with COPD are, in fact, treated very differently than those with asthma as a diagnosis of COPD effectively removes the temproary nature of asthma. Most interesting is the split on age in which younger patients are classified as at risk of ED episodes. Normally, elderly patients experience more severe symptoms and are generally considered more likely to seek emergency medical care.

we observe fuller classification trees. Well Controlled patients are simply kept as a pure node. Less than 4% of patients with well controlled asthma experienced an ED event; this prevalence rate makes the target too small to acheive significant improvment in risk while favoring small trees. Those categorized as “Not Well Controlled” were further divided into race as in the original model. Significant differentiation was seen in the group categorized as “Poorly Controlled” with the following added divisions. This does better the performance measures, but not by much. In the course of trying to better these scores, we attempted forced divisions in several categories, including HHS region, race and age. While each tree greatly differed in structure and thus information, the performance measures remained comparable to or less than the initial, optimal tree and will therefore not be further explored in this report.

4 Summary and Discussion

Within the work of this report, we have explored a raw, publically avaliable data set, the BRFSS Asthma Call Back Survey. Following the guidelines of the National Heart, Lung and Blood Association's EPR3, a measure of asthma control was created and the patients from 2006 to 2009 were classified as having "Well Controlled","Not Well Controlled", or "Poorly Controlled" asthma. Taking cues from the literature and known correlations in asthma severity, the data was searched for these relationships, many of which proved true. In addition to factors extracted from the ACBS data, the asthma control score was included in the construction of a model for predicting ED events.

The tree resulting from implementing the rpart algorithm was remarkably simplistic: the majority of signal was isolated by fully separating patients by control levels. Examination of tree construction did reveal the appearance of desired (and expected) factors such as socioeconomic situation, age and comorbidities. We beleive that control implicitly accounted for demographic factors that would otherwise show up in tree construction. Although the model does not perform at an extremely high level, the compromise of sensitivity and overall error is acceptable in terms of real-world implementation and comperable to similar models in the literature. In addition, comparison to logistic regression, the most common approach to predictive modeling within the literature, proved the rpart model as being more effective overall.

Knowing that a larger tree would improve these measures by some degree, we then forced tree growth from the initial divition on asthma control. "Well Controlled" patients experienced so few ED episodes that no branching could produce enough improvment to overcome restrictions on tree size. Similarly, "Not Well Controlled" patients produced only one additional split on race, which had already appeared in the optimal tree. This is somewhat surprising; it seems reasonable to believe that "middle of the road" patients would have the most diverse set of defining characteristics. However, "Poorly Controlled" patients proved to be different enough as to encourage further branching. While the rpart algorithm does have distinct strengths (such as its treatment of missing data and ability to capture intricate dependences between variables) after the scope of this analysis, we beleive that it may not be the most effective classification scheme for ED events given basic demographic data. A model which does not so strongly favor small tree structures would probably prove more efficient and allow for deeper analysis.

Appendix A

Variables Utilized for Calculation of Control Levels

No.	Attribute	Code	Possible Values
Control Computation			
1	How many symptomatic days in past month	SYMP_30D	Numeric
2	Do you have symptoms all the time	DUR_30D	Yes,No
3	How many nights have you been woken up by asthma symptoms	ASLEEP30	Numerical
4	In past 3 months, how many asthma attacks have you had	EPIS_TP	Numerical
5	To what extent have you modified your daily activities due to asthma	ACT_DAYS	Not at all, a little, a moderate amount, a lot
6	Pill Corticosteroid prescription	PILL_CS	Yes, No
7	Short acting beta-2 antagonist inhaler	INH_B2AS	Numeric(count of Rx)
8	Times per day or per week patient uses ihaler corresponding to inhaler ID	ILP08_(inhaler ID)	Numeric (300 indicates days, 400 indicates weeks)

Appendix B

Definition of Terms Used in RPART algorithm

A	a single node
T	the collective tree
C	numer of classes
$c(A)$	class assigned to terminal node A
ϕ	chosen impurity function(gini, information, entropy etc)
$L(i,j)$	loss matrix for saying individual with true label i belongs in class j, $L(i,i) \equiv 0$
p_{iA}	, probability of the class of an individual given that the individual is in node A
$R(A)$	$= \sum_{i=1}^{C} p_{iA} L(i, c(A))$, risk of node A
$I(A)$	$= \sum_{i=1}^{C} \phi(p_{iA})$,impurity of node A
R_α	$= R(T) + \alpha \lvert T \rvert R(T_0)$, risk of tree for particular complexity parameter α

Appendix C

Variables Utilized for Predictive Model

No.	Attribute	Code	Possible Values
Pridictive Model			
9	Age at interview	AGE	Numeric
10	Bmi in kg and m	_BMI4	Numeric
11	Self-identified race	RACE2	White only, Black only, Hispanic, Asian, Native Hawaiian/Pacific Islander, American Indian/Alaskan Native, Multiracial, Other
12	Sex	SEX	Male, Female
13	Metropolitain Status Code	MSCODE	center of MSA,outside center city, suburb, MSA with no center city, not in MSA
14	Highest completed education	_EDUCAG	Did not graduate HS, graduated HS, Some college/technical school, graduated college/technical school
15	Household Income level	_INCOMG	<15K,15-25K,25-35K,35-50K,>50K
16	Smoking status	_SMOKER3	Current every day, Current some days, Former, Never
17	In the past week, has anyone smoked inside your house	S_INSIDE	Yes, No
18	Smelled mold or musty odor in house	ENV_MOLD	Yes, No
19	Pets in house	ENV_PETS	Yes, No
20	Is there anyone you consider your primary health care provider	PERSDOC2	More than One, One, No

21	Have you been told by a health care professional that you are depressed	DEPRESS	Yes, No
22	Do you use a humidifier in the house	DEHUMID	Yes, No
23	Has there been a time you wanted to see a physician for your asthma but could not due to cost	ASMDCOST	Yes, No
24	Has there been a time you needed to fill a prescription for your asthma but could not due to cost	ASRXCOST	Yes, No
25	Have you been without insurance for any time in the past year	INS2	Yes, No
Combined Variables, Predictive Model			
	Attribute	Code	Variables Combined
26	Pest Sightings in past month	PEST	C_ROACH and C_RODENT
27	Precationary measures against dust mites	DUST	E_PILLOW, MATTRESS, AIRCLEANER, HOTWATER
28	Comorbidities	COMORBID	COPD and EMPHY
29	ED event	EMER_MED	ER_VISIT and HOSP_VST
30	HHS region assignment	HHS	_STATE

References

[1] Akinbami, LJ. *Asthma Prevalence, Health Care Use, and Mortality: United States, 2005-2009.* National Health Statistics Report. No. 32, Jan 2011.

[2] Breiman, L., Friedman *Classification and Regression Trees.* Monterey: Wadsworth and Brooks/Cole,1984.

[3] Farion, Ken. *A Tree-Based Decision Model to Support Prediction of the Severity of Asthma Exacerbations in Children.* Journal of Medical Systems, Vol 34, No. 4, August 2010.

[4] Gunnells, Linda. *Very Poorly Controlled Asthma Among Adults in Washington State.* Washington State Journal of Public Health Practice, Vol. 3, No. 1, 2010.

[5] National Heart, Lung, and Blood Institute. *Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma 2007.*. Bethesda, MD: National Institutes of Health; August 2007. NIHPublication 07-4051.

[6] Raileanu, Laura Elena. *Theoretical comparison between the Gini Index and Information Gain criteria.* Annals of Mathematics and Artificial Intelligence, Vol. 41, No. 1 May 2004.

[7] Therneau, T. M. and Atkinson, E. J. *An introduction to recursive partitioning using the RPART routines.* Technical report, Mayo Foundation. 1997.

[8] Zahran, Hetice *Predictors of Asthma Self-Management Education among Children and Adults– 2006-2007 Behavioral Risk Factor Surveillance System Asthma Call-Back Survey.* Journal of Asthma, Vol. 49, No. 1, February 2012.

[9] Winer, Rachel *Asthma Incidence among Children and Adults: Findings from the Behavioral Risk Factor Surveillance System Asthma Call-Back Survey– United States, 2006-2008.* Journal of Asthma, Vol. 49, No. 1, February 2012.

[10] Zhang, Heping. *Recursive Partitioning and Applications, 2nd ed.* Dordrecht: Springer 2010.