Characterization of Patients who Suffer Asthma Exacerbations using Data Extracted from Electronic Medical Records

Blanca E. Himes, PhD[1,2,3,4], Isaac S. Kohane, MD, PhD [1,2,4], Marco F. Ramoni, PhD[1,2,4], Scott T. Weiss, MD, MS[3,4]

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA; [2]Children's Hospital Informatics Program, Harvard Medical School, Boston, MA; [3]Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; [4]Harvard Partners Center for Genetics and Genomics, Boston, MA

Abstract

The increasing availability of electronic medical records offers opportunities to better characterize patient populations and create predictive tools to individualize health care. We determined which asthma patients suffer exacerbations using data extracted from electronic medical records of the Partners Healthcare System using Natural Language Processing tools from the "Informatics for Integrating Biology to the Bedside" center (i2b2). Univariable and multivariable analysis of data for 11,356 patients (1,394 cases, 9,962 controls) found that race, BMI, smoking history, and age at initial observation are predictors of asthma exacerbations. The area under the receiver operating characteristic curve (AUROC) corresponding to prediction of exacerbations in an independent group of 1,436 asthma patients (106 cases, 1,330 controls) is 0.67. Our findings are consistent with previous characterizations of asthma patients in epidemiological studies, and demonstrate that data extracted by natural language processing from electronic medical records is suitable for the characterization of patient populations.

Introduction

Asthma, a chronic respiratory disease, is a serious disease affecting 20.5 million Americans and costing over $16 billion yearly in the US[1]. As the primary reason for asthma hospitalizations and emergency room visits, asthma exacerbations account for a large portion of asthma healthcare expenses[2, 3]. The proper identification of asthma patients who suffer from frequent exacerbations is of clinical importance both to monitor them more carefully and to treat them more aggressively in an effort to reduce exacerbations[4]. Attempts to characterize patients who suffer from frequent exacerbations have been made, but this group is still not well understood.

Electronic medical records have been widely heralded for their potential to improve the quality of patient care[5]. Less obvious is the use of such records to characterize patient populations and develop predictive tools with an eye to the future of personalized medicine. The Partners Healthcare Research Patient Data Registry (RPDR), which contains information on over three million patients that have been treated in Partners-affiliated hospitals, is one of the larger collections of electronic medical records from academic medical centers. The computational infrastructure developed for the "Informatics for Integrating Biology to the Bedside" center (i2b2)[6] has been used to develop an "asthma data mart" from the RPDR[7, 8]. The asthma data mart contains codified annotations of patient records (e.g. billing codes) that have been augmented with additional concepts/phenotypes extracted from the textual notes of medical records using Natural Language Processing (NLP) techniques[9]. In this work, we used data extracted from the asthma data mart to characterize and predict which asthma patients suffer from frequent exacerbations.

Methods

A set of 12,792 asthma patients from the i2b2 asthma data mart extracted from the Partners Healthcare System's RDPR was obtained. Records from these patients have been parsed for coded data and variables such as smoking history extracted by Natural Language Processing of unstructured text[9]. The collection and study of this data is approved by the Institutional Review Board of Partners Healthcare System.

Patients were included if they had been observed for at least 5 years as determined by the dates of the earliest and latest records available, were at least 18 years of age at the initial observation date, and had race, sex, height, weight, smoking history, and birthdate data available. Patients were determined to have had asthma exacerbations on the basis of International Classification of Diseases, Ninth Revision (ICD-9) codes for asthma (i.e. those beginning with 493) being used as admission diagnosis billing codes for hospitalizations or primary diagnosis billing codes for emergency room

visits. If a patient had at least one instance of an asthma code being used for admission and/or primary diagnosis, then the patient was classified as a case, otherwise the patient was considered a control. Demographic variables included race, sex, smoking history, BMI, and age at initial hospitalization or emergency room visit observed. The values of race, as extracted from medical records, were "White," "Black," "Hispanic," and "Asian." The smoking history was dichotomized into "Negative history" if a patient reported not being a current or past smoker in 90 percent of his/her records and "Positive history" otherwise. The BMI was calculated with the average height and weight measurements for each patient as weight in kilograms divided by height in meters squared. The BMI was categorized as follows: underweight (BMI<18), normal (18<=BMI<25), overweight (25<=BMI<30), obese (30<=BMI<40), and morbidly obese (BMI>=40). The age at initial observation was categorized into four groups: 18-44, 45-64, 65-74, and 75+ years old.

Patients were divided into two groups according to initial observation date. A cohort of 11,356 patients (1,394 cases, 9,962 controls), who were initially observed between 1988 and 1998, were selected to model asthma exacerbations. A future independent set of 1,436 patients (106 cases, 1,330 controls) had initial observation dates between 1999 and 2002. The independent group was used to test the predictive accuracy of the model created.

Single variable effects on asthma exacerbations were measured with odds ratios using median-unbiased estimation. Multivariate analyses were performed using forward stepwise logistic regression. The performance of predictive models was evaluated with receiver operating characteristic (ROC) curves. Predictive accuracy was measured as the area under the ROC curve (AUROC), and significance for this accuracy was obtained by comparing the classification ability of models obtained to random classification. All statistical analyses were performed in R[10].

Results

Demographic data (sex, race, BMI, smoking history and age at initial observation) from 11,356 asthma patients was extracted from billing codes and by Natural Language Processing from the i2b2 data mart. Patients were classified into asthma exacerbation cases and controls on the basis of asthma-related emergency department visits and overnight hospitalizations (1,394 cases, 9,962 controls). The effects of single variables on asthma exacerbations are shown in Table 1. Older age (65-74 and 75+ years) was related to decreased asthma exacerbations compared to younger age (18-44 years), odds-ratios of 0.64 and 0.68 respectively. Having smaller or larger BMI than normal is associated with asthma exacerbations, with the greatest risk in morbidly obese and underweight individuals, odds-ratios 2.51 and 2.03 respectively. Race had the strongest effect on asthma exacerbations, with Blacks having four times and Hispanics having three times the risk of exacerbations that Whites and Asians have. A positive smoking history conferred a one-and-a-half increased risk of exacerbations. Gender did not have any significant effect on asthma exacerbations.

	Referent Group	Odds Ratio (95% CI*)
Age		
45-64	18-44	0.96 (0.85, 1.08)
65-74	18-44	0.64 (0.52, 0.78)*
75+	18-44	0.68 (0.51, 0.89)*
BMI		
Underweight	Normal	2.03 (1.20, 3.26)*
Overweight	Normal	1.30 (1.11, 1.53)*
Obese	Normal	1.78 (1.53, 2.08)*
Morbidly obese	Normal	2.51 (2.06, 3.05)*
Race		
Asian	White	0.81 (0.43, 1.41)
Black	White	4.14 (3.58, 4.78)*
Hispanic	White	3.05 (2.62, 3.55)*
Sex	Male	1.03 (0.91, 1.18)
Smoking history	Negative	1.50 (1.31, 1.72)*

Table 1. Individual variable effects on asthma exacerbation. The referent group refers to that used as reference for the odds-ratio. 95% CI = 95 percent confidence interval. *p-value < 0.01.

A multivariable logistic regression model was created with a forward stepwise procedure using the demographic data to predict asthma exacerbations. The model found that race, BMI, smoking history, and age at initial observation are independent predictors of asthma exacerbations (Table 2). These results are similar to the relationships among demographic variables and asthma exacerbations found with single variable analysis. The 65-74 year-old age group had significantly less exacerbations than the younger age group, but the 45-64 and 75+ year-old age groups had a similar risk when adjusting for other variables. Being overweight was not significantly different than having normal BMI for asthma exacerbation risk, but being underweight, obese, and especially morbidly obese were associated with higher risks of asthma exacerbation. Race continued to have the strongest effect on asthma

exacerbations, with Blacks and Hispanics having a high risk of exacerbations compared to Whites and Asians. A positive smoking history increased the risk of exacerbations, independently of the other variables.

	Referent Group	Adjusted Odds Ratio (95% CI)
Race		
Asian	White	0.97 (0.54, 1.76)
Black	White	3.86 (3.33, 4.47)*
Hispanic	White	3.08 (2.64, 3.60)*
Smoking history	Negative	1.71 (1.48, 1.97)*
BMI		
Underweight	Normal	1.96 (1.18, 3.25)*
Overweight	Normal	1.15 (0.98, 1.36)
Obese	Normal	1.40 (1.19, 1.64)*
Morbidly obese	Normal	1.75 (1.43, 2.15)*
Age		
45-64	18-44	0.99 (0.87, 1.12)
65-74	18-44	0.74 (0.60, 0.91)*
75+	18-44	0.84 (0.63, 1.12)

Table 2. Independent predictors of asthma exacerbation, in order added to logistic regression model. The referent group refers to that used as reference for the adjusted odds-ratio. 95% CI = 95 percent confidence interval. *p-value < 0.01.

The predictive accuracy of the logistic regression model was tested by using it to predict asthma exacerbations in an independent set of 1,436 asthma patients (106 cases, 1,330 controls). The AUROC corresponding to the classification of these subjects is 0.67 (p-val < 10^{-9}).

To test the robustness of our model to changes in the definition of asthma exacerbations, we repeated the analysis using two stricter definitions of asthma exacerbations. Cases were defined as having (1) two or more asthma-related ER visits or hospitalizations and (2) three or more asthma-related ER visits or hospitalizations. Subjects with no asthma-related ER visits or hospitalizations remained as controls. With these more severe definitions of asthma exacerbations, the number of cases decreased and the AUROCs increased (Table 3).

Exacerbation Definition	Model Cases	Ind Cases	AUROC	p-value
1+ ER/hosp	1394	106	0.67	5.30E-09
2+ ER/hosp	668	39	0.71	2.88E-06
3+ ER/hosp	403	14	0.83	9.21E-06

Table 3. Change in predictive accuracy with more severe exacerbation definitions (i.e. increased number of ER visits/hospitalizations). For each exacerbation definition, there were 9962 controls for the model construction, and 1330 controls used in the AUROC calculation.

Discussion

A well-known goal of the use of electronic medical records is to improve the quality and efficiency of patient care[5]. More recently, electronic medical records have been acknowledged as a potential source to identify a large number of subjects for research studies. Such studies include those limited to data collected through electronic medical records (e.g. the understanding of individual disease courses and outcomes), but also extend to those requiring additional data (e.g. genetics studies of complex diseases). With these goals in mind, natural language processing tools have been used to extract data from electronic medical records for the identification of subjects with diseases including asthma[9], diabetes mellitus[11], and heart failure[12]. In this work, we used data extracted from the i2b2 asthma data mart to characterize and predict which asthma patients suffer from frequent exacerbations.

Of the phenotypes used in this work, smoking history is the most challenging to extract from EMRs. Previous work describes some of the problems involved in determining smoking status and how NLP has been used to successfully determine it[8, 9]. In addition to the challenge of extracting smoking status, there can be uncertainty in the veracity of patient reporting. We chose to define a negative smoking history as one where a patient had "never smoked" in at least 90% of the smoking histories extracted. A positive smoking history could contain a mixture of "current-smoker," "past-smoker," and less than 90% "never-smoked" in their smoking histories. Most often, positive smokers had a large percentage of "current-smoker" and "past-smoker" as extracted smoking histories. Because we looked at records from patients who were observed at least 5 years, it was deemed more significant to differentiate patients with positive smoking history from those with a negative history, than to differentiate current smokers from past smokers. Therefore, we chose a conservative definition of non-smokers to try to ensure that this group truly contained patients with a negative history of smoking.

Using univariable and multivariable analyses, we found that race, BMI, smoking history, and age at initial observation are predictors of asthma exacerbations, while gender had no significant effect. We found that Black and Hispanic patients have increased rates of exacerbation compared to whites, as has been observed previously[1, 13]. In our univariable analysis, patients with any BMI that was different than normal were at increased risk for asthma exacerbations. In the multivariable analysis, the underweight, obese, and morbidly obese patients were still at an increased risk for asthma exacerbations. Consistent with these findings, it has previously been reported that the percentage of overweight and obese subjects among asthma exacerbation patients is larger than that in the general population[14]. Despite the challenges in defining smoking status described in the previous paragraph, our result that a positive history of smoking increases the risk of asthma exacerbations is consistent with previous findings[15]. Our age-at-initial-observation results, which suggest that older age is associated with decreased asthma exacerbations, are consistent with rates recorded by the Amerian Lung Association between 1997 and 2004, where 45-64 year-olds have a similar rate of asthma attack prevalence compared to 18-44 year-olds, while 65+ year-olds have a lower rate of attacks than younger individuals[1]. We found no difference in exacerbation rates between men and women. Although exacerbation rates have generally been observed to be higher in adult women than adult men[1, 16], no gender differences have been observed in other studies[13].

The accuracy of our logistic regression model was tested by using it to predict asthma exacerbations in an independent set of patients. The corresponding AUROC was 0.67. This AUROC is significantly better than that of a random classifier (i.e. 0.5), but it is lower than that of a desirable clinical classifier. However, it demonstrates robustness of our model when applied to an individual set of patients, who were initially observed at future time points than our original set of patients.

Defining which patients are asthma exacerbators is challenging, particularly when using data from clinical records only, because patients exhibit a range of asthma severity, which is not clearly dichotomous. We chose a cutoff of 1 ER visit and/or hospitalization with a primary diagnosis code for asthma as the definition that would include as many asthma exacerbators as possible. For comparison, we constructed models using stricter definitions of exacerbation: 2 or more ER visits or hospitalizations, and 3 or more ER visits or hospitalizations (Table 4). The number of cases decreased significantly with these stricter definitions, but the multivariable analyses results remained consistent with the findings reported in Tables 1 and 2. The AUROCs when predicting exacerbations in the independent set of patients increased from 0.67 to 0.71 and 0.83, suggesting that patients suffering from the most exacerbations are more demographically homogeneous. A larger number of patients with more severe exacerbations would be necessary to carefully address patient differences among the definitions.

In a previous study in which electronic medical records were used to predict mortality in reactive airway disease patients, a multivariate model was constructed using a comprehensive set of clinical variables that included comorbidities, medication use, and diagnostic test results[17]. These records were available from a computer based-system and did not have to be retrieved using NLP tools as the data for this study did. In the future, the increasing standardization of EMR data entry may improve the ease with which clinical data is extracted from EMR records, thereby reducing some of the obstacles that must be overcome when using NLP on free text records. Meanwhile, advances in NLP tailored for clinical records will help to extract useful data from existing EMRs. As additional medical record data extracted by i2b2 tools becomes available, more comprehensive models to understand and predict asthma exacerbations will be developed.

Conclusion

We have used data extracted from electronic medical records with natural language processing tools developed by i2b2 to characterize asthma patients who suffer from frequent exacerbations. We found that race, BMI, smoking history, and age at initial observation are independent predictors of asthma exacerbations and that these variables predict exacerbations in an independent group of patients with an AUROC of 0.67. Our findings are consistent with previous characterizations of asthma patients in epidemiological studies, supporting the use of data extracted by natural language processing from electronic medical records for the characterization of patient populations.

Acknowledgements

We thank Shawn Murphy, MD, PhD, from Massachusetts General Hospital at Harvard Medical School, for facilitating access to the i2b2 asthma data mart. This work was supported by the following NIH grants: 5U54LM008748-02 (National Centers for Biomedical Computing), 2U01HL065899 (NHLBI) and 2T15LM007092-16 (NLM).

References

1. American Lung Association. Trends in Asthma Morbidity and Mortality. New York, NY: Epidemiology and Statistics Unit, Research and Program Services, American Lung Association; 2006.
2. Lane S, Molina J, Plusa T. An international observational prospective study to determine the cost of asthma exacerbations (COAX). Respir Med 2006;100(3):434-50.
3. Gendo K, Lodewick MJ. Asthma economics: focusing on therapies that improve costly outcomes. Curr Opin Pulm Med 2005;11(1):43-50.
4. Heaney LG, Robinson DS. Severe asthma treatment: need for characterising patients. Lancet 2005;365(9463):974-6.
5. Committee on Quality of Health Care in America IoM. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington, D.C.: National Academy Press; 2001.
6. i2b2: Informatics for Integrating Biology and the Bedside. 2005. (http://www.i2b2.org.)
7. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc 2006:1040.
8. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(1):14-24.
9. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006;6:30.
10. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2007.
11. Wilke RA, Berg RL, Peissig P, et al. Use of an electronic medical record for the identification of research subjects with diabetes mellitus. Clin Med Res 2007;5(1):1-7.
12. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 2007;13(6 Part 1):281-8.
13. Griswold SK, Nordstrom CR, Clark S, Gaeta TJ, Price ML, Camargo CA, Jr. Asthma exacerbations in North American adults: who are the "frequent fliers" in the emergency department? Chest 2005;127(5):1579-86.
14. Thomson CC, Clark S, Camargo CA, Jr. Body mass index and asthma severity among adults presenting to the emergency department. Chest 2003;124(3):795-802.
15. Silverman RA, Boudreaux ED, Woodruff PG, Clark S, Camargo CA, Jr. Cigarette smoking among asthmatic adults presenting to 64 emergency departments. Chest 2003;123(5):1472-9.
16. Schatz M, Clark S, Camargo CA, Jr. Sex differences in the presentation and course of asthma hospitalizations. Chest 2006;129(1):50-5.
17. Tierney WM, Murray MD, Gaskins DL, Zhou XH. Using computer-based medical records to predict mortality risk for inner-city patients with reactive airways disease. J Am Med Inform Assoc 1997;4(4):313-21.