Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data

Wei Wu, PhD,[a] Eugene Bleecker, MD,[b] Wendy Moore, MD,[b] William W. Busse, MD,[c] Mario Castro, MD,[d] Kian Fan Chung, MD,[e] William J. Calhoun, MD,[f] Serpil Erzurum, MD,[g] Benjamin Gaston, MD,[h] Elliot Israel, MD,[i] Douglas Curran-Everett, PhD,[j] and Sally E. Wenzel, MD[j,k]

Pittsburgh, Pa, Winston-Salem, NC, Madison, Wis, St Louis, Mo, London, United Kingdom, Galveston, Tex, Cleveland, Ohio, Boston, Mass, and Denver, Colo

Background: Previous studies have identified asthma phenotypes based on small numbers of clinical, physiologic, or inflammatory characteristics. However, no studies have used a wide range of variables using machine learning approaches.
Objectives: We sought to identify subphenotypes of asthma by using blood, bronchoscopic, exhaled nitric oxide, and clinical data from the Severe Asthma Research Program with unsupervised clustering and then characterize them by using supervised learning approaches.
Methods: Unsupervised clustering approaches were applied to 112 clinical, physiologic, and inflammatory variables from 378 subjects. Variable selection and supervised learning techniques were used to select relevant and nonredundant variables and address their predictive values, as well as the predictive value of the full variable set.
Results: Ten variable clusters and 6 subject clusters were identified, which differed and overlapped with previous clusters. Patients with traditionally defined severe asthma were distributed through subject clusters 3 to 6. Cluster 4 identified patients with early-onset allergic asthma with low lung function and eosinophilic inflammation. Patients with later-onset, mostly severe asthma with nasal polyps and eosinophilia characterized cluster 5. Cluster 6 asthmatic patients manifested persistent inflammation in blood and bronchoalveolar lavage fluid and exacerbations despite high systemic corticosteroid use and side effects. Age of asthma onset, quality of life, symptoms, medications, and health care use were some of the 51 nonredundant variables distinguishing subject clusters. These 51 variables classified test cases with 88% accuracy compared with 93% accuracy with all 112 variables.
Conclusion: The unsupervised machine learning approaches used here provide unique insights into disease, confirming other approaches while revealing novel additional phenotypes. (J Allergy Clin Immunol 2014;133:1280-8.)

Key words: *Asthma phenotyping, variable analysis, unsupervised approaches, supervised machine learning approaches*

The definition of asthma (appropriate symptoms in association with reversible airflow limitation) allows a heterogeneous group of patients to be included under this term.[1] Clinical and statistical efforts have assigned patient phenotypes, with recent emphasis on

From [a]the Lane Center for Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh; [b]the Center for Human Genomics, School of Medicine, Wake Forest University, Winston-Salem; [c]the Division of Allergy and Immunology, University of Wisconsin, Madison; [d]the Division of Pulmonary & Critical Care Medicine, Washington University School of Medicine, St Louis; [e]the National Heart & Lung Institute, Imperial College, London; [f]the Department of Internal Medicine, University of Texas Medical Branch, Galveston; [g]the Department of Pulmonary, Allergy and Critical Care Medicine, Cleveland Clinic, Cleveland; [h]the Division of Pediatric Pulmonology, and Allergy/Immunology, Department of Pediatrics, School of Medicine, Case Western Reserve University, Cleveland; [i]the Pulmonary Division, Brigham and Women's Hospital, Boston; [j]National Jewish Medical and Research Center, University of Colorado Health Sciences Center, Denver; and [k]the Asthma Institute, Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh.

Supported by National Institutes of Health grants R01GM087694, RO1-HL69174, HL69116, HL69130, HL69149, HL69155, HL69167, HL69170, HL69174, HL69349, M01 RR018390, M01RR07122, M01 RR03186, and HL087665.

Disclosure of potential conflict of interest: W. Wu has received research support from the National Institutes of Health (NIH). E. Bleecker has been supported by a National Heart, Lung, and Blood Institute (NHLBI) Severe Asthma Research Program (SARP) grant and has received research support from the NIH. W. Moore has received research support from the NHLBI. W. W. Busse has been supported by the NIH/NHLBI; is a Board member for Merck; has consultancy arrangements with Amgen, Novartis, GlaxoSmithKline, MedImmune, Genentech, Boston Scientific, and ICON; has received one or more grants from or has one or more grants pending with the NIH/National Institutes of Allergy and Infectious Diseases (NIAID) and the NIH/NHLBI; and has received royalties from Elsevier. M. Castro has been supported by one or more grants from the NIH and the American Lung Association (ALA); has received support for travel from the NIH; has consultancy arrangements with Asthmatx/Boston Scientific, Genentech, IPS, Pulmagen, and Sanofi-Aventis; has received research support from Asthmatx/Boston Scientific, Amgen, Ception/Cephalon/Teva, Genentech, Medimmune, Merck, Novartis, GlaxoSmithKline, Sanofi-Aventis, and Vectura; has received one or more payments for lecturing from or is on the speakers' bureau for Pfizer, Merck, GlaxoSmithKline, Genentech, and Asthmatx/Boston Scientific; and has received royalties from Elsevier. K. F. Chung is a Board member for GlaxoSmithKline, Gilead, and NERC; has received research support from NERC, the NIH, and BHF; and has received one or more payments for lecturing from or is on the speakers' bureau for GlaxoSmithKline, AstraZeneca, and Merck. W. J. Calhoun has received research support from and travel support from the NHLBI. B. Gaston has received research support from the NIH, has received lecture fees from Aerocrine, and has patents with RRI, Galleon, and N30. E. Israel has consultancy arrangements with Cowen & Co, Infinity Pharmaceuticals, MedImmune (now AstraZeneca), Merck, NKT Therapeutics, Ono Pharmaceuticals, Regeneron Pharmaceuticals, TEVA Specialty Pharmaceuticals, Gilead Sciences, Johnson & Johnson, and Novartis; has received research support from Aerovance, Amgen, 3 Research (Biota), Genentech, MedImmune, and Novartis; and has received one or more payments for lecturing from or is on the speakers' bureau for Merck, Novartis, and Genentech. S. E. Wenzel has received research support from Amgen, Array, GlaxoSmithKline, MedImmune, and Sanofi; has received one or more consulting fees or honoraria from Actelion, Gilead, Amgen, GlaxoSmithKline, MedImmune, and Teva; and has received one or more payments for travel/accommodations/meeting expenses from Sanofi. The rest of the authors declare that they have no relevant conflicts of interest.

Received for publication March 23, 2013; revised October 8, 2013; accepted for publication November 11, 2013.

Available online February 28, 2014.

Corresponding author: Wei Wu, PhD, Lane Center for Computational Biology, School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. E-mail: weiwu2@cs.cmu.edu. Or: Sally E. Wenzel, MD, University of Pittsburgh Asthma Institute, Pulmonary, Allergy and Critical Care Medicine, NW 931 Montefiore, 3459 Fifth Ave, Pittsburgh, PA 15213. E-mail: wenzelse@upmc.edu.

0091-6749/$36.00

http://dx.doi.org/10.1016/j.jaci.2013.11.042

Abbreviations used
AQLQ: Asthma Quality of Life Questionnaire
BAL: Bronchoalveolar lavage
FENO: Fraction of exhaled nitric oxide
HC: Healthy control subject
HCU: Health care use
INFOGAIN: Information gain
SA: Severe asthma
SARP: Severe Asthma Research Program

statistical efforts. Seven clustering studies have been reported,[2-8] with 2 specifically including extremely well-characterized adults with severe asthma (SA).[2,3] The Leicester study used a *k*-means clustering approach[3] that was limited to 16 variables but included sputum eosinophil counts to identify 4 severe and 2 mild-to-moderate asthma clusters. Moore et al[2] performed a hierarchical clustering of asthmatic patients from the Severe Asthma Research Program (SARP), reducing 628 variables to 34 core variables to identify 5 clusters. Variables related to inflammatory markers were not included. Thus additional approaches are needed incorporating greater numbers of variables, inflammatory markers, or both.

Machine learning techniques have recently been applied to human diseases. Algorithms have assisted in selecting features from thousands of genes to facilitate biomarker identification and accurate patient diagnosis. For example, a feature selection framework based on information theory involves ranking features by using a correlation measurement information gain (INFOGAIN) and then selecting relevant and nonredundant features by using a Markov blanket algorithm.[9] This framework was applied to a leukemia microarray data set to obtain subsets of nonredundant features from 7130 genes that distinguished acute lymphocytic from acute myeloid leukemia with high accuracy.[10] When the top 3 informative and nonredundant genes were selected to predict leukemia types, 100% classification accuracy was achieved. Therefore machine learning techniques offer promising approaches to understand complex diseases.

Although a clustering analysis was previously performed in SARP, blood/bronchoscopic inflammatory characteristics and IgE and fraction of exhaled nitric oxide (FENO) values were not included because the majority had not undergone these procedures/tests. Additionally, healthy control subjects (HCs) were excluded because the majority of variables analyzed were clinically related to asthma. This second SARP analysis performed an unsupervised clustering of more than 100 variables on 378 asthmatic patients and HCs who had undergone bronchoscopy to incorporate inflammatory variables. A feature selection framework selected 51 relevant and nonredundant variables. Predictive values of the full set of 112 variables and the selected 51 variables were assessed for their ability to distinguish subject clusters.

METHODS

Patient population

Subjects were all part of SARP. Characterization of the subjects can be found in the Methods section in this article's Online Repository at www.jacionline.org. SA was defined by 2000 American Thoracic Society workshop criteria (see the Methods section in this article's Online Repository).[11] Other asthmatic patients were divided into 4 groups on the basis of FEV_1 percent predicted values and inhaled corticosteroid use, as previously described.[12] HCs from SARP were included as well. This clustering was limited to subjects with "lung" variables, including bronchoalveolar lavage (BAL) fluid cell counts and FENO values (see the Methods section in this article's Online Repository).

Computational and statistical analysis

Data preprocessing. Variables with 5% or greater missing data were excluded. Missing values in variables with less than 5% missing data were added by using a k-NNimpute algorithm.[13]

Cluster analysis. A *k*-means clustering method was applied to the preprocessed data to partition subjects (including HCs). A Ward agglomerative hierarchic clustering method applied to the preprocessed data grouped clinical variables.

Statistical tests. Data for continuous variables were log transformed to improve normality of distribution. Welch *t* statistics,[14] which allow for data in different groups with unequal variances and are popularly used in microarray analysis,[15] were used for $\log_2$-transformed data from continuous variables. χ^2 Tests were used for data from categorical variables to find differences between subjects in the total SARP cohort and those in this cluster analysis. ANOVA and pairwise *t* tests were used for continuous variables with $\log_2$-transformed data, Kruskal-Wallis and pairwise Wilcoxon rank sum tests were used for categorical ordinal variables, and χ^2 tests and pairwise Fisher exact tests were used for categorical binary variables to find significant differences between subject clusters. *P* values were adjusted for multiple testing by using a false discovery rate procedure.[16] *P* values of less than .05 were considered significant. Data presented in the text and figures are on the original scale.

Ranking variables by using INFOGAIN. INFOGAIN[17] measures how well a variable predicts subject cluster labels, as determined by using *k*-means clustering. Higher INFOGAIN values identify which variables best predict subject cluster labels. Calculating INFOGAIN values requires continuous variables to undergo discretization. After calculating their INFOGAIN values, variables were ranked according to their values.

Selection of relevant and nonredundant variables. A set of relevant variables was selected if their INFOGAIN values were greater than a threshold of 0.05; redundant variables were removed by using a Markov blanket algorithm.[18]

Classification of clustered subjects. Subjects whose cluster labels were assigned by means of *k*-means clustering were classified by using a multiclass support vector machine algorithm. The 378 subjects were split into a training data set (80% subjects) and a test data set (20% subjects). Leave-one-out cross-validation was performed to estimate classification errors.

Further details for the following methods can be found in the Methods section in this article's Online Repository: data preprocessing, cluster analysis, statistical tests, discretization and ranking variables by using INFOGAIN, selection of relevant nonredundant variables, and classification of clustered subjects.

RESULTS

Demographics

Only 378 subjects (of 1685 total) with BAL, blood, and FENO data were analyzed. These subjects, including HCs and patients with mild asthma, moderate asthma, and SA, did not generally differ from the total cohort (see Table E1 in this article's Online Repository at www.jacionline.org) but were older and with slightly better lung function likely caused by limitations on research bronchoscopy in subjects less than 18 years old or with more severe obstruction.

Clustering results

Three hundred seventy-eight subjects were clustered into 6 groups (Table I). One hundred twelve variables were clustered