Published in final edited form as:
J Asthma. 2020 November ; 57(11): 1155–1167. doi:10.1080/02770903.2019.1642352.

Prodromal clinical, demographic, and socio-ecological correlates of asthma in adults: A 10-year statewide big data multi-domain analysis

Jennifer N Fishe, MD[1], **Jiang Bian**[2], **Zhaoyi Chen**[3], **Hui Hu**[3], **Jae Min**[3], **Francois Modave**[4], **Mattia Prosperi**[3,*]

[1]Department of Emergency Medicine, University of Florida College of Medicine – Jacksonville, FL, 32209, USA.

[2]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville FL 32610, USA.

[3]Department of Epidemiology, College of Medicine & College of Public Health and Health Professions, University of Florida, Gainesville FL 32610, USA.

[4]Center for Health Outcomes and Informatics Research, Loyola University Chicago, Chicago IL 60153, USA.

Abstract

Objectives: To identify prodromal correlates of asthma as compared to chronic obstructive pulmonary disease and allied-conditions (COPDAC) using a multi domain analysis of socio-ecological, clinical, and demographic domains.

Methods: This is a retrospective case-risk-control study using data from Florida's statewide Healthcare Cost and Utilization Project (HCUP). Patients were grouped into three groups: asthma, COPDAC (without asthma), and neither asthma nor COPDAC. To identify socio-ecological, clinical, demographic, and clinical predictors of asthma and COPDAC, we used univariate analysis, feature ranking by bootstrapped information gain ratio, multivariable logistic regression with LogitBoost selection, decision trees, and random forests.

Results: A total of 141,729 patients met inclusion criteria, of whom 56,052 were diagnosed with asthma,85,677 with COPDAC, and 84,737 with neither asthma nor COPDAC. The multi-domain approach proved superior in distinguishing asthma versus COPDAC and non-asthma/non-COPDAC controls (area under the curve (AUROC) 84%). The best domain to distinguish asthma from COPDAC without controls was prior clinical diagnoses (AUROC 82%). Ranking variables from all the domains found the most important predictors for the asthma versus COPDAC and controls were primarily socio-ecological variables, while for asthma versus COPDAC without

[*]correspondence to Mattia Prosperi, *MEng, PhD* (m.prosperi@ufl.edu), Department of Epidemiology, College of Public Health and Health Professions & College of Medicine, University of Florida, 2004 Mowry Road, Gainesville 32610-0231, Florida, USA, phone: +1-352-273-5860.

Author contribution: MP designed the study, carried out the analysis, drafted and revised the manuscript. JF, JB, ZC, HH, JM, and FM drafted and revised the manuscript. All authors read and approved the final manuscript.

controls, demographic and clinical variables such as age, CCI, and prior clinical diagnoses, scored better.

Conclusions: In this large statewide study using a machine learning approach, we found that a multi-domain approach with demographics, clinical, and socio-ecological variables best predicted an asthma diagnosis. Future work should focus on integrating machine learning-generated predictive models into clinical practice to improve early detection of those common respiratory diseases.

Keywords

asthma; COPD; prediction; multi-domain; machine learning

Introduction

Asthma is one of the most prevalent disease worldwide, affecting over 300 million individuals.(1). Asthma and asthma-related mortality cost the United States healthcare system over $50 billion between 2008 and 2013. (2) Asthma patients often suffer from other comorbidities, with subsequent adverse health outcomes.(3) Increasing comorbidities correlate with higher numbers of unscheduled asthma healthcare encounters and the risk of death.(4) Therefore, identification of asthma predictors is necessary to improve patient outcomes through primary prevention, acute treatment, and long-term care management.

However, asthma has multiple causative pathways and disease endotypes.(5)’ (6,7) As such, diagnosing asthma is difficult given its frequent comorbidities and varied symptoms which may overlap with other respiratory conditions such as chronic obstructive pulmonary disease (COPD)(termed asthma COPD overlap syndrome (ACOS)). Additionally, there is great heterogeneity amongst patients with ACOS, (8,9) and therefore COPD and COPD and allied-conditions (COPDAC) may be mistaken for asthma/ACOS. The complexity of distinguishing asthma/ACOS from COPDAC may lead to high proportions of misdiagnosis and treatment failure.(10)

Attempts to identify asthma/ACOS predictors and distinguish asthma/ACOS from COPDAC using clinical scores, physiologic and laboratory testing, and socio-ecologicalfactors have produced promising candidate variables. (11-15) However, most prior studies limited prediction analysis to a single domain (e.g. genetics, lung function, allergen screening), using information from other domains to reduce confounding. We have previously shown that integration of multiple domains better explains variance of asthma onset and can be used effectively to develop composite risk models with high sensitivity and specificity.(16) An asthma versus COPDAC prediction model which contains modifiable factors across multiple domains would be useful to clinicians and patients alike in terms of prevention and treatment strategies.

Therefore, this study’s objective was to use big data for a multi-domain analysis of prodromal correlates of asthma/ACOS, as compared to COPDAC and patients with neither asthma nor COPDAC. We merged healthcare claims data with geo-located ecological and sociodemographic indicators in an exploratory analysis using statistical machine learning.

Our study includes ten years of statewide data from Florida, the state with the third-largest population,(17) including over 21 million people seen in primary care clinics and hospitals.

Methods

This study conforms to Helsinki's declaration and has been approved by the University of Florida's institutional review board with protocol number IRB201701906 (exempt). No written consent was obtained from the participants as this is a secondary analysis using existing database.

The Healthcare Cost and Utilization Project (HCUP)'s State Inpatient Databases (SID), State Ambulatory Surgery and Services Databases (SASD), and State Emergency Department Databases (SEDD) for the state of Florida, United States of America, between 2005 and 2014 (inclusive) were purchased (18). SID, SASD and SEDD contains anonymized, inpatient, outpatient, and emergency room visits data, with patients' demographics, residence (five-digit zip-codes), insurance status, hospital information, diagnoses, and procedures billed (encoded through the International Classification of Diseases, 9th revision (ICD-9) ontology.(19)). HCUP contains a unique de-identified patient code that is constant and presentlongitudinally across multiple visits ("VisitLink"), and this was used to index individual patients and link data longitudinally in this study. The HCUP databases did not contain questionnaire, laboratory, or pharmacy data.

The observational unit of our study was the single patient aged 18 years or older. We designed a case-risk-control study by grouping patients into three groups: i) patients who had a diagnosis of asthma with or without COPDAC, ii) patients who had a COPDAC diagnosis in absence of asthma (ever), and iii) patients who did not have either asthma or COPDAC ever recorded in the merged databases. Asthma was defined as a twice-confirmed diagnostic code ICD9 493.x. A twice-confirmed diagnosis of asthma was chosen to increase specificity in absence of codes for the ordering of pulmonary function tests.(20) We chose to include in the asthma diagnosis patients who may or may not have a COPDAC diagnosis given that the clinical overlap could produce patients with a prior misclassification of COPD. We therefore specifically chose to twice-confirm the asthma diagnosis in order to best ensure those patients did have asthma. COPDAC was defined as any code in the range 490.x-492.x (bronchitis, chronic bronchitis, emphysema) and 494.x-496.x (bronchiectasis, extrinsic allergic alveolitis, and other chronic airway obstruction not elsewhere classified). The baseline time for asthma patients was the first asthma diagnosis date, in presence or absence of any COPDAC. For patients diagnosed with COPDAC in absence of asthma, the baseline was matched to the distribution of years in the asthma population sample, i.e. patients being seen in the same year(s). Figure 1 displays the patient selection criteria and a visual representation of baseline times and data collection. Other study inclusion criteria were: non-missing age, non-missing gender or race/ethnicity, availability of zip-code, and availability of medical records for at least five years before diagnosis year. We also extracted patients without asthma/COPDAC who met the inclusion criteria as random controls, in an up to 2:1 ratio with the asthma sample, also matching the distribution of diagnostic years.

For all included patients, we associated their Charlson’s Comorbidity Index (CCI),(21) all previous diagnoses and procedures using both three-digit and full-digit codes (selecting those with frequency >1%), and a set of sociodemographic and environmental (“social-ecological”) indices obtained from the American Community Survey (n=65 indices, including education, income, poverty, and foreign-born) (22) and U.S. Environmental Protection Agency (n=8 air quality indices, including median Air Quality Index, atmospheric levels of particulate matter 2.5, particulate matter 10, sulfur dioxide, nitrogen dioxide, ozone, and carbon monoxide) (23) corresponding to the zip-code of residence. Since patients usually had more than one visit and possibly more than one residence over the study period, the zip-code most frequently reported prior to baseline was used. All statistical analysis was conducted at the zip-code level, but results were also aggregated by county. Social-ecological indices were extracted from aforementioned sources for every year (2005-2014) and matched to patient zip-codes by year of diagnosis. Missing values in the social-ecological indices were imputed by finding the nearest neighbor(s) matched by zip-code and year of variable collection, prioritized first on time and then on location. For comparison, we also used multiple imputations with a time series cross-sectional model.(24)

To identify demographic, clinical, and sociol-ecological predictors of asthma and COPDAC, we used univariate analysis, feature ranking by bootstrapped information gain ratio (a measure of entropy reduction in a target variable achieved by learning the value of another one, closely related to the Kullback-Leibler divergence),(25) multivariable logistic regression with LogitBoost selection, decision trees, and random forests.(26) Model parameters (number of features, number of features randomly selected on each tree split, minimum number of instances per tree leaf, tree pruning, and number of trees) were optimized through a bootstrapped grid search. We analyzed the importance of each domain (i.e. demographic, clinical, socio-ecological) singularly and together, using the area under the receiver operating characteristic (AUROC), and sensitivity, and specificity. Variable importance in random forest was assessed using corrected impurity.(27) Out-of-bag data from bootstrapping was employed to assess model performance and generalizability on unseen data, using false discovery rate adjustment and a corrected t-test for model comparison.(28) The programming language java (29) and the statistical suite R (30), including libraries *Amelia, ranger, ROCR,* and *RWeka* were used to process data and perform all analyses. R libraries *choroplethr, choroplethrMaps,* and *noncensus* were used for the spatial mapping.

Results

Population Characteristics

The 2005-2014 Florida HCUP SID/SASD/SEDD contained a total of 21,091,289 patients after dataset linkage and cleaning. There were 141,729 patients meeting the inclusion, of whom 56,052 were diagnosed with asthma and 85,677 with COPDAC (without asthma). There were 84,737 patients with neither asthma nor COPDAC (random controls). Of note, a perfect 2:1 ratio of controls to asthma could not be achieved due to sequential exclusion constraints. For the included years 2005-2014, there was an average of 11,210 asthma cases each year (standard deviation (SD) 2,489). The raw prevalence of asthma in this population (using the Florida 2010 census population greater than 18 years of age as the denominator) was 95.0 per 100,000 (SD 21.9 per 100,000). The peak number of cases was observed in 2012 (n=13,291).

Table 1 shows population characteristics by outcome group. Compared to COPDAC and controls, the asthma group had a higher proportion of females (71.5%) and Hispanics (14.4%), as well as the youngest median age (43 years). The COPDAC patients had a larger proportion of white patients (80.7%) and a higher median CCI (2) compared to asthma and controls. Control patients had the highest proportion of African Americans (30.4%). There were notable differences in mortality. During the follow-up time since diagnosis, 9.9% of COPDAC patients died as compared to 2.8% of patients in the asthma group and 0.4% in the controls. The Kaplan-Meier estimate yielded a four-year survival probability of 79.7% (79.1%-80.3%) for COPDAC, 94.0% (93.6%-94.4%) for asthma, and 95.7% (94.8%-96.6%) for control patients.

A total of 383 three-digit ICD-9 diagnostic codes, 536 full-digit ICD-9 diagnostic codes, and 106 procedure codes were found, all above 1% frequency. The most frequent diagnostic codes and procedures from prior clinic encounters are displayed in Figure 2. In the asthma group, the most common diagnoses prior to the baseline were chronic airway obstruction, not elsewhere classified (ICD-9: 496), and bronchitis, not specified as acute or chronic (ICD-9: 490). The most common diagnoses for the COPDAC group prior to baseline were chronic airway obstruction, not elsewhere classified (ICD-9: 496), and chronic bronchitis (ICD-9: 491). In the asthma population, the three most frequent diagnoses were symptoms involving respiratory system and other chest symptoms (ICD-9: 786), essential hypertension (ICD-9: 401), and other symptoms involving abdomen and pelvis (ICD-9: 789), while the three most frequent procedures were esophagogastroduodenoscopy with closed biopsy (ICD-9: 4516), application of splint (ICD-9 code 9354), and injection or infusion of other therapeutic or prophylactic substance (ICD-9: 9929).

Univariate Analysis

We subsequently performed univariate analysis to identify diagnoses, procedures, demographic, and other socio-ecological variables exhibiting the highest information gain ratio (IGR) in relation to a subsequent diagnosis of asthma versus COPDAC. In Table 2, we show variables in the clinical diagnoses and procedures domain with the highest IGR, along with bootstrapped frequency and mean values. Of note, when the HCUP data were merged with ecological variables from other sources, 2,670 observations were excluded due to inability to match zip-code to a Florida zip-code. Comparing prodromal clinical diagnoses for asthma versus COPDAC and control groups, diseases of the respiratory system and other disorders of female genital tract were more frequent in asthma, and diseases of arteries, arterioles, and capillaries were more frequent in COPDAC. When restricting the comparison to asthma versus COPDAC without controls, the diagnoses complications of pregnancy, childbirth, and the puerperium were more frequent in asthma. For clinical procedures, we observed procedures related to pregnancy complications were more frequent in asthma, and that procedures related to cardiovascular diseases were more frequent in COPDAC.

In order to assess the stability of the variable importance ranking through IGR, we repeated the univariate analysis using two other indices (calculated from univariate logistic regression) namely the Akaike Information Criterion (AIC) and the AUROC. For the AIC, results were similar to those obtained by using the IGR. For AUROC, clinical diagnoses related to diseases of the respiratory system (codes 460-519), symptoms, signs, and ill-defined conditions (codes 780-799), as well as injuries (codes 800-999), tended to be ranked higher.

We then analyzed socio-ecological prodromal correlates of asthma versus COPDAC, and random controls. The top-scoring variables in terms of IGR included air quality indices, such as particulate matter and carbon monoxide levels, and social indices related to income, age and gender. When limiting the comparison to asthma versus COPDAC without controls, the air quality indices did not show high IGR. Figure 3 shows the county-level 2010-2014 prevalence of asthma and COPDAC diagnoses, and air quality indices. From a geospatial perspective, there were “hotspots” where the prevalence of asthma as well as COPDAC was significantly higher than the state average, specifically in Wakulla, Franklin, Union and Bradford counties located in the northwest area of Florida.

Multivariable Analysis

We then fitted multivariable models on single and merged domains, using the top variables in terms of IGR for each domain (200 variables, optimized through a bootstrapped grid search). For a multi-domain analysis comparing asthma versus COPDAC and controls, the optimal random forest (500 trees, and a number of variables for each split equal to the inverse hyperbolic sine of the total) outperformed the boosted logistic regression (containing more than a hundred variables after feature selection) and the decision tree in terms of AUROC (Figure 4 A). However, for the comparison of asthma to COPDAC without controls, random forest, boosted logistic regression and decision tree all performed similarly (Figure 4 B), yielding average AUROCs of 83%, 84%, and 82%, respectively. The simplified tree had an AUROC of 81%. The differences in means were significant at the 0.01 level. The logistic regressor included 157 variables, while the simplified tree had a size of 25 nodes with 6 variables.

Figure 5 shows the discriminatory power of each domain and of domains together (for the single domain analyses, we used random forests). When distinguishing asthma versus COPDAC and controls, the domain with the highest discriminatory power was socio-ecological (AUROC 74%), followed by clinical diagnoses (AUROC 70%), demographics (AUROC 69%), clinical procedures (AUROC 61%), and CCI (AUROC 57%). Merging domains together improved by overall performance by 10% (AUROC 84%). When comparing asthma to COPDAC without controls, the highest performing domain was clinical diagnoses (AUROC 82%), followed by demographics (AUROC 78%), CCI (AUROC 73%), clinical procedures (AUROC 71%), and socio-ecological (AUROC 61%). For that comparison, merging domains only improved overall performance by 1% (AUROC 83%). As a sensitivity analysis, we removed all variables related to prior diagnoses of COPDAC, and the overall AUROC decreased slightly (3% or less). The most important prior diagnoses of COPDAC were: bronchitis, not specified as acute or chronic (490), and chronic airway obstruction, not elsewhere classified (496). Also, model performance was similar when changing the imputation method from nearest neighbor to time series cross sectional.

We then ranked the importance of variables from all the domains using random forests and corrected Gini impurity (Figure 6). The most important predictors for the asthma versus COPDAC and controls were primarily socio-ecological variables, while for asthma versus COPDAC without controls, demographic and clinical variables such as age, CCI, and prior clinical diagnoses, scored better.

The simplified decision trees for predicting asthma are displayed in Figure 7. However, for the simplified decision tree, its performance is inferior in distinguishing asthma versus COPDAC and controls, as well asthma versus COPDAC without controls (AUROC 71% and 81%, respectively, compared to the best performing models with AUROC 83% and 84%, respectively).

Discussion

This study presents a multi-domain, big data analysis of predictors of asthma versus COPDAC (with and without controls). Our novel approach required a comprehensive linkage of clinical and socio-ecological domains to add social and environmental variables related to diseases of the respiratory tract to the domains analyzed. Clinically, our findings have future utility in creating comprehensive risk prediction models incorporating data from the patient and their environment. Such models will have utility for clinicians and patients who wish to distinguish future asthma versus COPDAC, and design treatment and prevention strategies accordingly.

This study presents two main analyses: predicting asthma versus COPDAC and controls, and predicting asthma versus COPDAC without controls. For the latter, the multi-domain approach combining clinical, demographic, and socio-ecological data had superior performance (AUROC 84%, 10% higher than best single domain model). The latter comparison’s best performing model was clinical diagnoses (AUROC 82%), and merging other domains did not significantly improve the models performance. This is an interesting finding given that the most common prodromal diagnoses for asthma and COPDAC patients were fairly similar (asthma: chronic airway obstruction, not elsewhere classified, and bronchitis, not specified as acute or chronic; COPDAC: chronic airway obstruction, not elsewhere specified, and chronic bronchitis). However as also detailed in our results, those similar diagnoses may differentiate between the two respiratory ailments by virtue of being applied to the younger, more female and Hispanic asthma population versus the older, more male and white COPDAC population.

Further to this, our findings of lipid and cardiovascular disease more common to COPDAC patients likely reflects the increased risk of cardiovascular disease in patients with COPDAC.(31) Our findings of significantly lower survival rates in COPDAC patients may reflect that increased incidence of cardiovascular disease and COPDAC patient’s higher median age. The clinical utility of those results however, is confounded by the common lifestyle associations with both COPDAC and cardiovascular disease such as smoking, which may be involved in the pathogenesis of both conditions, and unfortunately was not a variable in this dataset.(32)

Similarly, the top clinical diagnoses and procedures that included pregnancy, childbirth, or complications thereof were more frequent in asthma. This likely reflects the high incidence of asthma exacerbations during pregnancy and the female predominance of the asthma sample.(33) Pregnant women with asthma have higher risk for hemorrhage and cesarean section (as reflected in our data).(33) The increase in asthma exacerbation frequency may be due to physiologic (particularly immunologic) changes in females during pregnancy. (34) It is unknown whether improved asthma diagnostic sensitivity and symptom control prior to conception helps prevent or decrease asthma-related pregnancy complications.(35) Notably, this study’s pregnancy-related findings were only significant only using the IGR method, and not the odds-ratio or the AUROC method.

Both the aforementioned findings for COPDAC and asthma largely reflect known characteristics of both diseases. It is unclear whether those characteristics are physiologically-linked or manifestations of clinicians’ bias and tendencies when treating patients (i.e., higher index of suspicion for COPDAC in elderly patients, or more vigilant treatment (and thereby diagnostic coding) of asthma in pregnancy). However, when utilizing big data for epidemiologic studies, such findings (whether causative by a physiologic mechanism or by a clinician’s association) can be useful in deriving predictive models.

With regard to socio-ecological variables, multivariate variable importance ranking by means of corrected Gini impurity obtained from random forests contained a number of top-scoring socio-ecological variables in both comparisons. Socio-ecological variables were the highest performing single domain in comparing asthma versus COPDAC and controls. With validation and construction of a future clinical prediction rule in mind, our findings that socio-ecological domains have a high discriminatory power is important. However, those environmental domains, particularly sociodemographic and air quality indices, contributed strongly to the data variance. This warrants a more thorough investigation of ecological determinants to identifying additional sources (e.g. integrating data from National Centers for Environmental Information),(36) and optimizing the geographic resolution (e.g. census block vs. zip-code vs. county level). Despite that, our findings on environmental predictors and air quality (particulate and carbon monoxide content) are immediately useful for clinicians and public health stakeholders in Florida and elsewhere in advocating for air quality improvements, particularly in the “hot spot” counties identified in this study.

While our models have significant predictive value, they all contain a large number of variables. Some have suggested a maximum of 10 variables for use in clinical practice.(37) However, a review of operative clinical decision models found a median of 27 variables.(38) Therefore currently there is no ‘simple score’ which predicts asthma in the general population with high sensitivity and specificity. The logistic main effects model has good performance (79% AUROC) but also utilizes a high number of variables (even after feature shrinkage). Further to this, the logistic main effects model is superseded by the decision tree and random forest, both nonlinear methods. Even the best decision tree had a high number of nodes that do not facilitate quick clinical interpretation. Overall, prediction of asthma versus COPDAC arises from a multiplicity of factors that are difficult to pare down in a simplified model at this time. Therefore future work should focus on integrating multi-domain data into a simplified, clinically usable tool that can be prospectively validated.

Limitations

This work has limitations which merit discussion. It is a retrospective analysis of an observational dataset from a single state, although Florida is the third most populated state in the US. (17) HCUP combines race and ethnicity into one variable, which limited our demographic analysis. Given that this was a large retrospective database, there is no way to definitively confirm whether the included patients had asthma, COPDAC, or neither asthma nor COPDAC. We chose to use ICD codes as a proxy for those diagnoses (and lack of diagnoses), even though ICD codes are not a perfect proxy for asthma and COPDAC. However, studies have demonstrated good correlation of ICD codes with patient questionnaires,(39) and that a confirmed code improves accuracy. Therefore, we chose to twice confirm the asthma diagnosis for patients included in the asthma group. The HCUP database does not include questionnaire data (such as smoking history or other atopy), physiologic data such as BMI, pharmacy data, laboratory data, or results of spirometry tests. Since some of the strongest predictors of asthma are lung function measurements and diagnoses at a younger age,(40,41) this is a shortcoming of using that dataset. We analyzed only ten years of Florida HCUP data, which may have excluded patients only seen earlier in the study timeframe and/or patients who do not regularly seek medical care. Additionally, apart from survival, we did not examine other patient outcomes as the objective of the study was to look at predictors of asthma versus COPDAC.

Our choice of the univariate feature scoring method used may have produced very different rankings (e.g. IGR vs. AUROC). However, we included over a third of the original variable space in the multivariable analysis, and the multivariable models' performance were not affected. However, the variable importance ranking obtained from the random forest showed different rankings than the univariate IGR, which warrants further investigation on the intra-domain and inter-domain variable mechanics behind that difference.

With regard to specific domain analysis, we included a specific set of socio-ecological indices and made broad assumptions on the indices' characteristics over time and space. For instance, for missing value imputation, the nearest neighbor method for air quality indices may not be the best choice given the sparsely located air monitors in Florida. The usage of a single pivot year to set socio-ecological indices may be acceptable for socio-economic factors, since they are relatively stable over time, but may introduce misclassifications for other indices such as air pollution given their large temporal variations. Also, due to the lack of residential mobility data and use of a single zip-code over all encounters, we may have introduced non-differential misclassification.(42)

Conclusion

In this large statewide study which linked clinical and socio-ecological domains, we found prodromal correlates of future asthma (female gender and pregnancy-related complications) versus COPDAC (cardiovascular disease). We also found that combining multiple (as opposed to single) domains does yield more sensitive and specific prediction models for asthma versus COPDAC and controls, but that clinical diagnoses as a single domain was the best model for asthma versus COPDAC without controls. However, the large number of variables in our simplified decision trees precludes immediate clinical application of those models. In order to predict and prevent asthma and its significant and costly complications, future work should focus on validating our findings beyond Florida, further exploring socio-ecological domain variables, and creating a streamlined prediction tool for clinicians.

Acknowledgments

Declaration of interest: This work was supported by the grant 5UL1TR001427-04 from NIH-NCATS, by the University of Florida (UF) One Health Center, and by UF “Creating the Healthiest Generation” Moonshot initiative, which is supported by the UF Office of the Provost, UF Office of Research, UF Health, UF College of Medicine and UF Clinical and Translational Science Institute. The authors declared there is no conflict of interests.

References

1. Vos T, Allen C, Arora M, Barber RM, Brown A, Carter A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1545–602. [PubMed: 27733282]
2. Nurmagambetov T, Kuwahara R, Garbe P. The economic burden of asthma in the United States, 2008-2013. Ann Am Thorac Soc. 2018;15(3):348–56. [PubMed: 29323930]
3. Steppuhn H, Langen U, Keil T, Scheidt-Nave C. Chronic disease co-morbidity of asthma and unscheduled asthma care among adults: results of the national telephone health interview survey German Health Update (GEDA) 2009 and 2010. Nat Publ Gr. 2014;23(1):22–9.
4. Sumino K, Bartle B, Au DH, Castro M, Lee TA, Author C. Coexisting chronic conditions associated with mortality and morbidity in adult patients with asthma. J Asthma. 2014;51(3):306–14. [PubMed: 24432868]
5. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. Current Allergy and Asthma Reports. 2015.
6. Svenningsen S, Nair P. Asthma Endotypes and an Overview of Targeted Therapy for Asthma. Front Med [Internet]. 2017 [cited 2017 Nov 15];4 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5622943/pdf/fmed-04-00158.pdf
7. Lötvall J, Akdis CA, Bacharier LB, Bjermer L, Casale TB, Custovic A, et al. Asthma endotypes: A new approach to classification of disease entities within the asthma syndrome. J Allergy Clin Immunol [Internet]. 2011 [cited 2018 Aug 10];127(2):355–60. Available from: https://ac.els-cdn.com/S0091674910018580/1-s2.0-S0091674910018580-main.pdf?_tid=ab126da7-1bbf-457c-a26d-8382f6e46dd2&acdnat=1533929191_58b62de1e5c8ad94c90a63df9eba5c42 [PubMed: 21281866]
8. Leung JM, Sin DD. Asthma-COPD overlap syndrome: pathogenesis, clinical features, and therapeutic targets. Bmj [Internet]. 2017 [cited 2018 Aug 10];358:j3772 Available from: http://www.bmj.com/lookup/doi/10.1136/bmj.j3772 [PubMed: 28947632]
9. Postma DS, Rabe KF. The Asthma–COPD Overlap Syndrome. N Engl J Med [Internet]. 2015 [cited 2018 Aug 10];373(13): 1241–9. Available from: http://www.nejm.org/doi/10.1056/NEJMra1411863 [PubMed: 26398072]
10. Löwhagen O Diagnosis of asthma - New theories. J Asthma [Internet]. 2015 [cited 2018 Aug 10];52(6):538–44. Available from: http://www.tandfonline.com/action/journalInformation?journalCode=ijas20http://informahealthcare.com/jas [PubMed: 25478896]
11. Wildfire JJ, Gergen PJ, Sorkness CA, Mitchell HE, Calatroni A, Kattan M, et al. Development and validation of the composite asthma severity index - An outcome measure for use in children and adolescents. J Allergy Clin Immunol [Internet]. 2012 [cited 2018 Aug 10];129(3):694–701.

Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3294274/pdf/nihms350008.pdf [PubMed: 22244599]

12. Bateman ED, Buhl R, O'Byrne PM, Humbert M, Reddel HK, Sears MR, et al. Development and validation of a novel risk score for asthma exacerbations: The risk score for exacerbations. J Allergy Clin Immunol. 2015;135(6):1457–1464e4. [PubMed: 25258144]
13. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet [Internet]. 2011 [cited 2018 Aug 11];43(9):887–92. Available from: http://www.nature.com/authors/editorial_policies/license.html#terms [PubMed: 21804549]
14. Zhu Z, Lee PH, Chaffin MD, Chung W, Loh PR, Lu Q, et al. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat Genet [Internet]. 2018 [cited 2018 Aug 11];50(6):857–64. Available from: www.nature.com/naturegenetics [PubMed: 29785011]
15. Sinharoy A, Mitra S, Mondal P. Socioeconomic and Environmental Predictors of Asthma-Related Mortality. J Environ Public Health [Internet]. 2018 [cited 2018 Aug 11];2018 Available from: 10.1155/2018/9389570
16. Prosperi MC, Marinho S, Simpson A, Custovic A, Buchan IE. Predicting phenotypes of asthma and eczema with machine learning. BMC Med Genomics. 2014;7(SUPPL.1).
17. U.S. Census Bureau QuickFacts: Florida; UNITED STATES [Internet]. [cited 2019 Jan 2]. Available from: https://www.census.gov/quickfacts/fact/table/fl,US/PST045217
18. HCUP-US Home Page [Internet]. [cited 2019 Jan 2]. Available from: https://www.hcup-us.ahrq.gov/
19. Slee VN. The International Classification of Diseases: ninth revision (ICD-9). Annals of Internal Medicine. 1978;88(3):424–6. [PubMed: 629506]
20. Cooke CR, Joo MJ, Anderson SM, Lee TA, Udris EM, Johnson E, et al. The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. BMC Health Serv Res. 2011;11.
21. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J-C, et al. Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data. Med Care [Internet]. 2005;43(11): 1130–9. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00005650-200511000-00010 [PubMed: 16224307]
22. American Community Survey (ACS) [Internet]. [cited 2019 Jan 3]. Available from: https://www.census.gov/programs-surveys/acs/
23. Air Data: Air Quality Data Collected at Outdoor Monitors Across the US [Internet]. [cited 2019 Jan 3]. Available from: https://www.epa.gov/outdoor-air-quality-data
24. Honaker J, King G. What to do about missing values in time-series cross-section data. Am J Pol Sci. 2010;
25. Li L, Zhang X, Xue M. Explaining information gain and information gain ratio in information theory. ICIC Express Lett. 2013;7(8):2385–91.
26. Rodriguez-Perez M, Gonzalez-Dominguez J, Mataran L, Garcia-Perez S, Salvatierra D. Association of HLA-DR5 with mucocutaneous lesions in patients with rheumatoid arthritis receiving gold sodium thiomalate. J Rheumatol. 1994;21(1):41–3. [PubMed: 8151585]
27. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711–8. [PubMed: 29757357]
28. Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn. 2003;52(3):239–81.
29. java.com: Java + You [Internet]. [cited 2019 Jan 2]. Available from: https://java.com/en/
30. R: The R Project for Statistical Computing [Internet]. [cited 2019 Jan 2]. Available from: https://www.r-project.org/
31. Ghoorah K, De Soyza A, Kunadian V. Increased cardiovascular risk in patients with chronic obstructive pulmonary disease and the potential mechanisms linking the two conditions: A review. Cardiology in Review. 2013.
32. Bhatt SP, Dransfield MT. Chronic obstructive pulmonary disease and cardiovascular disease. Translational Research. 2013.

33. Ali Z, Hansen AV, Ulrik CS. Exacerbations of asthma during pregnancy: Impact on pregnancy complications and outcome. Journal of Obstetrics and Gynaecology. 2016.

34. Murphy VE, Schatz M. Asthma in pregnancy: A hit for two. Eur Respir Rev. 2014;

35. Wang G, Murphy VE, Namazy J, Powell H, Schatz M, Chambers C, et al. The risk of maternal and placental complications in pregnant women with asthma: A systematic review and meta-analysis. J Matern Neonatal Med. 2014;

36. NCEI. National Centers for Environmental Information [Internet]. 2015 [cited 2019 Jan 2]. p. 1–5. Available from: https://www.ncei.noaa.gov/access

37. Lee Y, Bang H, Kim DJ. How to Establish Clinical Prediction Models. Endocrinol Metab. 2016;

38. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. J Am Med Informatics Assoc. 2017;

39. Yang CL, To T, Foty RG, Stieb DM, Dell SD. Verifying a questionnaire diagnosis of asthma in children using health claims data. BMC Pulm Med. 2011;11.

40. Bui DS, Lodge CJ, Burgess JA, Lowe AJ, Perret J, Bui MQ, et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. Lancet Respir Med. 2018;

41. Owens L, Laing IA, Zhang G, Le Souëf PN. Infant lung function predicts asthma persistence and remission in young adults. Respirology. 2017;

42. Pennington AF, Strickland MJ, Klein M, Zhai X, Russell AG, Hansen C, et al. Measurement error in mobile source air pollution exposure estimates due to residential mobility during pregnancy. J Expo Sci Environ Epidemiol. 2017;

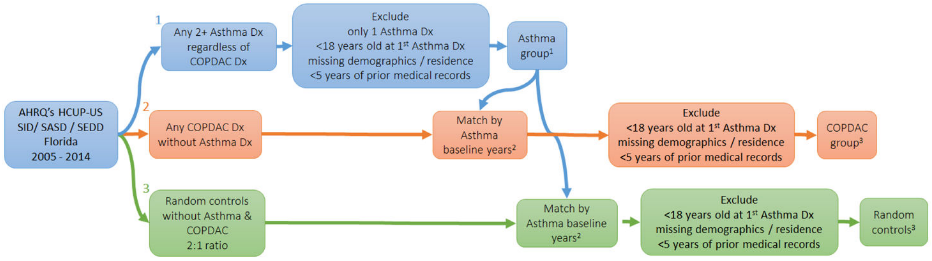
AHRQ's HCUP-US
SID/ SASD / SEDD
Florida
2005 - 2014
1
Any 2+ Asthma Dx regardless of COPDAC Dx
Exclude
only 1 Asthma Dx
<18 years old at 1st Asthma Dx
missing demographics / residence
<5 years of prior medical records
Asthma group1
2
Any COPDAC Dx without Asthma Dx
Match by Asthma baseline years2
Exclude
<18 years old at 1st Asthma Dx
missing demographics / residence
<5 years of prior medical records
COPDAC group3
3
Random controls without Asthma & COPDAC 2:1 ratio
Match by Asthma baseline years2
Exclude
<18 years old at 1st Asthma Dx
missing demographics / residence
<5 years of prior medical records
Random controls3

Figure 1.

Flowchart of data collection according to patient selection criteria, baseline times, and case matching.

[1]Baseline year is the year of first Asthma Dx (2005-2014).

[2]Using a stochastic algorithm.

[3]Baseline year is a random year according to Asthma baseline year distribution match.

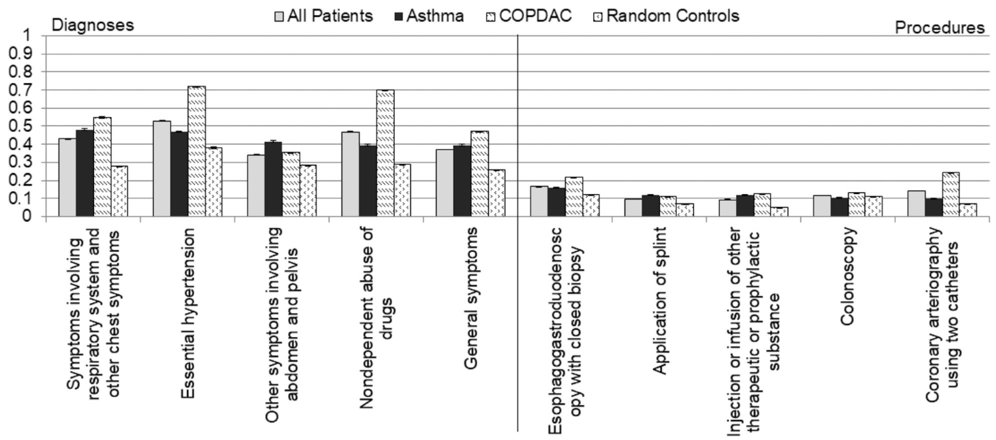

Figure 2.

Top-5 most frequent (A) diagnostic and (B) procedural codes called in the asthma, COPDAC, and random sample (non-asthma, non-COPDAC) populations.

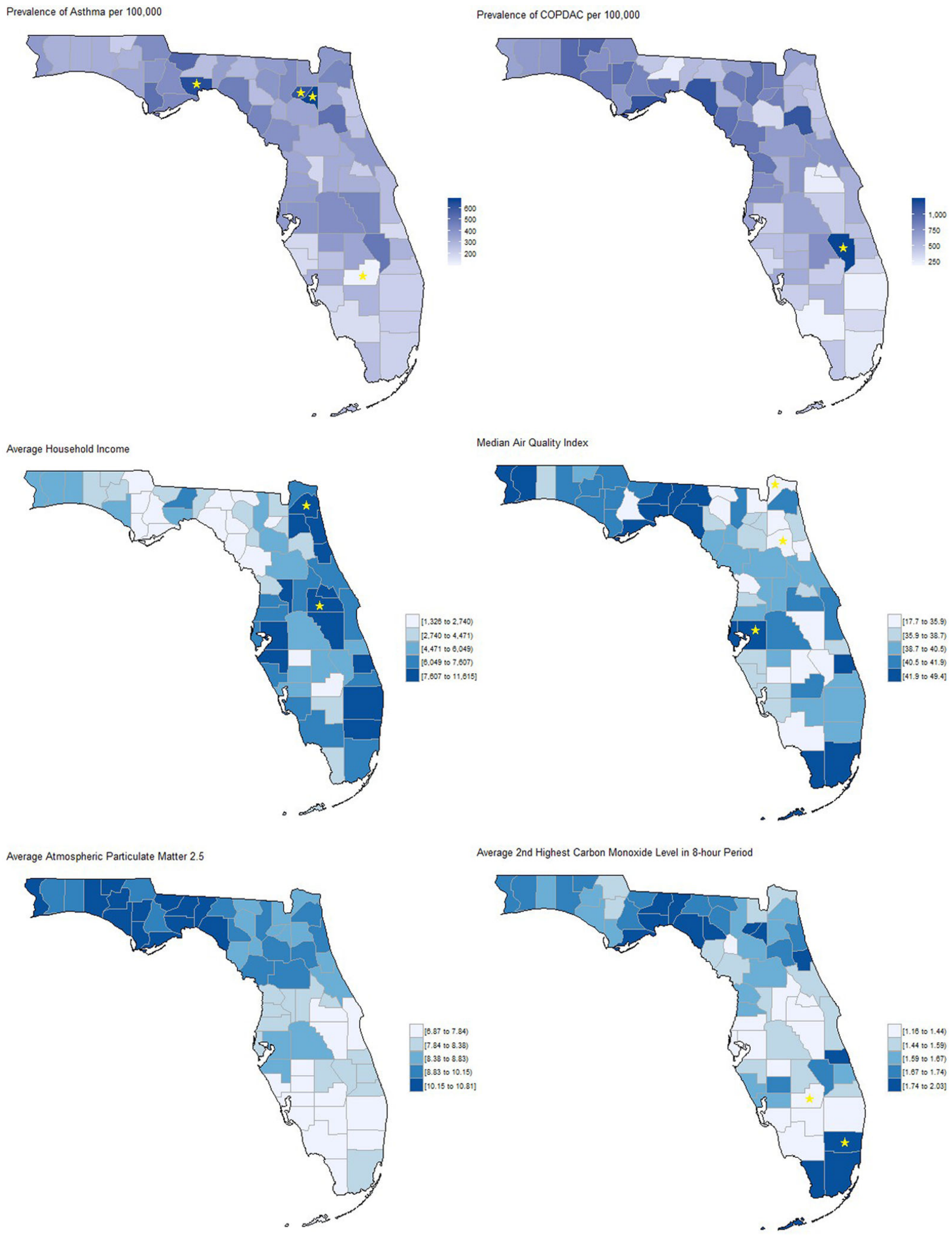

Figure 3.

Geographic distribution in Florida (county-level) between 2010 and 2014 of: (**A**) asthma prevalence; (**B**) COPDAC prevalence; (**C**) average household income; (**D**) median air quality index; (**E**) average particulate matter (diameter of less than 2.5 micrometers); (**F**) average 2nd highest carbon monoxide level in a 8-hour period by county. ★ indicate counties with values at least two standard deviations from the State mean.

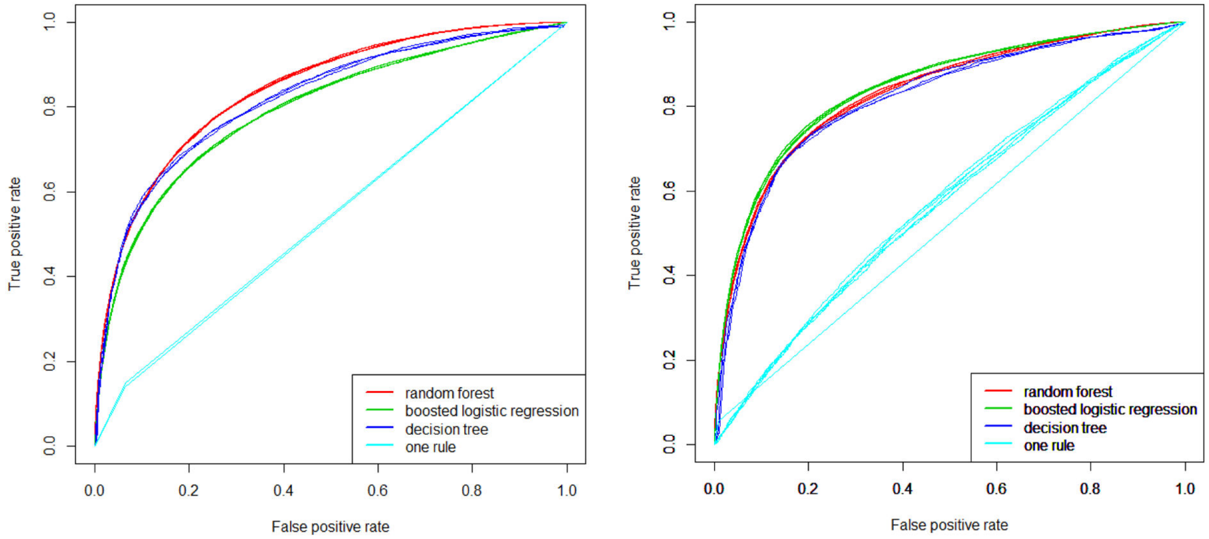

Figure 4.

Model selection using random forests, boosted logistic regression, and decision tree for: (**A**) asthma vs. COPDAC and random controls; (**B**) asthma vs. COPDAC. Results are averaged from 10 out-of-bag predictions over the corresponding bootstrap training sets.

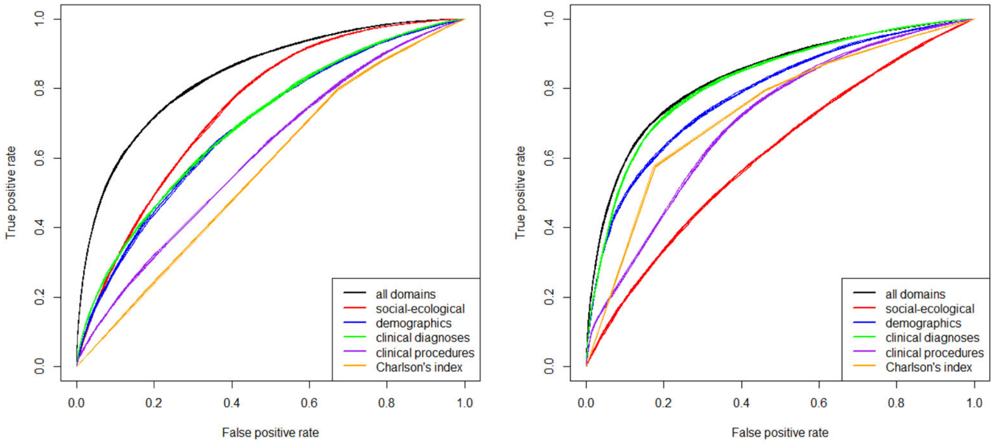

Figure 5.

Comparison of predictive power of domains independently and together, using random forests for: (**A**) asthma vs. COPDAC and random controls; (**B**) asthma vs. COPDAC. Results are averaged from 10 out-of-bag predictions over the corresponding bootstrap training sets.

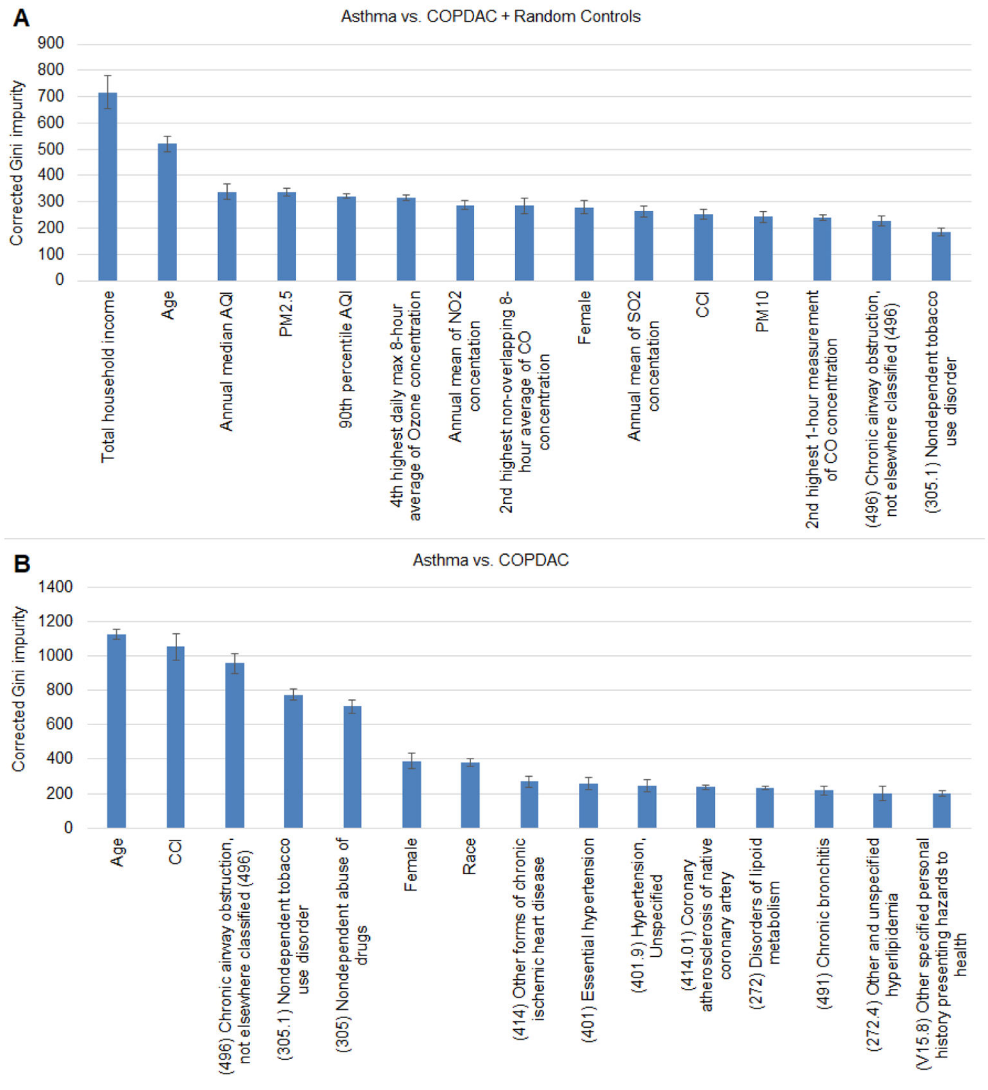

Figure 6.

Multivariate variable importance ranking by means of corrected Gini impurity obtained from random forests. Top panel (**A**) shows Asthma vs. COPDAC + random controls, bottom panel (**B**) shows Asthma vs. COPDAC. Plots show average and standard deviation of the most important 15 variables estimated from 10 bootstrap runs.

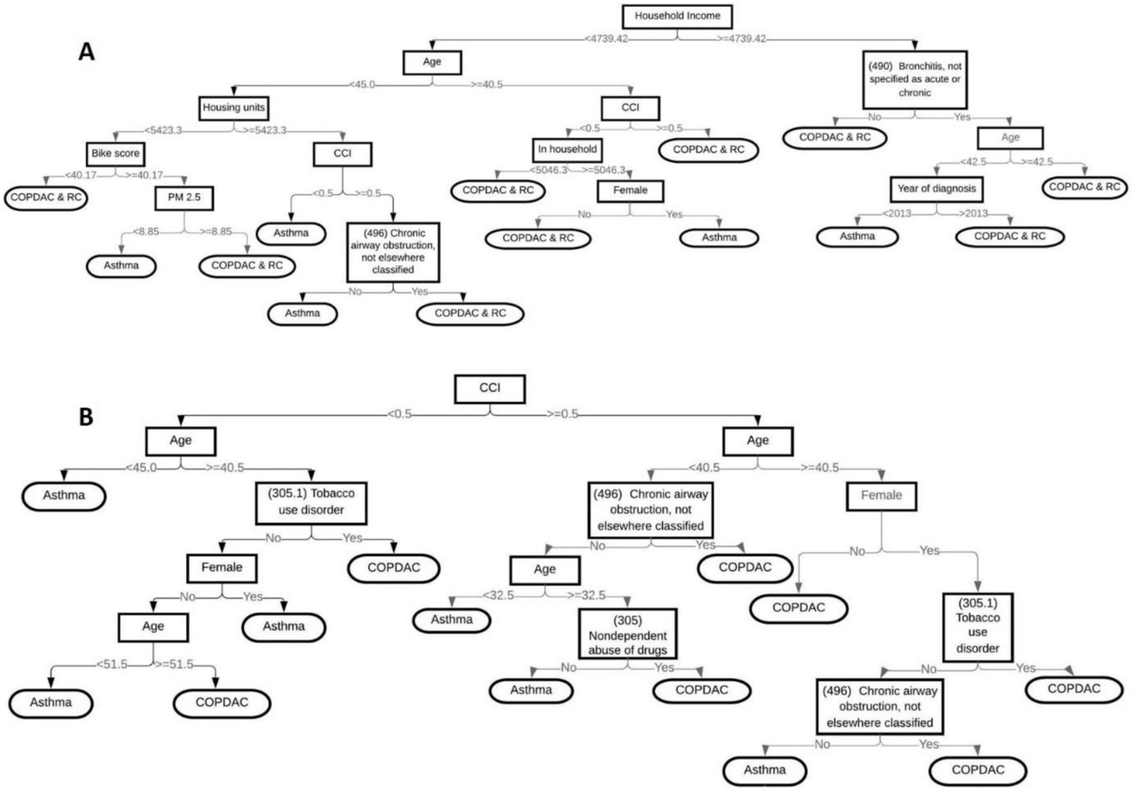

Figure 7.

Simplified decision tree models to distinguish (**A**) Asthma vs. COPDAC and random controls or (**B**) Asthma vs. COPDAC.

Table 1.

Population characteristics.

	Asthma	COPDAC (no asthma)	Random controls (no asthma, no COPDAC)
Total N	56052	85677	84737
Female N (%)	40079 (71.5%)	39644 (46.3%)	48034 (56.7%)
Race White N (%)	33156 (59.2%)	69130 (80.7%)	49188 (58%)
Race African American N (%)	13357 (23.8%)	9725 (11.4%)	25779 (30.4%)
Ethnicity Hispanic N (%)	8058 (14.4%)	5565 (6.5%)	6680 (7.9%)
Race Asian or Pacific Islander N (%)	276 (0.5%)	212 (0.2%)	744 (0.9%)
Race Native American N (%)	71 (0.1%)	91 (0.1%)	107 (0.1%)
Race Other N (%)	1134 (2%)	954 (1.1%)	2239 (2.6%)
Age median (interquartile range, IQR)	43 (30-53)	54 (47-60)	45 (31-56)
Year of asthma/COPDAC diagnosis (or baseline) median (IQR)	2012 (2011-2013)		
Years of prior medical history available in HCUP median (IQR)	6 (5-7)	6 (5-7)	6 (5-7)
Charlson’s comorbidity index median (IQR)	0 (0-1)	2 (1-4)	0 (0-0)
Bronchitis, not specified as acute or chronic (ICD-9 code 490)	7981 (14.2%)	10984 (12.8%)	N/A
Chronic bronchitis (ICD-9 code 491)	5773 (10.3%)	21371 (24.9%)	N/A
Emphysema (ICD-9 code 492)	1293 (2.3%)	7794 (9.1%)	N/A
Chronic airway obstruction, not elsewhere classified (ICD-9 code 496)	8467 (15.1%)	43174 (50.4%)	N/A
Obstructive chronic bronchitis without exacerbation (ICD-9 code 491.20)	866 (1.5%)	3230 (3.8%)	N/A
Obstructive chronic bronchitis with (acute) exacerbation (ICD-9 code 491.21)	4117 (7.3%)	15667 (18.3%)	N/A
Obstructive chronic bronchitis with acute bronchitis (ICD-9 code 491.22)	1719 (3.1%)	5111 (6%)	N/A
Other emphysema (ICD-9 code 492.8)	1236 (2.2%)	7397 (8.6%)	N/A

Table 2.

Top-10 variables on each domain (clinical diagnoses, procedures, and socio-ecological) showing the highest information gain ratio (IGR) with respect to the asthma diagnosis (univariate analysis, indices estimated through 10-fold replicates).

Asthma vs. COPDAC + Random Controls				
Clinical diagnoses (ICD-9 code)	**IGR mean (st.dev)**	**Asthma % (st.err)**	**COPDAC % (st.err)**	**Rand. Contr. % (st.err)**
Bronchitis, not specified as acute or chronic (490)	0.023 (0.0006)	14.3% (0.1)	12.9% (0.1)	NA
Noninflammatory disorders of ovary fallopian tube and broad ligament (620)	0.019 (0.0013)	8% (0.1)	2.6% (0.1)	4.6% (0.1)
Malignant neoplasm of trachea bronchus and lung (162)	0.019 (0.0007)	0.6% (0.03)	4.2% (0.1)	0.1% (0.01)
Acute bronchitis and bronchiolitis (466)	0.017 (0.0003)	16.8% (0.2)	13.2% (0.1)	4.1% (0.1)
Noninflammatory disorders of vagina (623)	0.017 (0.0018)	5.8% (0.1)	1.9% (0.05)	3.3% (0.1)
Acute pharyngitis (462)	0.017 (0.0009)	12.2% (0.1)	6.4% (0.1)	5.7% (0.1)
Acute upper respiratory infections of multiple or unspecified sites (465)	0.016 (0.0009)	13.1% (0.1)	8.5% (0.1)	4.9% (0.1)
Disorders of menstruation and other abnormal bleeding from female genital tract (626)	0.016 (0.0004)	8.1% (0.1)	2.5% (0.1)	5.4% (0.1)
Aortic aneurysm and dissection (441)	0.016 (0.001)	0.5% (0.03)	3.1% (0.1)	0.3% (0.02)
Atherosclerosis (440)	0.015 (0.0003)	2% (0.1)	9.3% (0.1)	1.3% (0.04)
Clinical procedures (ICD-9 code)	**IGR mean (st.dev)**	**Asthma % (st.err)**	**COPDAC % (st.err)**	**Rand. Contr. % (st.err)**
Other Fetal Monitoring (7534)	0.018 (0.0022)	1.9% (0.1)	0.3% (0.02)	1.2% (0.04)
Low Cervical Cesarean Section (741)	0.013 (0.001)	4.7% (0.1)	0.7% (0.03)	3.9% (0.1)
Angioplasty Of Other Non-Coronary Vessel (3950)	0.012 (0.0009)	1.2% (0.05)	5.4% (0.1)	0.9% (0.03)
Insertion Of Non-Drug-Eluting, Non-Coronary Artery Stent(s) (3990)	0.012 (0.0009)	0.7% (0.04)	3.6% (0.1)	0.4% (0.02)
Procedure On Single Vessel (0040)	0.012 (0.0006)	3.2% (0.1)	11.5% (0.1)	2.6% (0.1)
Arteriography Of Femoral And Other Lower Extremity Arteries (8848)	0.011 (0.0008)	1.6% (0.1)	6.3% (0.1)	1% (0.03)
Procedure On Two Vessels (0041)	0.011 (0.0007)	0.9% (0.04)	3.5% (0.1)	0.7% (0.03)
Percutaneous [Endoscopic] Gastrostomy (4311)	0.01 (0.0007)	0.5% (0.03)	2.1% (0.05)	0.3% (0.02)
Aortography (8842)	0.01 (0.0006)	2% (0.1)	7.6% (0.1)	1.3% (0.04)
Percutaneous Transluminal Coronary Angioplasty (0066)	0.01 (0.0004)	2.8% (0.1)	9.2% (0.1)	2.3% (0.1)
Asthma vs. COPDAC				
Clinical diagnoses (ICD-9 code)	**IGR mean (st.dev)**	**Asthma % (st.err)**	**COPDAC % (st.err)**	NA
Chronic airway obstruction, not elsewhere classified (496)	0.106 (0.0008)	15.1% (0.2)	50.4% (0.2)	
Outcome of delivery, single liveborn (V270)	0.099 (0.0025)	10.3% (0.1)	1.5% (0.04)	
Other current conditions in the mother classifiable elsewhere but complicating pregnancy childbirth or the puerperium (648)	0.088 (0.0021)	8.7% (0.1)	1.4% (0.04)	
Other complications of pregnancy not elsewhere classified (646)	0.077 (0.0017)	5.4% (0.1)	0.9% (0.03)	
Trauma to perineum and vulva during delivery (664)	0.077 (0.0043)	2.7% (0.1)	0.4% (0.02)	
Abnormality of organs and soft tissues of pelvis (654)	0.075 (0.0042)	3.6% (0.1)	0.5% (0.03)	
Hemorrhage in early pregnancy (640)	0.075 (0.002)	4% (0.1)	0.6% (0.03)	

Umbilical cord complications during labor and delivery (663)	0.074 (0.0045)	2.9% (0.1)	0.4% (0.02)	
Excessive vomiting in pregnancy (643)	0.07 (0.0042)	2.3% (0.1)	0.3% (0.02)	
Early or threatened labor (644)	0.07 (0.0025)	3% (0.1)	0.5% (0.02)	
Clinical procedures (ICD-9 code)	**IGR mean (st.dev)**	**Asthma % (st.err)**	**COPDAC % (st.err)**	
Other Manually Assisted Delivery (7359)	0.079 (0.0034)	5.3% (0.1)	0.8% (0.03)	
Repair Of Other Current Obstetric Laceration (7569)	0.076 (0.0057)	2.5% (0.1)	0.3% (0.02)	
Low Cervical Cesarean Section (741)	0.075 (0.0031)	4.7% (0.1)	0.7% (0.03)	
Other Artificial Rupture Of Membranes (7309)	0.07 (0.0034)	2.3% (0.1)	0.3% (0.02)	
Other Fetal Monitoring (7534)	0.067 (0.0032)	1.9% (0.1)	0.3% (0.02)	
Medical Induction Of Labor (734)	0.063 (0.0026)	2.2% (0.1)	0.3% (0.02)	
Procedure On Single Vessel (0040)	0.043 (0.0021)	3.2% (0.1)	11.5% (0.1)	
Angioplasty Of Other Non-Coronary Vessel (3950)	0.042 (0.0011)	1.2% (0.05)	5.4% (0.1)	
Aortography (8842)	0.04 (0.0014)	2% (0.1)	7.6% (0.1)	
Arteriography Of Femoral And Other Lower Extremity Arteries (8848)	0.04 (0.0006)	1.6% (0.1)	6.3% (0.1)	

Table 3.

Model performance

Data set	Model	Domain(s)	AUROC average (st.dev)	Sensitivity average (st.dev)	Specificity average (st.dev)	Cutoff average (st.dev)
Asthma vs. COPDAC + random controls	Boosted logistic regression	All	0.79 (0.002)	0.7 (0.004)	0.75 (0.002)	0.26 (0.001)
	Optimized decision tree		0.82 (0.003)	0.72 (0.009)	0.77 (0.011)	0.2 (0.014)
	Simplified decision tree		0.71 (0.002)	0.66 (0.001)	0.63 (0.003)	0.19 (0.002)
	One rule		0.54 (0.002)	0.14 (0.005)	0.93 (0.0002)	0.42 (0.005)
	Random forest	All	0.84 (0.0004)	0.75 (0.009)	0.77 (0.010)	0.27 (0.003)
		Clinical diagnoses	0.70 (0.0009)	0.63 (0.006)	0.65 (0.007)	0.26 (0.002)
		Clinical procedures	0.61 (0.002)	0.6 (0.003)	0.55 (0.003)	0.27 (0.001)
		Demographics	0.69 (0.003)	0.64 (0.005)	0.64 (0.002)	0.25 (0.003)
		Charlson's comorbidity index	0.57 (0.002)	0.8 (0.002)	0.33 (0.001)	0.27 (0.031)
		Social-ecological	0.74 (0.001)	0.72 (0.008)	0.64 (0.009)	0.27 (0.003)
Asthma vs. COPDAC	Boosted logistic regression	All	0.84 (0.002)	0.76 (0.009)	0.79 (0.007)	0.4 (0.01)
	Optimized decision tree		0.82 (0.003)	0.74 (0.004)	0.78 (0.011)	0.32 (0.021)
	Simplified decision tree		0.81 (0.001)	0.72 (0.013)	0.8 (0.016)	0.37 (0.069)
	One rule		0.56 (0.029)	0.5 (0.134)	0.61 (0.116)	0.42 (0.111)
	Random forest	All	0.83 (0.002)	0.74 (0.008)	0.79 (0.007)	0.39 (0.006)
		Clinical diagnoses	0.82 (0.001)	0.74 (0.003)	0.78 (0.003)	0.36 (0.003)
		Clinical procedures	0.71 (0.002)	0.68 (0.009)	0.64 (0.009)	0.43 (0.014)
		Demographics	0.78 (0.002)	0.69 (0.007)	0.74 (0.009)	0.36 (0.005)
		Charlson's comorbidity index	0.73 (0.001)	0.58 (0.002)	0.82 (0.002)	0.68 (0.002)
		Social-ecological	0.61 (0.002)	0.57 (0.01)	0.59 (0.012)	0.39 (0.004)