Predicting Olfactory Loss In Chronic Rhinosinusitis Using Machine Learning

Vijay R. Ramakrishnan, MD[1], Jaron Arbet, PhD[2], Jess C. Mace, MPH, CCRP[3], Stephanie Shintani Smith, MD, MS,[4] Zachary M. Soler, MD, MSc[5], Timothy L. Smith, MD, MPH[3]

Department of Otolaryngology-Head and Neck Surgery, University of Colorado, Aurora, CO.[1]

Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado-Denver Anschutz Medical Campus, Aurora, CO.[2]

Department of Otolaryngology-Head and Neck Surgery, Oregon Health Sciences University, Portland, OR.[3]

Department of Otolaryngology-Head and Neck Surgery, Northwestern University, Chicago, IL.[4]

Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, Charleston, SC.[5]

Corresponding author:
Vijay R. Ramakrishnan, MD
Department of Otolaryngology, University of Colorado
12631 E 17th Ave, B205
Aurora, CO. 80045
Phone: 303-724-1965
Fax: 303-724-1961
E-mail: Vijay.Ramakrishnan@ucdenver.edu

Running Title: Machine Learning and CRS Olfactory Loss

Financial Disclosures: V.R.R., J.C.M., Z.M.S., T.L.S., and S.S.S. are supported by grants for this investigation from the National Institute on Deafness and Other Communication Disorders (NIDCD) and the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health, Bethesda, MD., USA [R01 DC005805 (T.L.S.), K23 DC014747 (V.R.R.), 1P01AI145818-01 (S.S.S.)]. Public clinical trial registration (www.clinicaltrials.gov) ID# NCT01332136. Contents are the authors' sole responsibility and do not necessarily represent official NIH views.

Conflicts of Interest: None related to this study

ABSTRACT

Objective

Compare machine learning (ML) based predictive analytics methods to traditional logistic regression in classification of olfactory dysfunction in chronic rhinosinusitis (CRS-OD), and identify predictors within a large multi-institutional cohort of refractory CRS patients.

Methods

Adult CRS patients enrolled in a prospective, multi-institutional, observational cohort study were assessed for baseline CRS-OD using a smell identification test (SIT) or brief SIT (bSIT). Four different ML methods were compared to traditional logistic regression for classification of CRS normosmics versus CRS-OD.

Results

Data were collected for 611 study participants who met inclusion criteria between April 2011 and July 2015. 34% of enrolled patients demonstrated olfactory loss on psychophysical testing. Differences between CRS normosmics and those with smell loss included objective disease measures (CT and endoscopy scores), age, sex, prior surgeries, socioeconomic status, steroid use, polyp presence, asthma, and aspirin sensitivity. Most ML methods outperformed traditional logistic regression in terms of predictive ability. Top predictors include factors previously reported in the literature, as well as several socioeconomic factors.

Conclusion

Olfactory dysfunction is a variable phenomenon in CRS patients. ML methods outperform traditional logistic regression in classification of normosmia versus smell loss in CRS, and are able to include numerous risk factors into prediction models. These results carry implications for basic science and clinical research in hyposmia secondary to sinonasal disease, the most common cause of persistent olfactory loss in the general population.

INTRODUCTION

Olfactory loss is common, affecting up to 25% of the adult population, with chronic rhinosinusitis (CRS) being a major cause of persistent olfactory loss [1]. Smell loss is a hallmark of CRS, particularly in those with nasal polyps, and is one of the cardinal diagnostic symptoms for the disease [2-4]. Curiously, this symptom does not affect all CRS patients. Just as CRS is a heterogeneous disease [5], previous literature has estimated a 20-80% prevalence of olfactory loss in CRS cohorts, based on the sensitivity of olfactory assessment used and the proportion of enrolled subjects with nasal polyps (CRSwNP). A recent study using well-defined CRS diagnostic criteria and Sniffin' Sticks objective testing showed a 44% prevalence of smell loss in CRSsNP subjects and 58% in CRSwNP [6]. Risk factors associated with CRS-OD identified in this study included disease severity score on computed tomography (CT) and the presence of comorbid conditions such as diabetes, allergy, and asthma. Adding a healthy control group for comparison, Schlosser et al. noted CRS-OD to affect CRSwNP patients more than CRSsNP, and identified age as an additional risk factor [7]. Given that most existing studies consist of single-institution cohorts, with subjects undergoing variable treatment regimens, and different assessments of smell function, a systematic review and meta-analysis was performed by Kohli et al. [8] to clarify the presence of OD in CRS and identify consistent risk factors. The authors included 47 studies in the analysis, and identified age, CT score, and presence of nasal polyps as the three most important risk factors for CRS-OD in the published literature. Similarly, olfactory response to CRS treatment is highly variable, with resolution of olfactory disturbance after standard therapies occurring in approximately 40% of refractory CRS patients [9].

Olfactory dysfunction in CRS is potentially complex and the interplay among possible mechanisms is poorly understood. The simple model of conductive loss due to impaired odorant transport to the olfactory cleft in CRS may explain a portion of the disease for some patients. Local and systemic inflammation likely contribute to varying degrees in particular individuals, with underlying factors such as age and presence of comorbid diabetes influencing the capacity for olfactory sensory neuron regeneration. Given the intricate and multifactorial nature of CRS-OD, compounded by inherent challenges of human cohort studies, it is no surprise that putative predictors and their relative importance have been inconsistently characterized across studies. Existing single-institution studies carry a significant risk of sampling bias and their use of univariate regression models are insufficient to identify predictor variables for CRS-OD. These analyses cannot account for all possible predictor variables or interactions among variables, and therefore, may offer limited insights. Additional attempts at categorizing outcomes with multivariable regression approaches that were not developed for predictive performance, or assessed as such, cannot be relied upon [10].

Traditional statistical approaches can model how system variables relate to one another and generate corresponding mathematical metrics of statistical significance. However, novel data analytics approaches including machine learning (ML) methods have shown improved classification accuracy and can be utilized to predict future outcomes, rather than focusing on individual components within a system. ML approaches are particularly useful when there is a role for including complex interactions that may otherwise be ignored or dismissed as noise when using traditional statistics [11], such as in the case of CRS-OD.

Understanding the heterogeneity of CRS-OD and interpreting multi-dimensional risk factors for outcomes prediction is foundational for future clinical care and research. The objective of this study was to test different ML-based predictive analytics approaches against traditional regression modeling for classification accuracy of OD in CRS patients. A secondary goal was to

explore top predictors in high-performing models to identify common predictor variables of importance that may help explain the heterogeneity of OD found among CRS patients.

MATERIALS and METHODS

Study Population

Adult patients (≥18 years old) were recruited into a prospective, multi-center, observational cohort study, for which multiple reports have been previously published [12-15]. All study patients were diagnosed with medically refractory CRS as defined by the American Academy of Otolaryngology-Head and Neck Surgery 2007 guidelines [4]. Performance sites consisted of tertiary care sinus centers located within academic hospital systems in North America including: Oregon Health & Science University (OHSU; Portland, OR; eIRB#7198), Stanford University (Palo Alto, CA; IRB#4947), and the Medical University of South Carolina (Charleston, SC; IRB#12409), with central study coordination at OHSU. The Institutional Review Board (IRB) at each performance site governed all study protocols, informed consent documentation, and data safety monitoring.

At original study enrollment, participants were screened for demographics, social and medical history (**Table 1**). The diagnoses of asthma, allergy, and aspirin-exacerbated respiratory disease (AERD) were made by evaluation of the medical record, presence of classic symptoms and self-report of prior testing. Subjects with comorbid cystic fibrosis, ciliary dyskinesia, or autoimmune disease were included in the analysis. Patient reported outcome measures (PROMs) were collected during study enrollment, but were not used in the predictive analytics given the overlap between particular symptom ratings and objective smell loss [22-25].

Objective Measures of Disease Severity

Diagnostic measures of disease severity were collected as standard of care, and were also used for investigational purposes. These well-established objective measures have been recommended for clinical study [16], and include Lund-Kennedy nasal endoscopy scoring [17] and high-resolution CT imaging graded using the Lund-Mackay scoring system [18].

The main outcome of interest in the current study is olfactory dysfunction as measured by psychophysical testing. Olfactory function was assessed using the brief Smell Identification Test (bSIT) in the years 2011-2013, and the 40-item Smell Identification Test (SIT) in the years 2013-2015 (Sensonics Inc., Haddon Heights, NJ). Higher SIT total scores reflect a better sense of smell, and categorical ratings of "smell loss" (anosmia + hyposmia) and "normosmia" were assigned based on established scales [19-21].

Data Management and Statistical Analyses

Protected health information was previously removed, and study data were safeguarded using a unique study identification number assignment for each participant. Study data were securely transferred from a HIPAA compliant, relational database (Access, Microsoft Corp, Redmond, WA.) to the University of Colorado Department of Otolaryngology-Head and Neck Surgery password protected research server, according to specifications of a Data Use Agreement between OHSU and the University of Colorado.

Study data were evaluated descriptively using count (%) for categorical variables and median and interquartile range (IQR; 25th quantile, 75th quantile) for numeric variables. Comparing variables between normosmics and those with smell loss, Fisher's exact test was used for categorical variables and Wilcoxon rank sum test was used for numeric variables. Both the

original p-values and Benjamini-Hochberg multiple testing adjusted p-values ("q-values") were reported [26]. Statistical comparisons with q-values < 0.05 were regarded as statistically significant. All statistical and data analyses were completed using the R software program [27]. Machine learning models were compared to simpler logistic regression model with backwards step-wise variable selection using Akaike information criteria (AIC). The step-wise variable selection was re-performed during each iteration of the cross-validation procedure (details later), in order to prevent biased estimates of classification accuracy [28].

Machine Learning Approaches to Classification

Four different machine learning predictive analytics models were applied, in order to compare classification accuracy to logistic regression. All classifiers were fit using the caret R package [29], which provides a unified framework for tuning and evaluating classification accuracy. *Random forests (RF)* was chosen as it allows for non-linear and interaction effects; conditional inference trees were used with permutation-based variable importance scores for unbiased variable selection [30]. Similar to traditional logistic regression, least absolute shrinkage and selection operator (*LASSO*) assumes only linear and additive affects (without interaction) where regression coefficients are shrunk toward zero to perform variable selection [31,32]. In this model, unimportant variables are given coefficients equal to zero, effectively removing them from the model. Along these lines, multivariate adaptive regression splines (*MARS*) was also applied in a similar fashion to traditional logistic regression, but it uses stepwise methods for variable selection and allows for nonlinear and interaction effects [33]. Lastly, the *Support Vector Machine* (*SVM*) approach was applied with a radial basis kernel, which allows for non-linear and interaction effects [34].

Repeated 10-fold cross validation (repeated 5 times [35]) was used to tune and internally validate the classification accuracy of each model. The following measures of accuracy were considered: area under receiver operating characteristic curve (AUC), sensitivity (proportion of subjects with olfactory dysfunction that were correctly classified), and specificity (proportion of subjects without olfactory dysfunction that were correctly classified). Both sensitivity and specificity used a probability cutoff of 0.50. Pairwise two-sample t-tests were used to compare cross-validation accuracy measures between models using the Benjamini-Hochberg method to adjust for multiple testing. To deal with the fact that the two classes are unbalanced (dysosmia vs normosmia), under-sampling was used within the cross-validation procedure, as implemented in the caret R package [29,36]. Over-sampling was also considered, but produced nearly equal classification accuracy in terms of AUC (results not shown). The method of "surrogate splits" [37] was used to handle missing values in RF. The missForest R package [38] was used to impute missing values for all other models, while excluding the outcome in order to prevent over-fitting.

To identify the important predictor variables included in each classification algorithm, permutation-based variable importance scores were used to rank variables from most to least important in RF, while the absolute value of the standardized regression coefficient was used for LASSO. The IML R package [39] was used to calculate variable importance scores for SVM using a permutation method. All variable importance scores were scaled from 0 to 100, where the most important variable for the model has a score of 100. Of note, for a categorical variable with more than two categories, RF and LASSO calculate separate variable importance scores for each category, whereas the approach used for SVM only calculates an overall score.

RESULTS

Final Cohort Characteristics

A total of 611 study participants met all inclusion criteria and were prospectively enrolled between April 2011 and July 2015. Demographics and comorbid disease characteristics, examined by olfactory classification, are described in **Table 1**. The overall median age for the final cohort was 51 years (IQR= 38.3, 61.7) years with slight female preponderance (53% female). Approximately one-third of the cohort exhibited nasal polyps (35%).

Study participants were stratified by olfactory testing category. Statistically significant differences ($q<0.05$) were observed between the normosmic and hyposmic cohorts in terms of objective disease measures (CT and endoscopy scores), age, sex, previous surgery, socioeconomic status, corticosteroid dependence, asthma/COPD, and AERD. Sinonasal conditions including nasal polyps, allergic fungal rhinosinusitis, and deviated septum were also significant factors.

Smell loss vs Normosmia Classification Accuracy

A logistic regression model for classification of baseline olfactory dysfunction demonstrated acceptable accuracy in the context of CRS (AUC 0.707). Three of the four machine learning models outperformed the traditional logistic regression model (**Table 2**). The SVM approach was the most accurate, with a relatively good classification accuracy (AUC 0.754) and much higher sensitivity than other the methods. LASSO and Random Forest also outperformed logistic regression, whereas MARS had the lowest AUC. **Figure 1** illustrates two-sample t-tests to compare measures of classification accuracy between all models (across the 50 cross-validation resampling estimates), using the Benjamini-Hochberg method to adjust for multiple testing. Note that SVM, RF and LASSO did not have significantly different mean AUC, but they had significantly higher mean AUC compared to logistic regression and MARS. SVM had significantly higher sensitivity compared to all other models. Specificity did not significantly differ between models, although SVM did have the lowest mean specificity.

Machine Learning Predictions: Top variables

ML algorithms are adept are recognizing patterns within complex data, but the lack of *a priori* framework results in the inability to understand *why* the algorithm made its decisions. The ML models each arrived at somewhat different relative importance of a given predictor variable. SVM, the model with the largest AUC, considered many variables to be influential, with 20 variables having variable importance scores >25 (**Fig 2**). In Random Forest and LASSO models, fewer variables were important and the most weight was placed on the presence of nasal polyps and the nasal endoscopy severity score (**Fig 3 and Suppl Fig 1**).

Top predictors across the approaches include corticosteroid dependence, prior surgery, nasal polyps, AERD, age, sex, and smoking history. Inflammatory disease severity, as measured by CT and nasal endoscopy scores, were also strong predictors of CRS-OD.

Socioeconomic factors such as insurance status, household income, and race are among top predictors across the three models displayed. Age was a less important variable in all models. Comorbid medical conditions including asthma, diabetes, septal deviation, alcohol use, and sleep apnea, are also of moderate importance. Although the presence of nasal polyps, asthma, and AERD, are significantly associated with CRS-OD, the presence of allergic rhinitis by history or by skin testing appear to carry relatively minimal weight in the CRS-OD classifications.

DISCUSSION

CRS is a heterogeneous process in terms of presentation, subtypes, natural history and response to treatments [2]. Current data from a multi-institutional CRS cohort studying olfactory loss in CRS with a sensitive olfactory test (Sniffin' Sticks) demonstrates that nearly 70% of the CRS population has significant olfactory loss, with 20% exhibiting complete anosmia [7]. The pathophysiology of OD in CRS is ascribed to both sensorineural and conductive effects of local tissue inflammation, on top of the myriad factors that can affect olfaction in the general population [1,40]. In previous studies, factors associated with worse baseline olfactory function in CRS patients included nasal polyposis, asthma, age, smoking, and eosinophilia [41]. Other works have shown CRS-OD response treatment in ~40% of patients, with risk factors such as olfactory dysfunction severity, nasal polyps, female gender, high socioeconomic status and non-smoking associating with better quality of life results. In multivariate regression, only nasal polyps and degree of baseline olfactory dysfunction maintained statistical significance, highlighting the need for robust sample size and analytical methods to glean useful information from human study [10].

Here, we utilized ML-based data science approaches to interrogate whether novel predictive analytics approaches could outperform a traditional statistical method of logistic regression in CRS-OD classification. Top predictors across the approaches contain many variables associated with olfactory function in previous studies, including, nasal polyps, prior surgery, AERD, age, sex, corticosteroid dependence, and smoking history. Inflammatory disease severity, as measured by CT and nasal endoscopy scores, are not surprisingly very important predictors of CRS-OD. Socioeconomic factors such as insurance status, household income, and race were among the important variables across the three top performing models. Since age was less important in all models, insurance status appears to be an independent predictor (i.e., Medicare insurance is not simply a proxy for advanced age). Comorbid medical conditions including asthma, diabetes, septal deviation, alcohol use, and sleep apnea, are also of moderate importance. Notably, these factors can be influenced by medical care and lifestyle modification, suggesting that other approaches to improve general health and well-being may be considered as part of a holistic approach to CRS-OD management. Although the presence of nasal polyps, asthma, and AERD, are significantly associated with CRS-OD here and in prior literature, the presence of allergic rhinitis by history or by skin testing appeared to carry relatively minimal weight in the CRS-OD classifications. This observation suggests that medical therapies for atopic diseases--including immunotherapy and targeted biologics--may not necessarily provide benefit for olfactory function in CRS.

We observed improved prediction accuracy with most ML approaches, as measured by AUC and sensitivity/specificity. ML is defined as the use of algorithms to break down data, learn from it, and then make a determination or prediction about some aspect. Whereas traditional statistical methods focus on modeling how system variables relate to one another and what statistical inference (e.g. significance in p-values) can be made, the goal of machine learning (ML) is not interpretation of individual components but prediction of future outcomes. In doing so, ML provides a novel approach to uncover previously unrecognized patterns among CRS-OD patients and thereby offers numerous advantages over regression analyses, which have been traditionally employed in studies of CRS disease severity and outcomes [42].

In this study, four different supervised ML algorithms were used map features of interest to the outcome or "label" of olfactory loss, and three outperformed traditional regression. In particular, the SVM model had improved accuracy that was driven largely by improved

sensitivity. The improved classification accuracy may be attributed to the ability of these ML models to handle high-dimensional data in which more features exist than observations, and include non-linear and interaction effects. In contrast, traditional statistical modeling with a large number of potential features requires some form of dimension reduction or variable selection, and exploration of interactions and non-linear effects is challenging with more than a small number of predictors. As a result, we believe that ML and artificial intelligence data analytics are well-suited to prime research and eventually for the application of precision medicine in olfactory disorders, given the idiosyncrasies, nuances, and numerous possible predictor variables.

There are some limitations of the current study that should be addressed with regard to ML. Adequate enrollment by sex and race are common sampling issues in healthcare ML [43,44]. Details of this cohort have been extensively published and include even distributions by age and sex. Racial/ethnic minorities and low socioeconomic status groups are underrepresented perhaps as a function of geography and referral patterns of the enrollment sites, but also result from the general lack of accessible and affordable healthcare and racial disparities inherent within the U.S. healthcare system [45]. A related concern is that models created from our cohort data may "overfit" when new or unseen clinical data are applied. Certainly, these approaches should be further validated in independent studies, but nonetheless demonstrate the utility of data science approaches to uncover patterns within expansive human data in a noisy disease process. To incorporate a large enough patient dataset across three sites over several years of prospective data collection, psychophysical testing included classification of smell loss by either SIT or bSIT. Although the SIT may be more sensitive despite both of these tests being extensively validated in the literature, our prior analysis in a CRS cohort has demonstrated excellent diagnostic accuracy of the bSIT (AUC=0.873; 95% CI: 0.819, 0.927) [21]. Additionally, there may be important data inputs (i.e., molecular endotyping data) that are not yet clinically available. Serum markers, such as peripheral eosinophilia and serum IgE, have been proposed for CRS in the past but have generally been disappointing and do not appear to correlate with tissue levels [46,47]. Some of the clinical data input fields may associate phenotypes with likely molecular endotypes, but such overlap is not universal [5]. Future goals include testing whether additional accuracy is gained with the addition of molecular based biomarkers to the input data.

CONCLUSION

Here we demonstrate the importance of novel predictive analytics approaches in the study of clinical olfactory disorders. Founded in data science principles rather than traditional statistics, ML models generally outperformed logistic regression. Consistently observed predictors across models include a weighted interest in socioeconomic status, and other potentially modifiable conditions such as asthma/AERD. Further studies will be important to reproduce and validate these findings, and might explore why certain novel predictor variables appear to be important. Such approaches may have value for both clinical counseling and guidance for future basic science and translational research.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Miranda Kroehl, PhD, John Rice, PhD, and Krithika Suresh, PhD, for biostatistical expertise in machine learning and prediction models. We would also like to thank Drs. Peter Hwang, MD, Rod Schlosser, MD, and Luke Rudmik, MD, for their contributions in original study enrollment and ongoing collaboration.

342 **REFERENCES**

343

344 1. Hummel T, Whitcroft KL, Andrews P et al. Position Paper on Olfactory Dysfunction.
345 Rhinology 2016 31;56(1):1-30.

346 2. Orlandi RR, Kingdom TT, Hwang PH, et al. International Consensus Statement on Allergy
347 and Rhinology: Rhinosinusitis. Int Forum Allergy Rhinol. 2016;6 Suppl 1:S22-209. Section X.B.

348 3. Fokkens WJ, Lund VJ, Mullol J, et al. European position paper on rhinosinusitis and nasal
349 polyps 2012. Rhinol. Suppl 2012;23: 1-298.

350 4. Rosenfeld RM, Piccirillo JF, Chandrasekhar SS, et al. Clinical practice guideline (update):
351 adult sinusitis. Otolaryngol Head Neck Surg. 2015;152(2 Suppl):S1-39.

352 5. Akdis CA, Bachert C, Cingi C, et al. Endotypes and phenotypes of chronic rhinosinusitis: a
353 PRACTALL document of the European Academy of Allergy and Clinical Immunology and the
354 American Academy of Allergy, Asthma & Immunology. J Allergy Clin Immunol
355 2013;131(6):1479-90.

356 6. Soler ZM, Kohli P, Storck KA, Schlosser RJ. Olfactory Impairment in Chronic Rhinosinusitis
357 Using Threshold, Discrimination, and Identification Scores. Chem Senses 2016;41(9):713-719.

358 7. Schlosser RJ, Smith TL, Mace JC, et al. Factors driving olfactory loss in patients with chronic
359 rhinosinusitis: a case control study. Int Forum Allergy Rhinol 2020;10(1):7-14.

360 8. Kohli P, Naik AN, Harruff EE, et al. The prevalence of olfactory dysfunction in chronic
361 rhinosinusitis. Laryngoscope 2017; 127(2):309-320.

362 9. DeConde AS, Mace JC, Alt JA, Schlosser RJ, Smith TL, Soler ZM. Comparative effectiveness
363 of medical and surgical therapy on olfaction in chronic rhinosinusitis: a prospective, multi-
364 institutional study. Int Forum Allergy Rhinol. 2014;4(9):725-33.

365 10. Katotomichelakis M, Simopolous E, Tripsianis G, et al. Predictors of quality of life
366 outcomes in chronic rhinosinusitis after sinus surgery. Eur Arch Otorhinolaryngol 2014;
367 271(4):733-41.

368 11. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods
369 2018;15(4):233-234.

370 12. Smith TL, Kern R, Palmer JN, et al. Medical therapy vs surgery for chronic rhinosinusitis: a
371 prospective, multi-institutional study with 1-year follow-up. Int Forum Allergy Rhinol
372 2013;3(1):4-9.

373 13. DeConde AS, Mace JC, Alt JA, et al. Longitudinal improvement and stability of the SNOT-
374 22 survey in the evaluation of surgical management for chronic rhinosinusitis. Int Forum Allergy
375 Rhinol. 2015; 5(3):233-239.

376 14. Soler ZM, Smith TL, Alt JA, et al. Olfactory-specific quality of life outcomes after
377 endoscopic sinus surgery. Int Forum Allergy Rhinol. 2016; 6(4): 407-413.

378 15. Ramakrishnan VR, Mace JC, Soler ZM, Smith TL. Examination of high-antibiotic users in a
379 multi-institutional cohort of chronic rhinosinusitis patients. Int Forum Allergy Rhinol.
380 2017;7(4):343-51.

16. Hopkins C, Hettige R, Soni-Jaiswal A, et al. Chronic rhinosinusitis outcome measures (CHROME), developing a core outcome set for trials of interventions in chronic rhinosinusitis. Rhinology 2018;56(1):22-32.

17. Lund VJ and Kennedy DW. Quantification for staging sinusitis. The Staging and Therapy Group. Ann Otol Rhinol Laryngol Suppl. 1995;167:17-21.

18. Lund VJ and Mackay IS. Staging in rhinosinusitis. Rhinology. 1992;31(4):183-4.

19. Doty RL. The Smell Identification Test Administration Manual. 3. Haddon Heights, NJ: Sensonics, Inc; 1995.

20. Doty RL. The Brief Smell Identification Test Administration Manual. Haddon Heights, NJ: Sensonics, Inc; 2001.

21. El Rassi E, Mace JC, Steele TO, et al. Sensitivity analysis and diagnostic accuracy of the Brief Smell Identification Test in patients with rhinosinusitis. Int Forum Allergy Rhinol. 2016;6(3):287-92.

22. Piccirillo JF, Merritt MG, Jr, Richards ML. Psychometric and clinimetric validity of the 20-Item Sino-Nasal Outcome Test (SNOT-20) Otolaryngol Head Neck Surg. 2002;126(1):41–47.

23. Hopkins C, Gillett S, Slack R, et al. Psychometric validity of the 22-item Sinonasal Outcome Test. Clin Otolaryngol. 2009;34(5):447–454.

24. Senior BA, Glaze C, Benninger MS. Use of the rhinosinusitis disability index (RSDI) in rhinologic disease. Am J Rhinol. 2001; 15(1):15-20

25. Simopoulos E, Katotomichelakis M, Gouveris H, Tripsianis G, Livaditis M, Danielides V. Olfaction-associated quality of life in chronic rhinosinusitis: adaptation and validation of an olfaction-specific questionnaire. Laryngoscope 2012;122:1450–1454.

26. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 1995;57(1):289-300.

27. R Core Team. 2019. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

28. Hastie T, Tibshirani R, and Friedman J. The wrong and right way to do cross-validation. Elements of Statistical Learning, Data Mining, Inference, Prediction, Springer, NY 2009: 245-247.

29. Kuhn, M. Building predictive models in R using the caret package. Journal of statistical software 2008; 28(5):1-26.

30. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 2007;8(1):25.

31. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 1996;58(1):267-288.

32. Friedman J, Hastie T, and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 2010;33(1):1.

33. Friedman JH. Multivariate adaptive regression splines. Annals of Statistics 1991;1-67.

421 34. Cristianini N and Shawe-Taylor J. An introduction to support vector machines and other
422 kernel-based learning methods. Cambridge university press, 2000.

423 35. Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out
424 and bootstrap. Computational statistics & data analysis 2009;53(11):3735-3745.

425 36. He H and Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and
426 data engineering 2009;21(9):1263-1284.

427 37. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: Rationale, application,
428 and characteristics of classification and regression trees, bagging, and random forests.
429 Psychological methods 2009; 14:323.

430 38. Stekhoven DJ and Bühlmann, P. MissForest—non-parametric missing value imputation for
431 mixed-type data. Bioinformatics 2012;28(1):112-118.

432 39. Molnar C, Casalicchio G, Bischl B. iml: An R package for interpretable machine learning. J
433 Open Source Software 2018;3(26), 786.

434 40. Rombaux P, Huart C, Levie P, Cingi C, Hummel T. Olfaction in Chronic Rhinosinusitis.
435 Curr Allergy Asthma Rep 2016;16(5):41.

436 41. Litvack JR, Fong K, Mace J, et al. Predictors of olfactory dysfunction in patients with
437 chronic rhinosinusitis. Laryngoscope 2008;118:2225–30.

438 42. Michie D, Spiegelhalter D, Taylor C. Machine Learning, Neural, and Statistical
439 Classification. New York, NY: Ellis Horwood; 1994.

440 43. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med
441 2019;380(14):1347-1358.

442 44. Cutillo CM, Sharma KR, Foschini L, et al; MI in Healthcare Workshop Working Group.
443 Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and
444 transparency. NPJ Digit Med 2020;3:47.

445 45. 2018 National Healthcare Quality and Disparities Report. Content last reviewed April 2020.
446 Agency for Healthcare Research and Quality, Rockville, MD.
447 https://www.ahrq.gov/research/findings/nhqrdr/nhqdr18/index.html

448 46. Gitomer SA, Fountain CR, Kingdom TT, et al. Clinical Examination of Tissue Eosinophilia
449 in Patients with Chronic Rhinosinusitis and Nasal Polyposis. Otolaryngol Head Neck Surg
450 2016;155(1):173-8.

451 47. Settipane GA. Nasal polyps and immunoglobulin E (IgE). Allergy Asthma Proc
452 1996;17(5):269-73.

453

453 **Table 1. Study population.** Collected demographics and clinical metadata used as inputs
454 for classification models are shown here for the overall cohort, and those with and without
455 olfactory dysfunction. Count (%) and Fisher's exact test were used for categorical variables,
456 while median (25th quantile, 75th quantile) and Wilcoxon rank sum test were used for numeric
457 variables. Data sorted from smallest to largest q-value (multiple testing adjusted p-value).
458

Characteristic	Overall, n = 611	n	Olfactory Dysfunction		p-value	q-value
			Yes, n = 203	No, n = 408		
CRS with POLYPS		611			**<0.001**	**<0.001**
No	399 (65%)		86 (42%)	313 (77%)		
Yes	212 (35%)		117 (58%)	95 (23%)		
TOTAL ENDOSCOPY SCORE	6.0 (4.0, 8.5)	607	8.0 (4.0, 11.0)	4.0 (2.0, 7.0)	**<0.001**	**<0.001**
TOTAL CT SCORE	12.0 (7.0, 16.0)	582	15.0 (10.0, 19.0)	11.0 (6.0, 14.0)	**<0.001**	**<0.001**
AERD		611			**<0.001**	**<0.001**
No	567 (93%)		173 (85%)	394 (97%)		
Yes	44 (7.2%)		30 (15%)	14 (3.4%)		
AGE	51.0 (38.3, 61.7)	611	56.6 (44.6, 65.4)	49.1 (36.9, 59.9)	**<0.001**	**<0.001**
PRIOR SURGERY		610			**<0.001**	**<0.001**
0	278 (46%)		66 (33%)	212 (52%)		
1	153 (25%)		56 (28%)	97 (24%)		
2	90 (15%)		34 (17%)	56 (14%)		
3	40 (6.6%)		24 (12%)	16 (3.9%)		
4+	49 (8.0%)		23 (11%)	26 (6.4%)		
INSURANCE		610			**<0.001**	**<0.001**
Employer provided	373 (61%)		112 (55%)	261 (64%)		
Medicaid	24 (3.9%)		4 (2.0%)	20 (4.9%)		
Medicare	129 (21%)		63 (31%)	66 (16%)		
None	6 (1.0%)		5 (2.5%)	1 (0.2%)		
Private	62 (10%)		13 (6.4%)	49 (12%)		
State Assisted	8 (1.3%)		4 (2.0%)	4 (1.0%)		
VA Benefits	3 (0.5%)		0 (0%)	3 (0.7%)		
VA Benefits / Tricare	5 (0.8%)		2 (1.0%)	3 (0.7%)		
ASTHMA		611			**<0.001**	**<0.001**
No	397 (65%)		110 (54%)	287 (70%)		
Yes	214 (35%)		93 (46%)	121 (30%)		

Characteristic	Overall, n = 611	n	Olfactory Dysfunction		p-value	q-value
			Yes, n = 203	No, n = 408		
SEPTAL DEVIATION		611			**<0.001**	**0.001**
No	408 (67%)		155 (76%)	253 (62%)		
Yes	203 (33%)		48 (24%)	155 (38%)		
RECURRENT ACUTE SINUSITIS		611			**<0.001**	**0.003**
No	571 (93%)		199 (98%)	372 (91%)		
Yes	40 (6.5%)		4 (2.0%)	36 (8.8%)		
STEROID DEPENDENCE		611			**0.003**	**0.009**
No	555 (91%)		174 (86%)	381 (93%)		
Yes	56 (9.2%)		29 (14%)	27 (6.6%)		
SEX		611			**0.013**	**0.037**
Female	324 (53%)		93 (46%)	231 (57%)		
Male	287 (47%)		110 (54%)	177 (43%)		
AFRS		611			**0.019**	**0.049**
No	593 (97%)		192 (95%)	401 (98%)		
Yes	18 (2.9%)		11 (5.4%)	7 (1.7%)		
COPD		611			**0.020**	**0.049**
No	579 (95%)		186 (92%)	393 (96%)		
Yes	32 (5.2%)		17 (8.4%)	15 (3.7%)		
HOUSEHOLD INCOME		595			**0.027**	0.064
0-25000	96 (16%)		34 (17%)	62 (16%)		
26000-50000	102 (17%)		47 (24%)	55 (14%)		
51000-75000	120 (20%)		37 (19%)	83 (21%)		
76000-100000	107 (18%)		28 (14%)	79 (20%)		
100000+	170 (29%)		51 (26%)	119 (30%)		
Inferior Turbinate Hypertrophy		611			0.091	0.200
No	520 (85%)		180 (89%)	340 (83%)		
Yes	91 (15%)		23 (11%)	68 (17%)		
CF or ciliary dysfunction		611			0.117	0.241
No	588 (96%)		199 (98%)	389 (95%)		
Yes	23 (3.8%)		4 (2.0%)	19 (4.7%)		
ALLERGY by HISTORY		611			0.143	0.277
No	480 (79%)		152 (75%)	328 (80%)		

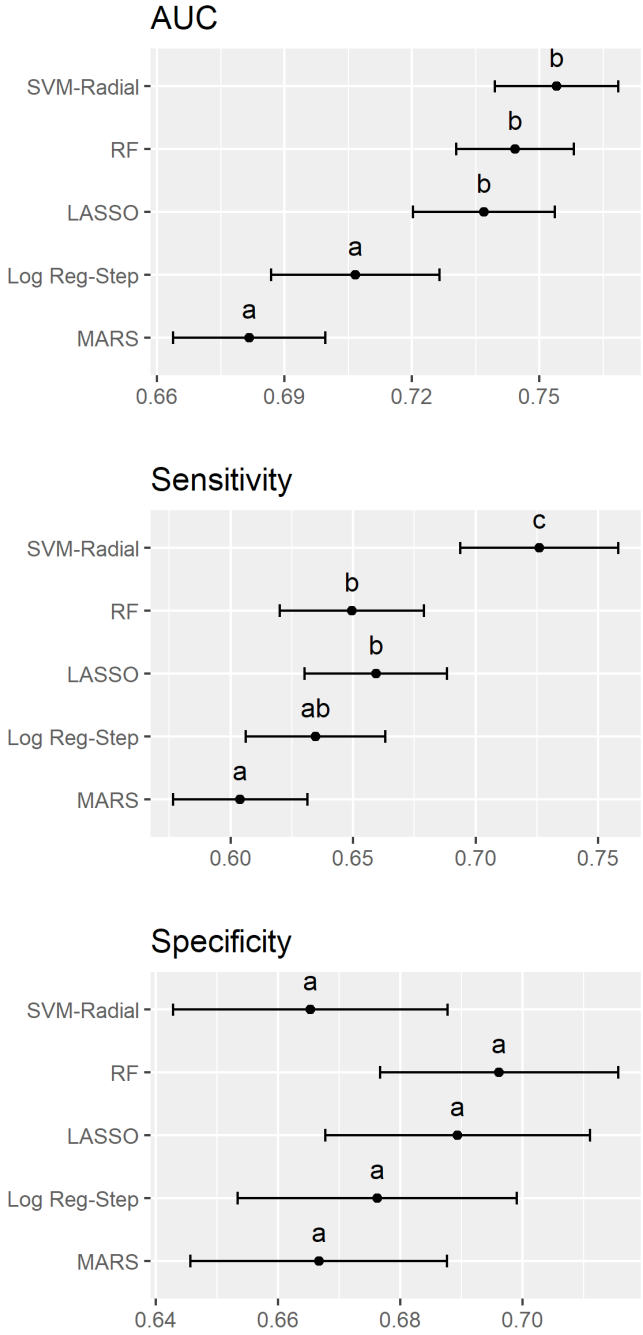
Characteristic	Overall, n = 611	n	Olfactory Dysfunction		p-value	q-value
			Yes, n = 203	No, n = 408		
Yes	131 (21%)		51 (25%)	80 (20%)		
SITE		611			0.189	0.349
#1	222 (36%)		84 (41%)	138 (34%)		
#2	253 (41%)		77 (38%)	176 (43%)		
#3	136 (22%)		42 (21%)	94 (23%)		
OSA by HISTORY		611			0.215	0.376
No	584 (96%)		191 (94%)	393 (96%)		
Yes	27 (4.4%)		12 (5.9%)	15 (3.7%)		
EDUCATION (Yrs)	16.0 (13.0, 17.0)	608	15.0 (13.0, 16.0)	16.0 (13.0, 17.0)	0.235	0.392
RACE		610			0.247	0.393
African American	49 (8.0%)		18 (8.9%)	31 (7.6%)		
American Indian/Alaska Native	5 (0.8%)		2 (1.0%)	3 (0.7%)		
Asian	24 (3.9%)		10 (4.9%)	14 (3.4%)		
Native Hawaiian/Pacific Islander	1 (0.2%)		1 (0.5%)	0 (0%)		
Other	25 (4.1%)		12 (5.9%)	13 (3.2%)		
White	506 (83%)		160 (79%)	346 (85%)		
GERD		611			0.303	0.460
No	570 (93%)		186 (92%)	384 (94%)		
Yes	41 (6.7%)		17 (8.4%)	24 (5.9%)		
ETHNICITY		611			0.352	0.493
Hispanic / Latino	21 (3.4%)		9 (4.4%)	12 (2.9%)		
Non-Hispanic / Latino	590 (97%)		194 (96%)	396 (97%)		
ALCOHOL Use		608			0.344	0.493
No	313 (51%)		109 (54%)	204 (50%)		
Yes	295 (49%)		92 (46%)	203 (50%)		
ALLERGY_TESTING		611			0.382	0.510
No	362 (59%)		115 (57%)	247 (61%)		
Yes	249 (41%)		88 (43%)	161 (39%)		
DEPRESSION		611			0.403	0.510
No	518 (85%)		176 (87%)	342 (84%)		
Yes	93 (15%)		27 (13%)	66 (16%)		

Characteristic	Overall, n = 611	n	Olfactory Dysfunction		p-value	q-value
			Yes, n = 203	No, n = 408		
FIBROMYALGIA		611			0.408	0.510
No	584 (96%)		192 (95%)	392 (96%)		
Yes	27 (4.4%)		11 (5.4%)	16 (3.9%)		
OSA BY TESTING		611			0.450	0.543
No	557 (91%)		188 (93%)	369 (90%)		
Yes	54 (8.8%)		15 (7.4%)	39 (9.6%)		
ALCOHOL USE (#)	0.0 (0.0, 36.0)	608	0.0 (0.0, 24.0)	0.0 (0.0, 36.0)	0.473	0.552
DIABETES		609			0.536	0.605
No	558 (92%)		183 (91%)	375 (92%)		
Yes	51 (8.4%)		19 (9.4%)	32 (7.9%)		
AUTOIMMUNE DISEASE		610			0.589	0.645
No	573 (94%)		188 (93%)	385 (94%)		
Yes	37 (6.1%)		14 (6.9%)	23 (5.6%)		
IMMUNODEFICIENCY		611			0.626	0.664
No	592 (97%)		198 (98%)	394 (97%)		
Yes	19 (3.1%)		5 (2.5%)	14 (3.4%)		
SMOKER (current or former)		607			1.000	1.000
No	574 (95%)		190 (95%)	384 (95%)		
Yes	33 (5.4%)		11 (5.5%)	22 (5.4%)		

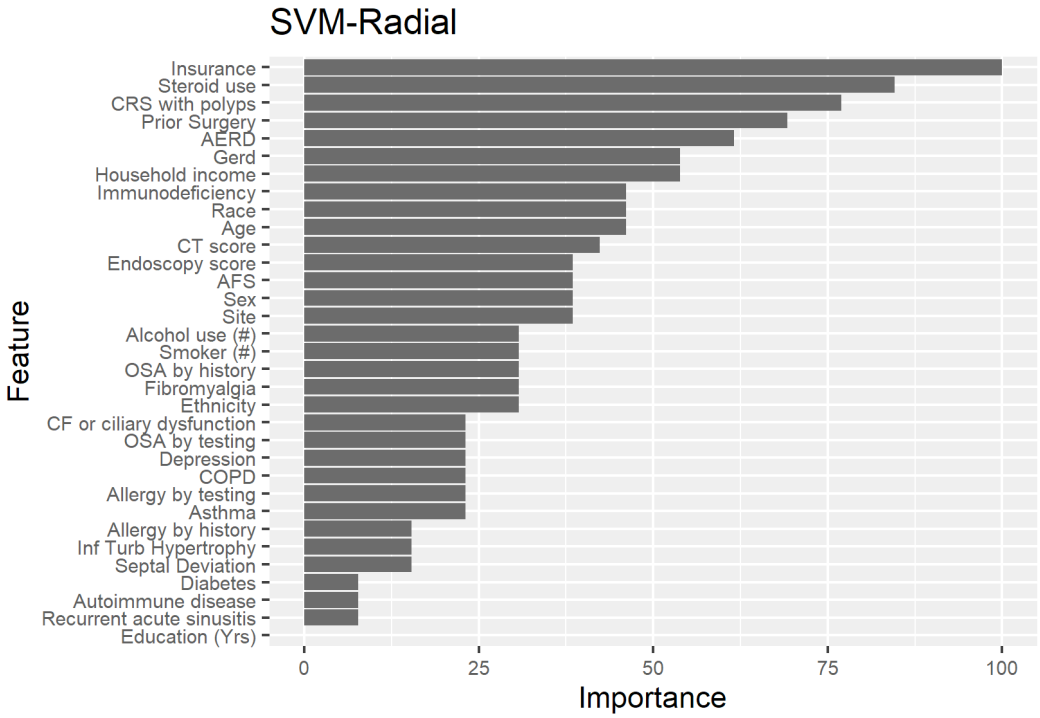
Table 2. **Comparing classification accuracy.** Three ML approaches outperform traditional logistic regression. *AUC = area under receiver operating characteristic curve; SVM-Radial = support vector machine with a radial basis kernel; Log Reg-Step = logistic regression with step-wise variable selection.*

Model	AUC	Sensitivity	Specificity
SVM-Radial	0.754	0.726	0.665
Random Forest	0.744	0.649	0.696
LASSO	0.737	0.659	0.689
Log Reg-Step	0.707	0.635	0.676
MARS	0.682	0.604	0.667

Figure 1. Comparing classification accuracy between methods. Dots correspond to the mean estimate (across all 50 cross-validation resampling estimates) and error bars are 95% confidence intervals for the mean. Pairwise comparisons were performed using Welch's 2-sample t-test with the Benjamini-Hochberg multiple testing adjusted p-values ("q-values"). Models with different letters (a,b,c) have significantly different means (q<0.05), while models sharing the same letter do not significantly differ (q>0.05). *SVM-Radial = support vector machine with a radial basis kernel; RF = random forest; Log Reg-Step = logistic regression with step-wise variable selection.*

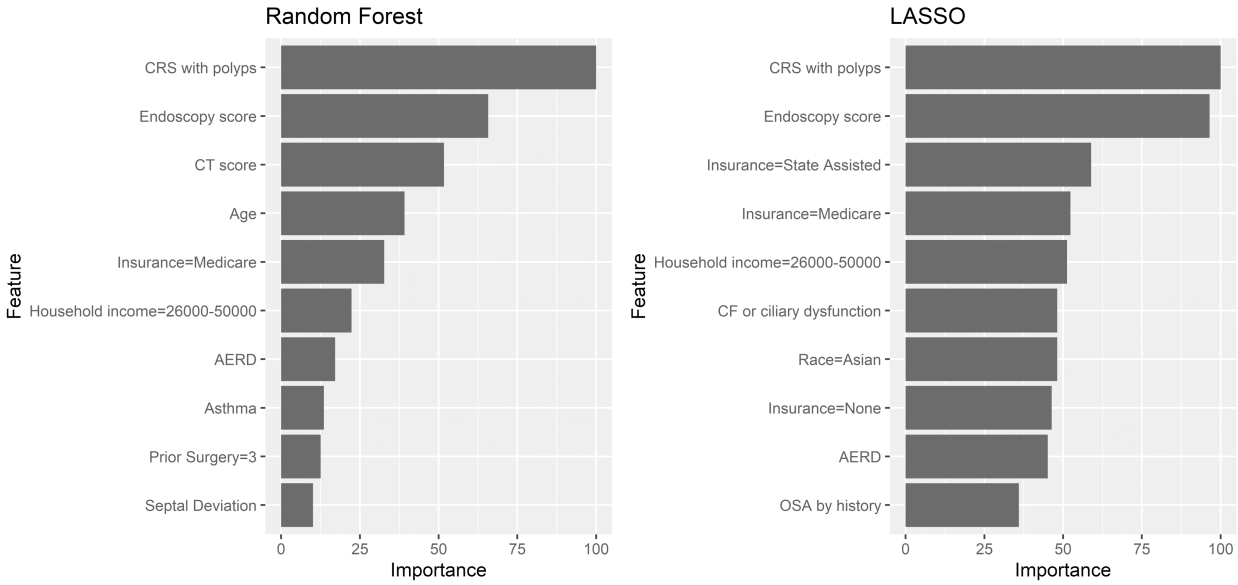


481 **Figure 2**. **Variable importance display for most accurate classification model (Support**
482 **Vector Machine with a radial basis kernel).** Note the inclusion of many predictor variables –
483 32 are included with > 10% variable importance, suggesting potential interaction between
484 predictors. *SVM-Radial = support vector machine with a radial basis kernel.*
485



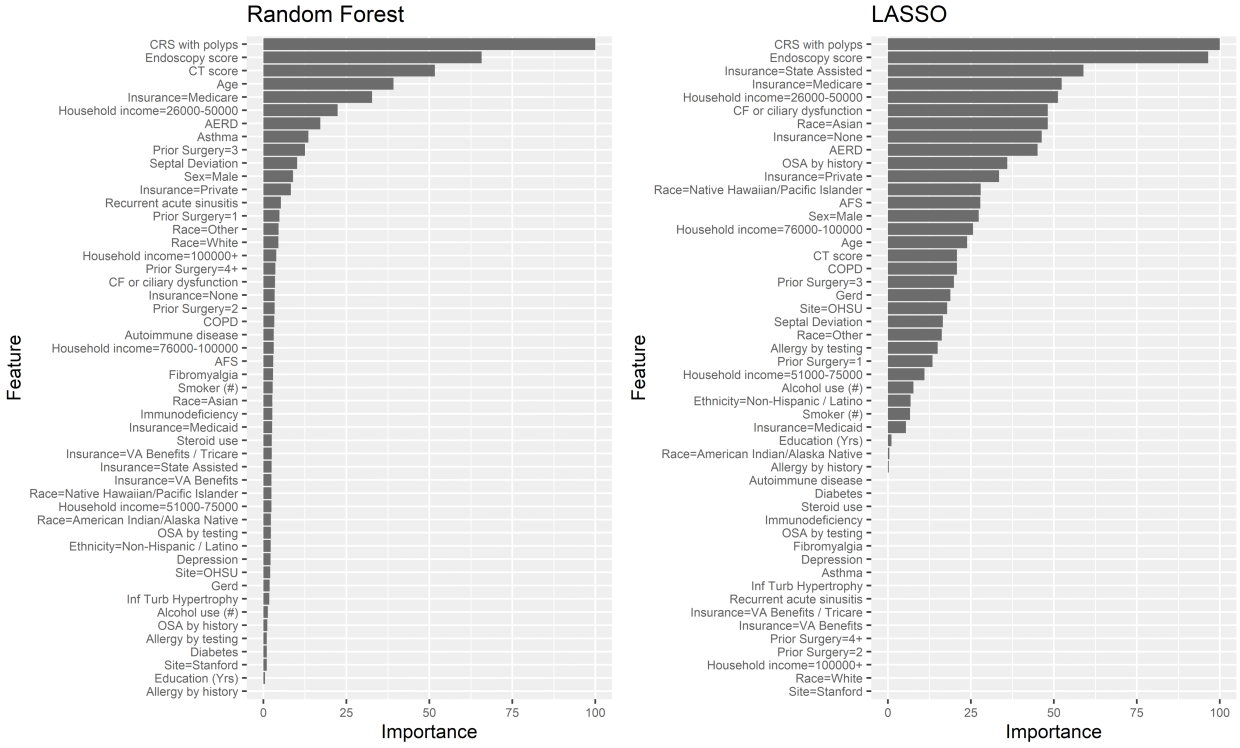
486
487
488
489

489 **Figure 3**. **Top 10 Predictors and Variable importance for Random Forest and LASSO**
490 **models.** *AERD = aspirin exacerbated respiratory disease; CF = cystic fibrosis.*
491



492
493

493 **Supplementary Figure 1**. **Complete list of predictors for Random Forest and LASSO**
494 **models.** Notably, the LASSO model incorporates significantly more features of interest than the
495 Random Forest algorithm (26 features with >10% importance in the prediction algorithm).
496
497



498