

Original Research

Improving the Likelihood of Identifying Alpha-1 Antitrypsin Deficiency Among Patients With COPD: A Novel Predictive Model Using Real-world Data

Daniel N. Pfeffer, PhD¹ Rahul Dhakne, MS² Omnya El Massad, MSc¹ Pulkit Sehgal, BS²
Thomas Ardiles, MD³ Michael O. Calloway, PhD⁴ M. Chris Runken, PharmD⁴ Charlie Strange,
MD^{5,6}

¹Data & Analytics, EVERSANA, Milwaukee, Wisconsin, United States

²Data & Analytics, EVERSANA, Pune, India

³Medical Affairs, Grifols Shared Services North America, Durham, North Carolina, United States

⁴Global Health Economics and Outcomes Research/Real-World Evidence, Grifols Shared Services North America, Durham, North Carolina, United States

⁵Division of Pulmonary, Critical Care, Allergy, and Sleep Medicine, Medical University of South Carolina, Charleston, South Carolina, United States

⁶AlphaNet, Coral Gables, Florida, United States

Address correspondence to:

M. Chris Runken, PharmD
Grifols, Shared Services North America
79 T.W. Alexander Drive
Durham, NC 27709
Phone: (919) 316-6343
Email: chris.runken@grifols.com

Running Head: Improving Identification of AATD in COPD

Keywords: alpha-1 antitrypsin deficiency; electronic health records; predictive model; detection; testing

Abbreviations: Alpha-1 antitrypsin deficiency (AATD); Alpha-1 antitrypsin (AAT); Chronic obstructive pulmonary disease (COPD); Serpin family A member 1 (*SERPINA1*); Real-world data (RWD); Electronic health records (EHR); International Classification of Diseases (ICD); eXtreme Gradient Boosting (XGBOOST); Area under the curve (AUC); Receiver operating characteristic (ROC); Shapley Additive exPlanations (SHAP)

Funding Support: This study was supported by Grifols Shared Services of North America, Inc.

Date of Acceptance: November 19, 2024 | **Published Online Date:** December 4, 2024

Citation: Pfeffer DN, Dhakne R, El Massad O, et al. Improving the likelihood of identifying alpha-1 antitrypsin deficiency among patients with COPD: a novel predictive model using real-world data. *Chronic Obstr Pulm Dis*. 2025; Published online December 4, 2024.

<https://doi.org/10.15326/jcopdf.2023.0491>

This article has an online supplement.

Pre-proof

Abstract

Background: Despite guideline recommendations, most patients with COPD do not undergo alpha-1 antitrypsin deficiency (AATD) testing and approximately 90% of people with AATD in the US remain undiagnosed. This study sought to develop a predictive model using real-world data to improve detection of AATD-positive patients in the general COPD population.

Methods: A predictive model using XGBoost was developed using the EVERSANA database, including longitudinal, patient-level medical claims, prescription claims, AATD-specific testing data, and electronic health records (EHR). The model was trained and then validated to predict AATD-positive status. Patients were coded as AATD positive based on the presence of any of the following criteria: 1) ≥ 2 AATD diagnosis codes in claims; 2) an AATD diagnosis code in the EHR; 3) a positive laboratory test for AATD; or 4) use of AATD-related medication. Over 500 variables were used to train the predictive model and >20 models were run to optimize the predictive power.

Results: 13,585 AATD-positive patients and 7,796 AATD-negative patients were included in the model. Inclusion of non-AATD laboratory test results was critical for defining cohorts and optimizing model prediction (e.g., respiratory comorbidities, and calcium, glucose, hemoglobin, and bilirubin levels). The final model yielded high predictive power, with an area under the receiver operating characteristic curve of 0.9.

Conclusions: Predictive modeling using real-world data is a sound approach for assessing AATD risk and useful for identifying COPD patients who should be confirmed by genetic testing. External validation is warranted to further assess the generalizability of these results.

Introduction

Alpha-1 antitrypsin deficiency (AATD) is an autosomal, co-dominant genetic disorder characterized by low serum protein levels of the protease inhibitor alpha-1 antitrypsin (AAT), resulting in an increased risk of chronic obstructive pulmonary disease (COPD), bronchiectasis, and/or liver disease.¹⁻³ AATD typically presents in individuals who inherited two disease-associated variant alleles of the gene that encodes AAT, serpin family A member 1 (*SERPINA1*), with over 200 variations associated with altered protein production, folding, or glycosylation.⁴ Notably, patients with AATD-associated COPD have higher rates of AATD-related comorbidities (i.e., pulmonary heart disease, chronic liver disease, cirrhosis, and liver transplantation) than patients with non-AATD-COPD.⁵ Therefore, the ability to distinguish AATD-associated COPD from non-AATD-COPD would help to more efficiently identify patients for confirmatory genetic testing, as well as those who may derive benefit from AATD-specific therapies and earlier treatment.

The estimated prevalence of AATD is 1% to 3% of patients with COPD, however, AATD is under-recognized, with approximately 90% to 95% of patients remaining undiagnosed.^{1,2,6,7} Brantly and colleagues suggested that targeting tests for symptomatic patients with bronchiectasis or cryptogenic liver disease and family members of known AATD patients would yield a higher detection rate.¹ Current guideline recommendations from the World Health Association, the American Thoracic Society, the European Respiratory Society, and the COPD Foundation advocate that all COPD patients be tested for AATD. Despite these universal guidelines and that AATD testing can be conveniently administered free of charge to patients in home self-test kits, a large majority of COPD patients do not undergo genetic testing.^{8,9} This lack of testing is primarily driven by the low detection rates (1 in 100) observed when testing broad COPD populations. As a

result, individuals with AATD frequently experience prolonged delays to diagnosis (estimated range: 5 to 8 years) from the onset of their COPD symptoms.⁶ This delay in diagnosis may result in significant disease exacerbations and progression, as well as shorter overall and transplant-free survival time.¹⁰⁻¹² Therefore, along with more opportunities to be tested, additional strategies are needed to improve the accuracy of identifying AATD-positive patients within COPD populations, to improve detection rates and to minimize delays in diagnosis. This model is one such strategy.

Augmentation therapy involves infusion of human plasma-derived AAT and is currently the only AATD-specific pharmacological intervention.¹³ However, there are many other benefits to establishing an AATD diagnosis that relate to smoking cessation, and avoidance of obesity, that may slow disease progression.¹⁴⁻¹⁶ Among patients with the most severe form of AATD, augmentation therapy is associated with decreased mortality risk, as well as reduced rates of lung density loss and emphysema progression.¹⁷⁻²¹

Developing computer models to assist in diagnosing AATD patients has been hampered by the small number of AATD-diagnosed patients in available real-world data (RWD).^{22,23} Recent attempts are aided by two important developments that help overcome this limitation. One is the advent of Big Data where the size and comprehensiveness of patient records and databases have grown.²⁴ A second advancement has been in the use of machine-learning programs for developing models. Notably, machine-learning based models have found recent success in predicting readmission after hospitalization for COPD, and predicting outcomes of AATD-associated liver disease.^{25,26} To aid in the detection of AATD-positive patients in the general COPD population, this study developed a novel, predictive model that uses both RWD and machine-learning algorithms. Individuals identified by use of this model might be candidates for lifestyle modification, AATD-related therapy, and/or clinical trials of novel therapies.

Pre-proof

Methods

Data sources and patient selection

In this study, US open claims data and electronic health records (EHR) were utilized to identify AATD-positive and AATD-negative patients with COPD using diagnosis codes, treatments prescribed, and laboratory tests. EVERSANA[®]'s EHR and Open-claims payer databases were linked via Datavant Tokenization to extract the model population. EVERSANA[®]'s EHR represents approximately 115 million US patients from >2,000 outpatient clinics and 50 hospitals running >30 different EHR software solutions between 2014 and 2022. Open-claims payer data (i.e., sourced from providers) were extracted from the Symphony Health IDV[®], a healthcare database that integrates deidentified data from physician practices, pharmacies, and hospitals for over 280 million active unique patients (2018-2022) in the US across commercial, Medicare, and Medicaid plans.

The model's initial population included patients with ≥ 1 COPD diagnosis code. Of these patients, those who met ≥ 1 of the following criteria were considered AATD positive: 1) ≥ 2 AATD diagnosis codes in claims; 2) an AATD diagnosis code in the EHR; 3) a positive laboratory test for AATD; or 4) the use of an AATD-related medication (i.e., augmentation therapy). A positive laboratory test for AATD was defined as having an abnormal serum AAT level (i.e., < 90 mg/dL) in the EHR and/or a confirmatory genotype test of the *SERPINA1* gene. Conversely, AATD-negative status was determined by a negative laboratory test for AATD. Together, the AATD-positive and AATD-negative groups constituted the final evaluable population. This population was then randomly divided, with 75% used to train the model and the remaining 25% to test and confirm the model.

Feature selection

Potential predictor features (i.e., multivariate, non-linear features), that are shared between the AATD-positive and AATD-negative groups, were used to build a predictive model. These features included common International Classification of Diseases (ICD) diagnosis codes (ICD-9/ICD-10), treatments grouped by generic name, and common laboratory tests grouped by component type. The results of the laboratory tests were grouped as follows: 1) low (i.e., bottom third of the expected range); 2) medium (i.e., middle third of the expected range); 3) high (i.e., upper third of the expected range); and 4) abnormal (i.e., any result that was abnormally higher or lower than the expected range). AAT levels were excluded as a laboratory feature since the model was built with the assumption that AAT laboratory values would not consistently be available. All features included in the predictive model were binary (i.e., if a patient had a diagnosis/treatment/laboratory result, the feature would be assigned a value of 1; otherwise, the value was 0), with the exception of age (continuous feature).

Predictive model development

A predictive model (predicting AATD positivity) was trained using the randomly selected training population (75% of the dataset) and eXtreme Gradient Boosting (XGBoost), a machine-learning technique.²⁷ This technique sequentially builds multiple decision trees, which allows for error corrections in a stepwise and gradual manner (i.e., the errors in early trees are corrected in trees that are generated later). The Boosted Decision Trees generally minimize the errors of the previous trees by using a gradient descent algorithm. XGBoost further enhances gradient-boosted decision trees by adding regularization terms on the final weights. Regularization terms help the model to prevent over-fitting, learn to handle sparse data, and provides general optimizations to decrease

the time required to train a model. Over 500 features were incorporated into the present model, and more than 20 models were run using different samples and exclusion rules, to optimize the model's predictive power. The performance of the final model was assessed in the randomly selected testing population (25% of the dataset) by examining the area under the curve (AUC) of the receiver operating characteristic (ROC; true positive rate vs. false positive rate) curve.

Assessment of feature importance

The features found to be most important in the predictive model were first identified using the default XGBoost feature importance scoring, which is based on the average information gain across all splits (decisions in the tree) in which the feature was used. The Shapley Additive exPlanations (SHAP) approach was utilized to quantify the contribution of each feature in correctly predicting whether a patient was AATD-positive or AATD-negative.^{28,29} The SHAP analysis was performed using the TreeExplainer method on the produced XGBoost model.

Results

Cohort definition and demographics

A total of 10,449,445 patients with ≥ 1 COPD code were identified in the combined open claims and EHR datasets. Of these, 2,902 patients were excluded from analysis, as they were classified as AATD-positive but had a negative test result. From the remaining 10,446,543 patients, 13,585 AATD-positive patients (0.1% of the cohort) and 7,796 AATD-negative patients were analyzed (**Figure 1**).

Among the AATD-positive cohort (**Table 1**), most patients were identified using ICD diagnosis codes, with a subset identified based on low serum AAT levels (i.e., generally <90 mg/dL, although some laboratories used a reference threshold of <100 mg/dL for abnormally low serum AAT levels; **Figure 2**). Nearly one-third (31.1%) of patients had received an AATD-specific medication. Cohort demographics are summarized in **Table 2**. In both groups, the median age at COPD diagnosis was 63 years and more than half were female (AATD-positive: 56.2%; AATD-negative: 58.3%). The median recorded patient history prior to a COPD diagnosis was 1 and 7 months in the AATD-positive and AATD-negative groups, respectively; both cohorts had a similar median post-COPD follow-up time (AATD-positive: 28 months; AATD-negative: 26 months). Many of the patients (40% to 50%) were located within the southern region of the US at diagnosis. Compared with the AATD-negative group, patients in the AATD-positive group had higher rates of bronchiectasis and emphysema, but lower rates of asthma and chronic bronchitis. Further, higher rates of cardiac arrhythmias and liver disease were observed in the AATD-negative group compared with the AATD-positive group. The latter may reflect the exclusion of the estimated

3.2% of patients with AATD who have independent liver involvement (i.e., without obstructive lung disease),³⁰ given the present model exclusively focuses on COPD patients.

Feature importance

Among the >500 features included in the initial predictive model, the inclusion of non-AATD-related laboratory test results was critical for model prediction (**Supplementary Figure 1**). First, among laboratory test results, key predictors of a positive or negative AATD diagnosis included calcium, glucose, hemoglobin, ferritin, and bilirubin levels. Normal calcium levels, and low bilirubin levels were indicators of being AATD negative. Abnormal or high glucose levels were associated with being AATD negative, whereas low or normal glucose levels had a small positive impact on predicting AATD-positive cases. In order to further characterize the impact of these features in distinguishing AATD positive or negative cases, SHAP values were estimated to identify each individual feature's magnitude and direction on the predictive model.

The features most likely to predict AATD cases included having emphysema, shortness of breath, as well as the use of albuterol sulfate and doxycycline hyclate (**Figure 3**). Features that were less likely to predict AATD-positive disease included hyperlipidemia, hypertension, esophageal reflux, low hemoglobin levels (i.e., bottom third of the expected range), smoking-related features, and older age (**Figure 3**).

Secondly, for laboratory features, the values were split into tertiles (i.e., low, medium, and high within the expected range) and abnormal (i.e., outside the expected range). Combining abnormally high or low patients into a single category resulted in a small degree of information loss relative

to the use of a separate group for each, although this likely had a negligible impact on the overall model results. Of the top laboratory results included in the model, hemoglobin, ferritin, and the hemogram/platelets/differential panel were less likely to be associated with being AATD positive, regardless of whether the values were within expected range (low, medium, high or abnormal). Ceruloplasmin showed a similar pattern, with the exception of the abnormal ceruloplasmin group (**Figure 4**). Having normal chloride levels was a good indicator of being AATD negative.

Cumulatively, these findings highlight the model's complexity, as there are multiple paths (i.e., trees) that a feature could take when predicting AATD positive or negative status, owing to the complex interactions between the model's features. For example, an AATD-negative patient could have a hemoglobin level that is within the lower third of the expected range in combination with a specific set of features, whereas an AATD-positive patient could have a similar hemoglobin level but in combination with a different set of features (**Figure 5** and **Figure 6**). Lastly, as the model assesses multiple features simultaneously, features that are correlated with each other (e.g., age and cardiovascular comorbidities, given their greater prevalence with increasing age) may be used as a proxy for these other features.

Performance of the predictive model

In this study, the final model was quite accurate in predicting known AATD-positive and AATD-negative patients. The final model had a combined AUC of the receiver operating characteristic curve of 0.9, indicating high predictive power for AATD diagnosis (**Supplementary Figure 2**). The model's sensitivity and specificity varied based on the chosen threshold. When a threshold of 0.49 was set, the sensitivity and specificity were both 0.82, meaning that 82% of AATD-positive

patients would be correctly identified, but simultaneously, 18% of AATD-negative patients would be incorrectly classified as AATD positive. Increasing the threshold to 0.68 reduced the sensitivity to 0.68 while increasing the specificity to 0.93 (i.e., 68% of AATD-positive patients would be correctly labeled, and 7% would be incorrectly labeled as AATD positive). Based on the current model performance, to identify 1 patient with confirmed AATD, subsequent confirmatory testing would be needed for 25 individuals categorized as likely AATD-positive patients (assuming a conservative 1% prevalence of AATD among patients with COPD). This is much better than the current 1 in 100 detection rate in real-world clinical settings.¹

Discussion

AATD is an under-diagnosed disease, even among COPD patients and the resulting delays in diagnosis and proper treatment lead to an increased risk of long-term complications and exacerbations.^{21,22} In fact, these data showed that of approximately 10.5 million COPD patients, only 21,381 had been tested. While this could be an artifact of the data and limitations in its collection and retention, the apparent lack of testing is notable. This study demonstrated that predictive modeling using RWD from EHRs and open-claims databases based on machine-learning techniques can improve predictive likelihood of AATD among COPD patients. Our model yielded high predictive power for AATD diagnosis with an area under the curve of the receiver operating characteristic curve of 0.9. Given these results, the next step will be to run the model in real world population demonstration projects. If validated, healthcare practitioners could then target individual patients who could benefit from AATD therapy, family testing of at-risk individuals, or lifestyle modifications (e.g., smoking cessation for lung disease and weight loss for liver disease).¹⁴⁻¹⁶

This model used common patient features that are almost universally collected at the patient level, as identified through ICD codes, generic drug name, and lab components. Features that were among the top predictive features for classifying AATD-positive patients included symptoms and comorbidities like shortness of breath and emphysema,³¹ as well as medications for COPD-related symptoms (albuterol sulfate and doxycycline).^{32,33} In contrast, nondependent tobacco use disorder, hypertension, and older age were features associated with a greater likelihood of not having AATD. In a systematic literature review by Miravittles and colleagues, exposure to tobacco smoke was shown to impact the severity of lung disease among patients with AATD.³¹ In one study, AATD patients with COPD had lower rates of cardiovascular and related comorbidities (e.g.,

hypertension) compared to patients with COPD alone.³⁴ Another study reported higher rates of chronic pulmonary heart disease in COPD patients with AATD.⁵ Further, features that may impact disease progression/prognosis among AATD-positive patients with COPD may not necessarily be key features in identifying AATD-positive patients from a broader population of patients with COPD. In other words, diagnostic associations may differ from features that may drive AATD progression.

In rare diseases, ROC curves for positive and negative predictive value require additional considerations when these diagnostic algorithms are transitioned to clinical practice. The current model generated 82% specificity and 82% sensitivity at a threshold of 0.49. Assuming 1% of patients with COPD have AATD, for every 1,000 patients suggested by the algorithm, an 82% capture would find 8 out of 10 individuals with AATD. However, 18% of the patients without AATD (990 patients) would require 178 negative tests. Therefore, the algorithm would improve testing yields from 1% to 4.3% (i.e., $8/[178+8]$). Therefore, higher specificity at the cost of lower sensitivity is important in rare disease testing. The value of predictive models is the ability to make decisions that complement the capability of any hospital system. Further, the ability to find carrier genotypes would identify families in which further genetic testing could occur for COPD or individuals with liver disease.

As opposed to prior strategies that identified potential AATD-positive patients using electronic medical record and electronic patient messages data,²² our model included non-AATD-specific lab test results, and some turned out to be critical predictors of being AATD positive. However, not all laboratory tests were found useful. Some, like normal bilirubin, abnormal sodium, abnormal chloride, medium glucose, were only moderately predictive of being AATD positive. Interestingly,

there were several laboratory test-features that were less predictive of being AATD positive regardless of their level (i.e., hemoglobin, ferritin, and ceruloplasmin).

The reason that some laboratory tests were included in the final model, while others did not, is also nuanced. Bilirubin levels that were within the lower third of the expect range were associated with being AATD negative. One would reason that this finding may reflect differences in subtle liver disease in adults with or without AATD. However, there is likely no impact on calcium levels from AATD lung or liver disease. The care providers ordering these tests ordered them for a reason that might have included having a complicated patient who needed a comprehensive metabolic panel. Therefore, the model reflects both objective and subjective aspects of the healthcare practitioner's mindset and the healthcare systems in which they live.

There are limitations to this project. There is no large, real-world database in which the AATD genetic status of all participants is known. Further, the present model was solely informed by data obtained from patients with COPD. The exclusion of patients with AATD who have independent liver involvement may have contributed to the lower rates of liver disease among AATD-positive individuals, and could reduce the generalizability of results to the minority of patients without apparent lung involvement. Therefore, there are inherent biases associated with the prediction of AATD patients with comorbid COPD.

Additionally, the current algorithm was designed to find carriers of one variant allele. Although ICD-10 codes are available for carrier status individuals, they are rarely used and cannot improve this limitation. Some of the ICD codes used for the project will have been placed by healthcare providers for genetic testing that ultimately showed normal genetics. Others would be placed to identify carriers in the medical record, a useful endeavor to remind clinicians of the increased risk

of lung and liver disease associated with the carrier state. Because other symptomatic individuals with lung or liver disease can reside in these families, cascade testing can follow recognition of an AAT carrier. Lastly, most patients were first captured during 2018-2019 ($n = 16,329$), with relatively few captured during 2020-2022 ($n = 3,636$), with the latter time period potentially reflecting lower rates of testing or medical visits during the COVID-19 pandemic.

Conclusions

The use of predictive modelling based on machine-learning algorithms and RWD shows promise in assisting physicians in identifying COPD patients who may be at high risk of having AATD. It holds promise for identifying which COPD patients are good candidates for genetic testing. In turn, reducing the number of genetic tests required to identify an AATD patient will reduce the duration of the undiagnosed disease journey. Ultimately, identifying at-risk patients earlier through targeted testing can result in slowing disease progression and improved genetic testing in the remainder of the family. External validation of this model is warranted to assess its generalizability to other real-world settings including among non-COPD populations.

Acknowledgments:

All authors participated in and contributed to study design, data interpretation, and manuscript review and preparation. Data development and analyses was primarily performed by authors associated with the Department of Data & Analytics at Eversana. Results of this model were previously presented at the CHEST Annual Conference, 2022, Nashville TN., as an abstract and poster. Joshua Gauci and Neeti Vashi (Eversana) are acknowledged for medical writing assistance.

Data availability:

The individual, de-identified participant data are available from Eversana. The statistical analysis plan is provided in the text. Additional information can be obtained from the corresponding author upon reasonable request.

Declaration of Interest:

All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship: all had full access to all the data, take responsibility for the study's integrity and accuracy of analyses, and have given their approval for its publication. Joshua Gauci and Neeti Vashi (Eversana) provided medical writing assistance.

Charlie Strange has grants paid to the Medical University of South Carolina from Dicerna, Grifols, Krystal, and Takeda and is a consultant to AstraZeneca, Vertex, and Wave Life Sciences in alpha-1 antitrypsin deficiency. He is a medical director at AlphaNet.

M Chris Runken, Thomas Ardiles and Michael Calloway are employees of Grifols, SSNA, Inc.

Rahul Dhakne, Omnya El Massad, Daniel Pfeffer and Pulkit Sehgal are current or former employees of Eversana Life Sciences, Inc.

References

1. Brantly M, Campos M, Davis AM, et al. Detection of alpha-1 antitrypsin deficiency: the past, present and future. *Orphanet J Rare Dis.* Apr 19 2020;15(1):96. doi:10.1186/s13023-020-01352-5
2. Gurevich S, Daya A, Da Silva C, Girard C, Rahaghi F. Improving Screening for Alpha-1 Antitrypsin Deficiency with Direct Testing in the Pulmonary Function Testing Laboratory. *Chronic Obstr Pulm Dis.* Apr 27 2021;8(2):190-197. doi:10.15326/jcopdf.2020.0179
3. Ghosh AJ, Hobbs BD. Recent advancements in understanding the genetic involvement of alpha-1 antitrypsin deficiency associated lung disease: a look at future precision medicine approaches. *Expert Rev Respir Med.* Feb 2022;16(2):173-182. doi:10.1080/17476348.2022.2027755
4. O'Brien ME, Murray G, Gogoi D, et al. A Review of Alpha-1 Antitrypsin Binding Partners for Immune Regulation and Potential Therapeutic Application. *Int J Mol Sci.* Feb 23 2022;23(5)doi:10.3390/ijms23052441
5. Sandhaus R, Strange C, Stone G, Runken MC, Blanchette CM, Howden R. Comorbidity Associations with AATD Among Commercially Insured and Medicare Beneficiaries with COPD in the US. *Int J Chron Obstruct Pulmon Dis.* 2020;15:2389-2397. doi:10.2147/COPD.S263297
6. Stoller JK, Brantly M. The challenge of detecting alpha-1 antitrypsin deficiency. *COPD.* Mar 2013;10 Suppl 1:26-34. doi:10.3109/15412555.2013.763782
7. Nakanishi T, Forgetta V, Handa T, et al. The undiagnosed disease burden associated with alpha-1 antitrypsin deficiency genotypes. *Eur Respir J.* Dec 2020;56(6)doi:10.1183/13993003.01441-2020

8. American Thoracic S, European Respiratory S. American Thoracic Society/European Respiratory Society statement: standards for the diagnosis and management of individuals with alpha-1 antitrypsin deficiency. *Am J Respir Crit Care Med*. Oct 1 2003;168(7):818-900. doi:10.1164/rccm.168.7.818
9. Sandhaus RA, Turino G, Brantly ML, et al. The Diagnosis and Management of Alpha-1 Antitrypsin Deficiency in the Adult. *Chronic Obstr Pulm Dis*. Jun 6 2016;3(3):668-682. doi:10.15326/jcopdf.3.3.2015.0182
10. Ashenurst JR, Nhan H, Shelton JF, et al. Prevalence of Alpha-1 Antitrypsin Deficiency, Self-Reported Behavior Change, and Health Care Engagement Among Direct-to-Consumer Recipients of a Personalized Genetic Risk Report. *Chest*. Feb 2022;161(2):373-381. doi:10.1016/j.chest.2021.09.041
11. Tejwani V, Nowacki AS, Fye E, Sanders C, Stoller JK. The Impact of Delayed Diagnosis of Alpha-1 Antitrypsin Deficiency: The Association Between Diagnostic Delay and Worsened Clinical Status. *Respir Care*. Aug 2019;64(8):915-922. doi:10.4187/respcare.06555
12. Meischl T, Schmid-Scherzer K, Vafai-Tabrizi F, et al. The impact of diagnostic delay on survival in alpha-1-antitrypsin deficiency: results from the Austrian Alpha-1 Lung Registry. *Respir Res*. Jan 27 2023;24(1):34. doi:10.1186/s12931-023-02338-0
13. Barjaktarevic I, Campos M. Management of lung disease in alpha-1 antitrypsin deficiency: what we do and what we do not know. *Ther Adv Chronic Dis*. 2021;12_suppl:20406223211010172. doi:10.1177/20406223211010172
14. Rahaghi FF. Alpha-1 antitrypsin deficiency research and emerging treatment strategies: what's down the road? *Ther Adv Chronic Dis*. 2021;12_suppl:20406223211014025. doi:10.1177/20406223211014025

15. Carpenter MJ, Strange C, Jones Y, et al. Does genetic testing result in behavioral health change? Changes in smoking behavior following testing for alpha-1 antitrypsin deficiency. *Ann Behav Med.* Feb 2007;33(1):22-8. doi:10.1207/s15324796abm3301_3
16. Clark VC, Marek G, Liu C, et al. Clinical and histologic features of adults with alpha-1 antitrypsin deficiency in a non-cirrhotic cohort. *J Hepatol.* Dec 2018;69(6):1357-1364. doi:10.1016/j.jhep.2018.08.005
17. Group TA--ADRS. Survival and FEV1 decline in individuals with severe deficiency of alpha1-antitrypsin. The Alpha-1-Antitrypsin Deficiency Registry Study Group. *Am J Respir Crit Care Med.* Jul 1998;158(1):49-59. doi:10.1164/ajrccm.158.1.9712017
18. Rahaghi F, Monk R, Ramakrishnan V, Strange C. American Thoracic Society International Conference Abstract: Alpha-1 Antitrypsin Augmentation Improves Survival In Severely Deficient Individuals With Low Lung Function. Available at: https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2014.189.1_MeetingAbstracts.A5794?role=tab. 2014;
19. McElvaney NG, Burdon J, Holmes M, et al. Long-term efficacy and safety of alpha1 proteinase inhibitor treatment for emphysema caused by severe alpha1 antitrypsin deficiency: an open-label extension trial (RAPID-OLE). *Lancet Respir Med.* Jan 2017;5(1):51-60. doi:10.1016/S2213-2600(16)30430-1
20. Chapman KR, Burdon JG, Piitulainen E, et al. Intravenous augmentation treatment and lung density in severe alpha1 antitrypsin deficiency (RAPID): a randomised, double-blind, placebo-controlled trial. *Lancet.* Jul 25 2015;386(9991):360-8. doi:10.1016/S0140-6736(15)60860-1

21. Fraughen DD, Ghosh AJ, Hobbs BD, et al. Augmentation Therapy for Severe Alpha-1 Antitrypsin Deficiency Improves Survival and Is Decoupled from Spirometric Decline - A Multi-National Registry Analysis. *Am J Respir Crit Care Med*. Aug 25 2023;doi:10.1164/rccm.202305-0863OC
22. Lam SW, Strange C, Brantly ML, Stoller JK. A Novel Detection Method to Identify Individuals with Alpha-1 Antitrypsin Deficiency: Linking Prescription of COPD Medications with the Patient-Facing Electronic Medical Record. *Chronic Obstr Pulm Dis*. Jan 27 2022;9(1):26-33. doi:10.15326/jcopdf.2021.0260
23. Rao S, Disilvio B, Dumont T. Alpha-1 antitrypsin deficiency: are we testing enough? *Chest*. 2021;160(4)(Supplement A1789)
24. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*. 2019/06/19 2019;6(1):54. doi:10.1186/s40537-019-0217-0
25. Goto T, Jo T, Matsui H, Fushimi K, Hayashi H, Yasunaga H. Machine Learning-Based Prediction Models for 30-Day Readmission after Hospitalization for Chronic Obstructive Pulmonary Disease. *COPD*. Dec 2019;16(5-6):338-343. doi:10.1080/15412555.2019.1688278
26. Meng L, Treem W, Heap GA, Chen J. A stacking ensemble machine learning model to predict alpha-1 antitrypsin deficiency-associated liver disease clinical outcomes based on UK Biobank data. *Sci Rep*. Oct 11 2022;12(1):17001. doi:10.1038/s41598-022-21389-9
27. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Available at: <https://arxiv.org/abs/1603.02754>. 2016;

28. Lundberg SM, Lee S. *Advances in Neural Information Processing Systems 30 (NIPS 2017): A Unified Approach to Interpreting Model Predictions*. 2017.
29. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell*. Jan 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9
30. Strange C, Stoller JK, Sandhaus RA, Dickson R, Turino G. Results of a survey of patients with alpha-1 antitrypsin deficiency. *Respiration*. 2006;73(2):185-90. doi:10.1159/000088061
31. Miravittles M, Herepath M, Priyendu A, et al. Disease burden associated with alpha-1 antitrypsin deficiency: systematic and structured literature reviews. *Eur Respir Rev*. Mar 31 2022;31(163)doi:10.1183/16000617.0262-2021
32. Bollmeier SG, Hartmann AP. Management of chronic obstructive pulmonary disease: A review focusing on exacerbations. *Am J Health Syst Pharm*. Feb 7 2020;77(4):259-268. doi:10.1093/ajhp/zxz306
33. Wang Y, Pera V, Boezen HM, et al. Real-World Effects of Antibiotic Treatment on Acute COPD Exacerbations in Outpatients: A Cohort Study under the PharmLines Initiative. *Respiration*. 2022;101(6):553-564. doi:10.1159/000520884
34. Fahndrich S, Biertz F, Karch A, et al. Cardiovascular risk in patients with alpha-1-antitrypsin deficiency. *Respir Res*. Sep 15 2017;18(1):171. doi:10.1186/s12931-017-0655-1

Table 1: Summary of Pathway to Identify AATD-positive Patients

Criterion	AATD-positive patients (n = 13,585) n (%)
Patients with ≥ 2 AATD diagnosis codes in claims	8,097 (59.6%)
Patients with diagnosis codes in EHR	4,103 (30.2%)
Patients with a positive laboratory test result	
AAT serum level	199 (1.46%)
Genotype testing	71 (0.5%)
Patients treated with AATD-specific therapy	4,231 (31.1%)

Abbreviations: AATD = alpha-1 antitrypsin deficiency; COPD = chronic obstructive pulmonary disease; EHR = electronic health record.

Table 2: Patient Cohort Demographics

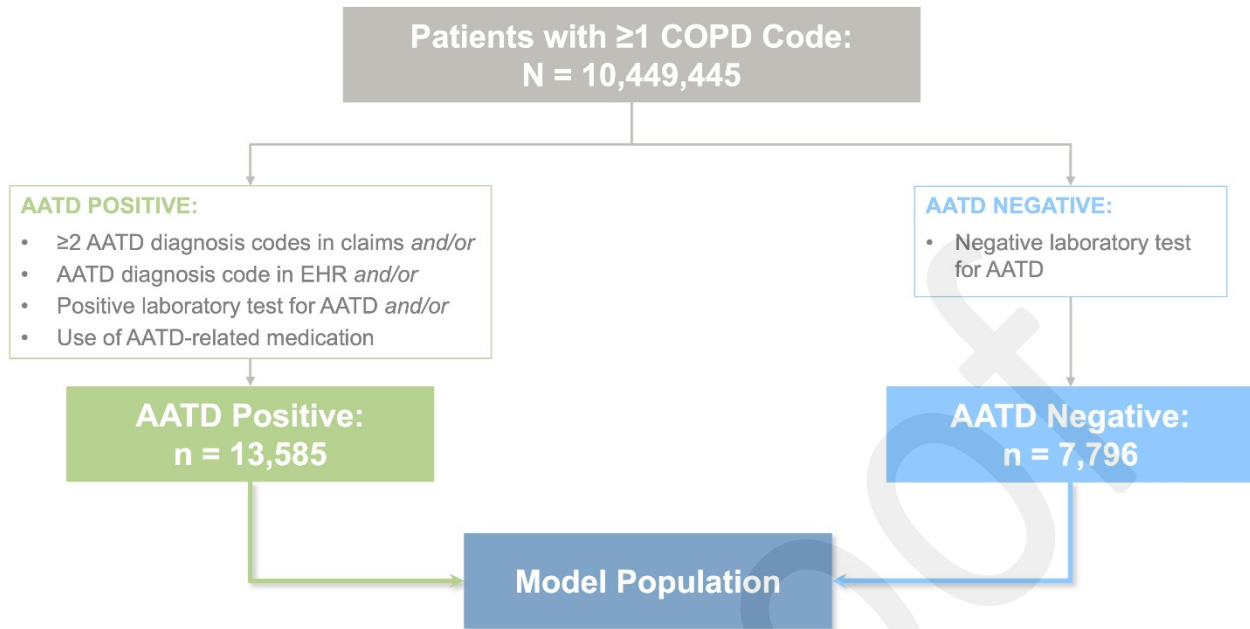
Parameter	AATD-positive patients (n = 13,585)	AATD-negative patients (n = 7,796)
Median age at COPD diagnosis, years	63	63
Median history prior to COPD diagnosis, months	1	7
Median post-COPD follow-up, months	28	26
Year of initial capture or first observed COPD diagnosis, n (%)		
2022	5 (0.04%)	11 (0.14%)
2021	570 (4.20%)	511 (6.56%)
2020	1,394 (10.26%)	1,145 (14.70%)
2019	3,271 (24.08%)	2,155 (27.66%)
2018 ^a	7,691 (56.61%)	2,942 (37.76%)
≤2017	654 (4.81%)	1,027 (13.18%)
Gender, n (%)		
Female	7,640 (56.2%)	4,548 (58.3%)
Male	5,945 (43.8%)	3,248 (41.7%)
Region, n (%)		
North Central	4,275 (31.47%)	1,209 (15.51%)
Northeast	2,389 (17.59%)	1,652 (21.19%)
South	5,170 (38.06%)	3,826 (49.08%)
West	1,734 (12.76%)	1,106 (14.19%)
Unknown	17 (0.13%)	3 (0.04%)
Other pulmonary diagnoses, n (%)		
Asthma	4,108 (30.24%)	2,895 (37.13%)
Bronchiectasis	1,288 (9.48%)	394 (5.05%)
Chronic bronchitis	2,213 (16.29%)	1,812 (23.24%)
Emphysema	6,480 (47.70%)	1,981 (25.41%)
Comorbidities, n (%)		
Cardiac arrhythmias	2,178 (16.03%)	1,822 (23.37%)
Pulmonary circulation disease	2,084 (15.34%)	1,095 (14.05%)
Liver disease	3,024 (22.26%)	2,475 (31.75%)

^a Note: For AATD-positive patients who were identified using the open-claims database, data extraction was between 2018 and 2022; as such, the large number of AATD-positive patients discovered in 2018 includes those diagnosed prior to 2018 who remained in the health plan.

Abbreviations: AATD = alpha-1 antitrypsin deficiency; COPD = chronic obstructive pulmonary disease.

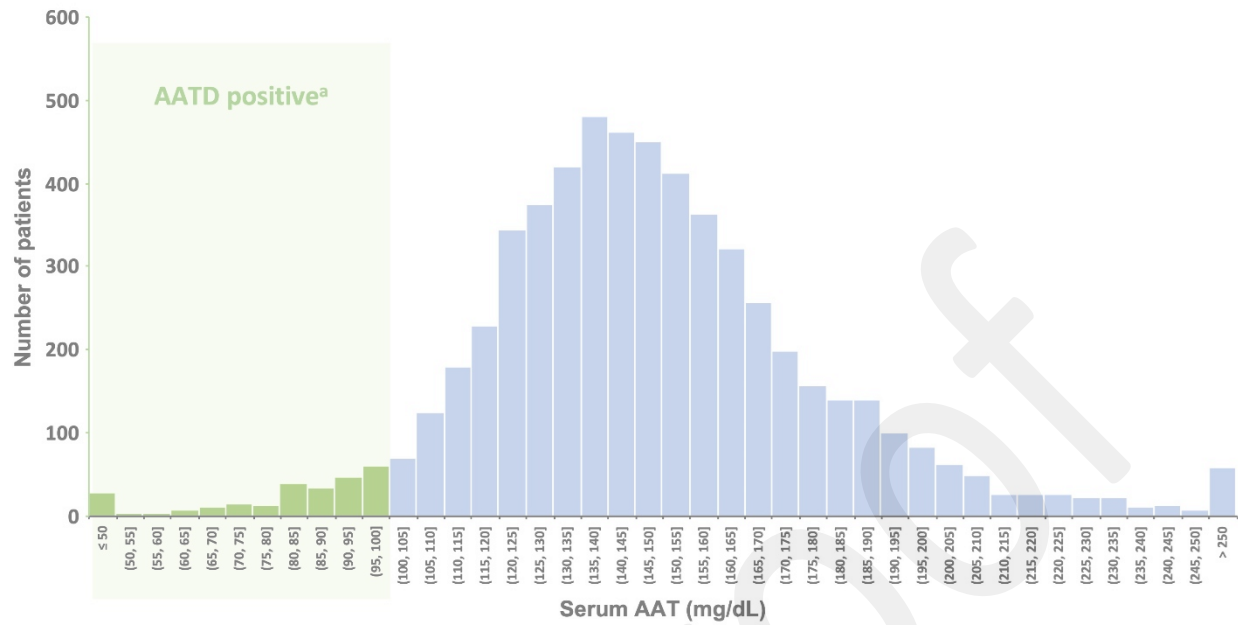
Pre-proof

Figure 1: Cohort Definition and Exclusion Rules



Abbreviations: AATD = alpha-1 antitrypsin deficiency; COPD = chronic obstructive pulmonary disease; EHR = electronic health record.

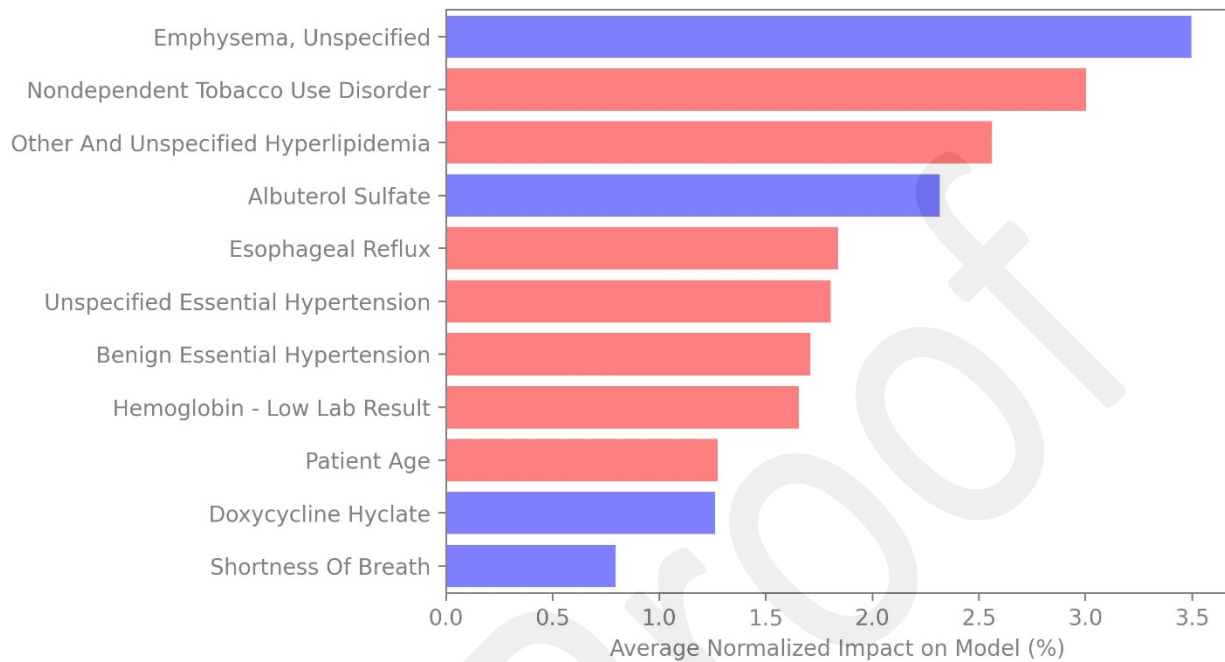
Figure 2: Distribution of Serum AAT Levels in the Model Population



Legend : green bars indicate AATD-positive patients; blue bars indicate AATD-negative patients.

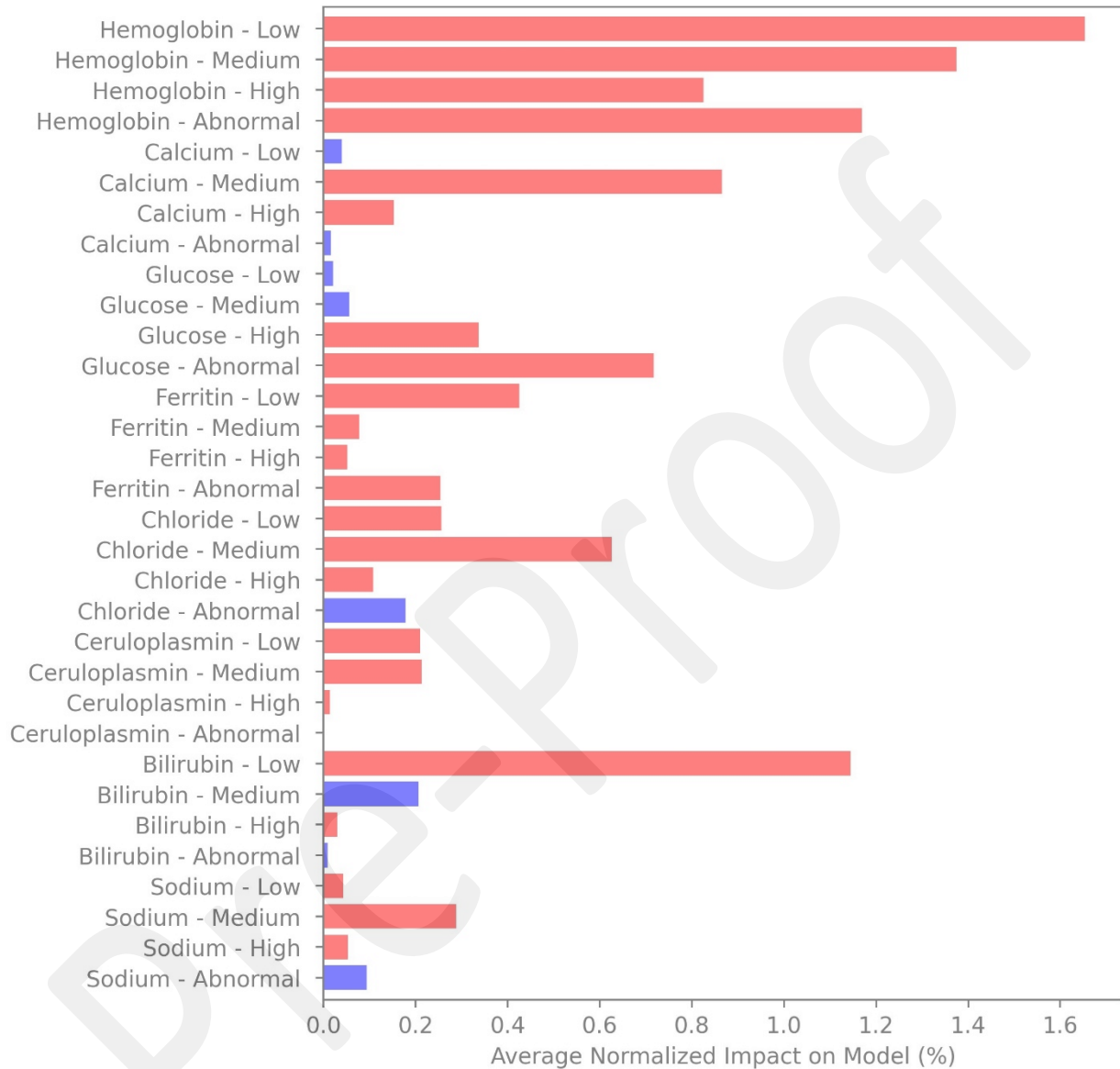
Note: ^a Due to variability in laboratory testing, the reference range for the detection of abnormal serum AAT levels from the EHR was generally <90 mg/dL, but some laboratories used <100 mg/dL.

Figure 3: Select List of Important Features Stratified by Contribution to the Predictive Model



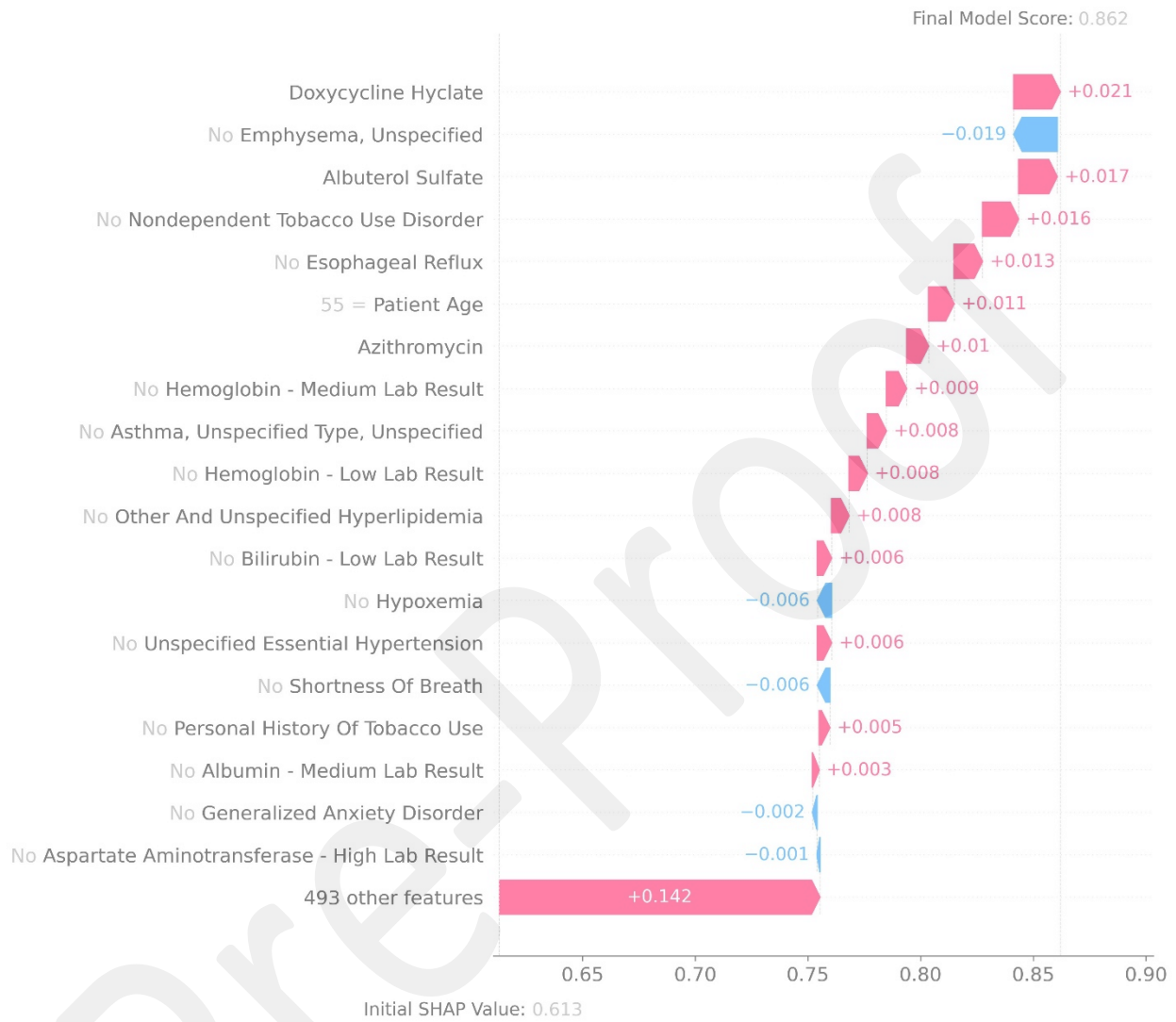
Note: Blue bars denote features that are more likely to predict AATD-positive disease, whereas red bars denote features that are less likely to predict AATD-positive disease (ie, lowers the score for AATD-positive disease given by the model for each patient).

Figure 4: Select List of Laboratory Test Results Stratified by Contribution to the Predictive Model



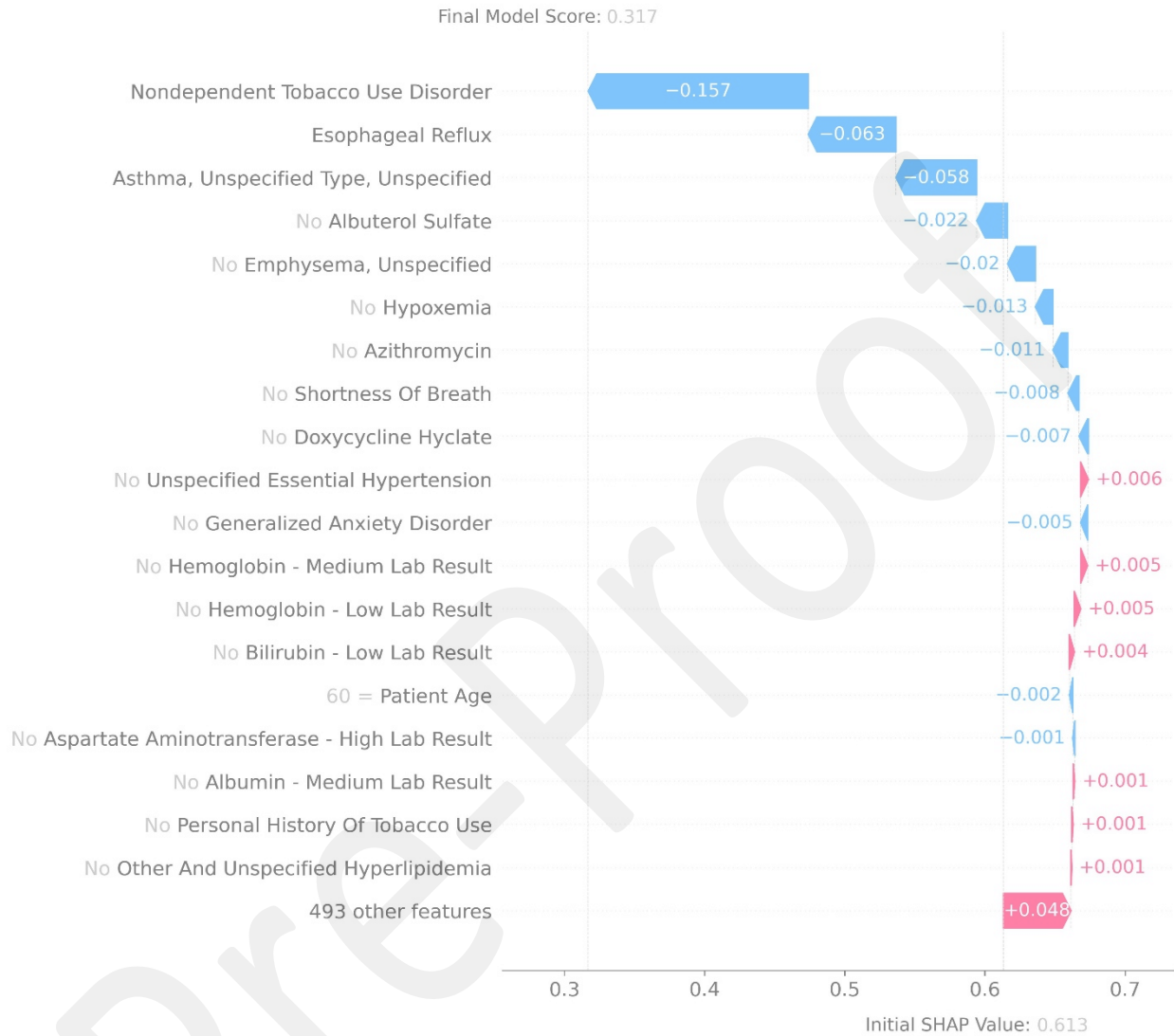
Note: Blue bars denote features that are more likely to predict AATD-positive disease, whereas red bars denote features that are less likely to predict AATD-positive disease (ie, lowers the score for AATD-positive disease given by the model for each patient).

Figure 5: SHAP Feature Importance Metrics for a Select Likely AATD-Positive Patient



Note: Positive values are denoted in pink and reflect greater probability of being AATD positive; conversely, negative values are depicted in blue and represent greater probability of being AATD negative.

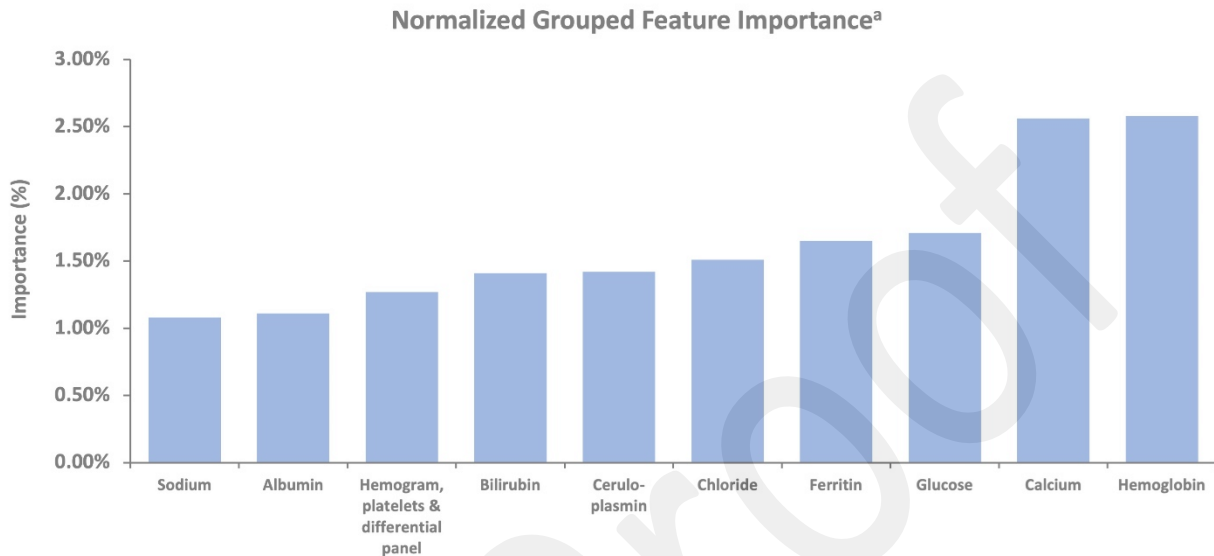
Figure 6: SHAP Feature Importance Metrics for a Select Likely AATD-Negative Patient



Note: Positive values are denoted in pink and reflect greater probability of being AATD positive; conversely, negative values are depicted in blue and represent greater probability of being AATD negative.

Online Supplement

Supplementary Figure 1: Select List of Important Laboratory Features Included in the Predictive Model



^a Feature importance is based on the average information gain for a given feature weighted by the total importance of all included features. Feature importance was grouped at the component level for laboratory test results, instead of the laboratory test component and resulting tertile scoring used in the model.

Supplementary Figure 2: Model Evaluation in the Testing Group

