Original Contribution

Development of Predictive Models for Airflow Obstruction in Alpha-1-Antitrypsin Deficiency


Initially submitted January 23, 2009; accepted for publication June 25, 2009.

Alpha-1-antitrypsin deficiency is a genetic condition associated with severe, early-onset chronic obstructive pulmonary disease (COPD). However, there is significant variability in lung function impairment among persons with the protease inhibitor ZZ genotype. Early identification of persons at highest risk of developing lung disease could be beneficial in guiding monitoring and treatment decisions. Using a multicenter, family-based study sample (2002–2005) of 372 persons with the protease inhibitor ZZ genotype, the authors developed prediction models for forced expiratory volume in 1 second (FEV1) and the presence of severe COPD using demographic, clinical, and genetic variables. Half of the data sample was used for model development, and the other half was used for model validation. In the training sample, variables found to be predictive of both FEV1 and severe COPD were age, sex, pack-years of smoking, bronchodilator responsiveness, chronic bronchitis symptoms, and index case status. In the validation sample, the predictive model for FEV1 explained 50% of the variance in FEV1, and the model for severe COPD exhibited excellent discrimination ($c$ statistic $= 0.88$).

Abbreviations: AAT, α1-antitrypsin; COPD, chronic obstructive pulmonary disease; GSTP1, glutathione S-transferase P1; FEV1, forced expiratory volume in 1 second; IL10, interleukin-10; NOS3, nitric oxide synthase 3; SNP, single nucleotide polymorphism; TNF, tumor necrosis factor.

Alpha-1-antitrypsin (AAT) deficiency is a genetic disorder associated with increased risk of chronic obstructive pulmonary disease (COPD). It is estimated that 1%–2% of COPD cases in the United States are due to AAT deficiency, which most commonly results from the inheritance of ZZ deficiency alleles in the protease inhibitor gene (SERPINA1) (1). Respiratory disease is the leading cause of death in persons diagnosed with AAT deficiency. However, population-based screening studies have demonstrated that most of the approximately 80,000 persons in the United States who carry the protease inhibitor ZZ genotype remain undiagnosed (2). Accurate prediction of the course of lung function decline in affected persons remains a significant challenge.

The natural history of the development of lung disease in AAT deficiency is highly variable (3, 4). This variability is partially but not fully explained by varying levels of exposure to cigarette smoke (5). A number of other clinical and demographic factors have been shown to be associated with forced expiratory volume in 1 second (FEV1) in AAT-deficient persons, and it is likely that modifier genes also contribute to the clinical variability of AAT deficiency (6–8).

Currently, the only specific treatment for COPD in the setting of AAT deficiency is intravenous pooled plasma-derived AAT augmentation, an expensive therapy whose cost per quality-adjusted life-year exceeds traditional cost-benefit thresholds (9). AAT augmentation therapy is typically reserved for persons with evidence of airflow obstruction and/or emphysema. If persons at high risk of lung disease could be prospectively identified, this information could be used to inform treatment and monitoring decisions.

Correspondence to Dr. Dawn L. DeMeo, Channing Laboratory, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115 (e-mail:dawn.demeo@channing.harvard.edu).
Few investigators have evaluated the ability of multivariate models to predict lung function in AAT deficiency. We hypothesized that clinical and demographic variables in combination with genetic marker data from 4 candidate genes would be capable of accurately predicting the pulmonary function of persons with AAT deficiency. To explore this hypothesis, we used multivariable regression to develop predictive models for FEV\textsubscript{1} and the presence of severe COPD, utilizing cross-sectional data from a family-based cohort study of persons with the protease inhibitor ZZ genotype.

**MATERIALS AND METHODS**

**AAT Genetic Modifier Study**

The AAT Genetic Modifier Study was a multicenter, family-based study consisting of adults who were homozygous for the protease inhibitor Z allele. Study participants had at least 1 sibling with the protease inhibitor ZZ genotype who was also enrolled in the study. A detailed description of the first phase of this study, which was performed from 2002 to 2005 and which was used in the current manuscript, has been published previously (6). For this analysis, 6 non-siblings (1 ZZ uncle, 4 ZZ parents, and 1 adult child of ZZ parents) were removed from the data set in order to limit familial relationships to siblings only; this resulted in a total of 372 individuals from 167 families. Each participant completed a modified version of the American Thoracic Society–Division of Lung Disease Respiratory Epidemiology Questionnaire and underwent spirometry testing according to American Thoracic Society standards (10, 11). More detailed information on questionnaire items is available in the Web Appendix, which is posted on the Journal’s Web site (http://aje.oxfordjournals.org/).

**Genotyping**

Candidate genes were selected on the basis of a previously reported association with airflow obstruction in the AAT Genetic Modifier Study (interleukin-10 (IL10) and tumor necrosis factor (TNF)) or on prior reports of an association with FEV\textsubscript{1} in AAT-deficient subjects (glutathione S-transferase P1 (GSTP1) and nitric oxide synthase 3 (NOS3)) (8, 12, 13). Briefly, for IL10 and TNF, we used HapMap data (http://www.hapmap.org/) from a white population to select a set of linkage-disequilibrium tagging single nucleotide polymorphisms (SNPs) with the Tagger program (14), using a pairwise \( r^2 \) cutoff of 0.8. SNPs with a minor allele frequency less than 5% were excluded from the analysis. A total of 17 SNPs were analyzed, comprising 10 SNPs from IL10, 5 SNPs from TNF, 1 SNP from GSTP1, and 1 SNP from NOS3 (see Web Table 1 (http://aje.oxfordjournals.org/)).

Genotyping was performed using the Sequenom platform (Sequenom, San Diego, California), and all data were checked for Mendelian consistency using the PedCheck program (15).

**Predictive model-building methods**

We created 2 predictive models: a linear regression model for prebronchodilator FEV\textsubscript{1} percent predicted and a logistic regression model for the presence of severe COPD. Severe COPD was defined as an FEV\textsubscript{1} less than 50% predicted. More detailed information on model-building methods is available in the Web Appendix (http://aje.oxfordjournals.org/).

The data set was randomly divided into a development sample and a validation sample with a 1:1 ratio between the development and validation samples, using the PROC SURVEYSELECT procedure in SAS, version 9.1.3 (SAS Institute Inc., Cary, North Carolina). Sampling was performed at the family level, resulting in familial correlations within each data set but not between data sets.

We constructed the predictive models by means of linear and logistic regression methods with a stepwise model-building approach, using an entry and exit criterion of \( P \leq 0.10 \). In addition to meeting the \( P \)-value criterion for model entry, a candidate variable also needed to demonstrate a consistent direction of effect in the development and validation samples in order to be included in the final model. We used generalized estimating equations methods to adjust for familial correlations. An initial set of variables, not including genetic polymorphisms, was selected on the basis of previously reported epidemiologic associations with FEV\textsubscript{1} (6, 16–19). These variables included age, sex, pack-years of cigarette smoking, bronchodilator responsiveness, chronic bronchitis symptoms, history of an asthma diagnosis before age 30 years, history of a pneumonia diagnosis before age 30 years, and index case status. We individually tested a series of interaction terms, using the same inclusion/exclusion thresholds, in order to assess age and sex interactions with other model variables.

Using model coefficients derived from the development sample, we quantified each model’s predictive performance in the validation sample. Discrimination was assessed by calculating \( r^2 \) and the \( c \) statistic for linear and logistic models, respectively, and calibration was assessed graphically for the logistic regression models.

In order to identify any additional contribution from incorporating data on genetic variation in IL10, TNF, GSTP1, and NOS3 into the predictive models, we tested each SNP and SNP-by-smoking interaction term sequentially in the models. We also tested for SNP-SNP interactions between genes by sequentially including a multiplicative interaction term for each possible SNP-SNP second-order interaction. We also tested for haplotype effects within a gene by performing haplotype analysis for the genes in which multiple markers were tested (IL10 and TNF). Haplotype analysis was performed using all typed markers within the gene of interest, as well as with adjacent 3-SNP sliding windows.

We performed power calculations with the Quanto program, assuming an additive genetic model, a 2-sided \( P \) value of 0.10, and a range of minor allele frequencies and allelic effect sizes (20). A power chart is presented in Web Figure 1 (http://aje.oxfordjournals.org/). Our study had 80% power to detect polymorphisms with odds ratios of 1.75 or higher for severe COPD or polymorphisms that explained 4% or more of the variance in FEV\textsubscript{1}. For odds ratios of 1.25 and polymorphisms that explained 1% of the FEV\textsubscript{1} variance, power was 40%.

RESULTS

Table 1 shows the baseline characteristics of the entire cohort of 372 study participants and the development and validation samples. The study population had a median prebronchodilator \( \text{FEV}_1 \) percent predicted of 58.2%, and 41% \((n = 154)\) of the subjects with the protease inhibitor ZZ genotype were identified as having severe COPD. While the prevalence of prior cigarette smoking was high, the median amount of tobacco exposure was relatively low, and very few participants were current smokers. The development and validation data samples had statistically significant differences in sex distribution and the ratio of \( \text{FEV}_1 \) to forced vital capacity.

The relation between smoking and \( \text{FEV}_1 \) was a nonlinear relation characterized by a steep decrease in \( \text{FEV}_1 \) associated with the first 20 pack-years (i.e., 19.9 pack-years) of personal tobacco exposure, followed by a less steep decline in \( \text{FEV}_1 \) associated with pack-years of 20 or greater (Figure 1). We therefore created 2 piecewise-linear variables for smoking with a breakpoint at 20 pack-years and included these variables in the model-building procedure. Use of these 2 variables allowed for the assignment of 2 different rates of \( \text{FEV}_1 \) decline, 1 for the first 20 pack-years of smoking and another for subsequent pack-year exposure. Only the variable capturing the first 20 pack-years of personal smoking exposure was selected into the models for \( \text{FEV}_1 \) and severe COPD, indicating that in our data set the first 20 pack-years of personal tobacco exposure seemed to most significantly affect lung function.

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Relation between forced expiratory volume in 1 second (\( \text{FEV}_1 \)) and pack-years of cigarette smoking in 231 smokers with the protease inhibitor ZZ genotype, AAT Genetic Modifier Study, 2002–2005. The figure shows a scatterplot with a smoothing curve demonstrating the trend in \( \text{FEV}_1 \) percent predicted with increasing pack-years of tobacco exposure, with a linear trend from 0 pack-years to 19.9 pack-years. Nonsmoking subjects with the protease inhibitor ZZ genotype were not included. AAT, \( \alpha_1 \)-antitrypsin.
Tables 2 and 3 show results from the model-building process for FEV₁ and severe COPD. Estimates of the effect sizes in the validation sample are included in the tables to allow for comparison of effect sizes between the development and validation sets; however, in our assessment of model performance in the validation sample, we used the β coefficients obtained from the development sample. The main-effects variables that were included in both models were age, sex, the first 20 pack-years of personal smoking exposure, bronchodilator responsiveness, symptoms of chronic bronchitis, and index case status. A diagnosis of pneumonia before age 30 years, chronic bronchitis, and index case status. A diagnosis of asthma diagnosis prior to age 30 years (0 = no, 1 = yes) was selected for the FEV₁ model for severe COPD. The age-by-pack-years interaction term was also associated with a protective effect, but this failed to reach statistical significance in the validation sample. Because the age-by-pack-years interaction was a significant predictor of FEV₁, we performed a sensitivity analysis excluding persons over age 60 years in order to assess the impact of age on the estimated model coefficients; the effect was minimal.

The results of testing for SNP main effects, SNP-by-smoking interactions, and SNP-SNP interactions are presented in Web Tables 2, 3, and 4 (http://aje.oxfordjournals.org/). When we tested SNPs and SNP-by-pack-year interaction terms for IL10, TNF, GSTP1, and NOS3 sequentially in the predictive models, 7 SNPs in IL10 and the NOS3 SNP rs1799983 were associated with FEV₁ or severe COPD in 1 of the models. However, none of these selected SNPs or SNP-by-smoking interactions met the model inclusion criteria of significant association in both samples in a consistent direction of effect.

When we tested for SNP-SNP interactions, 1 second-order SNP-SNP interaction between rs3024492 in IL10 and rs800610 in TNF met the threshold for model inclusion in the severe COPD model but not the FEV₁ model. Because this interaction was included in the severe COPD model, we also included each SNP as a main effect as well. Haplotype analysis performed in IL10 and TNF using all typed markers and 3-SNP sliding windows did not identify any haplotypes significantly associated with FEV₁ or severe COPD in both samples.
The predictive model for FEV₁ explained 50% of the variance in FEV₁ in the validation sample, a decrease of 4 percentage points from performance in the development sample. All model variables that were significantly associated with FEV₁ in the derivation sample were also associated with FEV₁ in the validation sample, with relatively stable effect-size estimates. The following variables were associated with lower levels of FEV₁: age, male sex, first 20 pack-years of smoking, and index case status. The interaction between age and the first 20 pack-years of smoking was associated with higher levels of FEV₁ as the value of the interaction term increased.

The predictive model for severe COPD had a statistic of 0.88 in the validation sample, a decrease from 0.91 in the development sample. All of the variables selected into the model in the development sample also met the threshold for significance in the validation sample, except for the age-by-pack-years interaction term. These variables were the same as those for the FEV₁ model, with the exception of pneumonia diagnosis and the age-by-pack-years interaction term. Figure 2 shows the receiver operator characteristic curve representing the predictive performance of the model in the validation data set. Model calibration is depicted in Figure 3, demonstrating that the model’s predictive probabilities were closely calibrated to the observed likelihood of severe COPD across prediction quartiles.

DISCUSSION

In a family-based cohort of 372 persons with the protease inhibitor ZZ genotype whose pulmonary function spanned a broad spectrum, we developed and tested predictive models for FEV₁ and the presence of severe COPD by incorporating demographic, clinical, and genetic information. In validation testing, these models explained 50% of the variance in FEV₁ and had a c statistic of 0.88 in models for severe COPD, a level of performance that matches or exceeds other, well-established clinical prediction tools (21, 22). The predictive variables common to both models were age, sex, first 20 pack-years of personal smoking exposure, bronchodilator responsiveness, the presence of chronic bronchitis symptoms, and index case status. Two interacting loci in *IL10* (rs3024492) and *TNF* (rs1800610) met the criteria for inclusion in the severe COPD model and yielded a marginal improvement in predictive performance (the c statistic increased from 0.86 to 0.88). When we tested for main effects and SNP-by-smoking interactions, none of the genetic markers in the 4 studied genes were incrementally informative in predicting FEV₁ or severe COPD.

In previous analyses from the AAT Genetic Modifier study cohort, researchers identified associations between FEV₁ and sex, asthma, chronic bronchitis, pneumonia, pack-years of personal cigarette smoking, and genetic polymorphisms in *IL10* and *TNF* (6, 12). As expected, results from our models...
corroborate the findings from previous analyses of this cohort, with the exception that asthma and the genetic polymorphisms in IL10 and TNF were not significant predictors in our models when analyzed as main effects. In the case of asthma, bronchodilator responsiveness was a correlated but more powerful predictor of FEV1. Once this variable was included in the model, asthma diagnosis provided no additional predictive information. However, in models without bronchodilator responsiveness, asthma before age 30 years was a significant predictor of FEV1 and severe COPD. Regarding associations with IL10, TNF, GSTP1, and NOS3, none of the polymorphisms analyzed in our study added incremental benefit to our models when analyzed as main effects. This could be due to a true lack of effect, confounding with variables included in the model, or differences in analytic methods between our study and previous studies (8, 12, 13). In addition, splitting our sample into development and validation samples limited our power to identify genetic associations.

A multiplicative interaction between loci in IL10 (rs3024492) and TNF (rs1800610) met the inclusion threshold for the severe COPD model but not the FEV1 model, though the direction of effect in the FEV1 model was consistent with the COPD findings. Both SNPs are located in introns. In the family-based analysis of this sample by DeMeo et al. (12), the TNF SNP but not the IL10 SNP was associated with FEV1. To our knowledge, the IL10 SNP has not been studied in other cohorts for an association with COPD. The TNF SNP has been studied in at least 4 case-control studies; in 1 of them, investigators reported increased COPD risk associated with the minor allele (23), and in the other 3, investigators reported no statistically significant association (24–26). While there is a biologic rationale for a link between AAT, IL10, and TNF (27), our finding should be interpreted cautiously given the low power of our study to detect SNP-SNP interactions and the large number of statistical tests performed. Until this interaction is replicated in larger cohorts, it should be viewed as a tentative association that may be attributable to chance.

Our finding of a nonlinear relation between pack-years and FEV1 may be due to a “floor effect”; namely, once FEV1 levels of 30%–40% predicted are reached, there is simply less lung function to lose and mortality rates are so high that few people survive to manifest lower levels of FEV1. This effect has been previously described in other clinical situations (28). While previous research has demonstrated a linear relation between pack-years of smoking and FEV1, there have been conflicting reports on the effect of a relation between pack-years and FEV1 in AAT deficiency (16, 19, 29). A floor effect in the smoking-FEV1 relation could explain the observation of a linear relation in healthier populations and the lack of a relation in more severely affected populations. For example, in the Swedish national AAT deficiency registry (16), the relation between FEV1 decline and pack-years was found to be dose-dependent, whereas in the more severely obstructed population studied by Hutchison et al. (19), there was no statistically significant relation between FEV1 decline and smoking exposure. Understanding and correctly modeling these relations is crucial to the development of accurate predictive models.
This study had several strengths. First, our data set was characterized by wide variability in lung function, which made it well suited for developing a predictive model. Second, the family-based nature of our study sample mirrored the situation in which a predictive model would probably be used—that is, to predict the lung function of index patients and their newly diagnosed siblings. Third, we were able to test genetic polymorphisms in 4 genes for their potential impact on lung function.

However, our study also had some limitations. First, our study population had only 1 set of spirometric values, so our predictions were for single FEV₁ values rather than longitudinal decline in FEV₁. Our predictive model can provide an initial assessment of disease severity by allowing a comparison of the subject’s observed and predicted FEV₁ values. In addition, using our models, an estimate of an individual’s future FEV₁ level or likelihood of developing severe COPD can be calculated; but this requires the assumption that the older members of the study population are comparable to the younger ones in terms of COPD susceptibility and environmental exposures. A model developed from longitudinal data would not be susceptible to this source of bias.

Second, given the modest size of our study population, we had limited power to detect small effects and to test for interactions between candidate variables. With a larger sample size, it is likely that some of the genetic risk factors, epidemiologic risk factors, and interaction terms that were not included in our models would have reached the levels of significance required for model entry.

Third, while our models exhibited good predictive performance, 50% of the variance in FEV₁ was not explained by our FEV₁ model. Our efforts represent an initial attempt at developing a predictive model for lung function level, and we have demonstrated that a limited number of predictive variables explain a significant proportion of the variability in lung function. Future models might include genetic data from genome-wide association analyses, other interactions between epidemiologic and genetic factors, or important clinical variables not assessed in our models, such as passive smoking exposure.

Fourth, we validated our model by splitting our study population into a development sample and a validation sample. However, data supporting some of our initial variables (asthma, pneumonia, and chronic bronchitis) came from an epidemiologic risk factors, and interaction terms that were not included in our models would have reached the levels of significance required for model entry. Further validation of these models rivals or exceeds that of other, well-established clinical predictive models, but there is still room for improvement in predictive accuracy. Further identification of epidemiologic and genetic risk factors for FEV₁ decline in AAT deficiency, larger sample sizes for model development and validation, and longitudinal data on lung function will allow for the development of more powerful risk prediction tools.

In summary, we have presented results from 2 predictive models for FEV₁ and severe COPD in AAT deficiency. These represent a set of predictive variables that performed well in 2 separate data samples, indicating that they are a stable set of variables with relatively uniform, predictable effects in our study population. The predictive performance of these models rivals or exceeds that of other, well-established clinical predictive models, but there is still room for improvement in predictive accuracy. Further identification of epidemiologic and genetic risk factors for FEV₁ decline in AAT deficiency, larger sample sizes for model development and validation, and longitudinal data on lung function will allow for the development of more powerful risk prediction tools.
This work was supported by a grant from the Alpha-1 Foundation (D. L. D.) and National Institutes of Health grants T32 HS00060 (P. J. C.), F32 HL094035 (P. J. C.), K08 HL072918 (D. L. D.), UL1 RR025752 (P. J. C.), RO1 HL68926 (E. K. S.), and P01 083069 (E. K. S.).

S. R. has participated as a speaker in programs organized by AstraZeneca, Boehringer-Ingelheim, GlaxoSmithKline, Otsuka, and Pfizer. He serves on advisory boards for Altana, AstraZeneca, Dey, GlaxoSmithKline, Novartis, Schering-Plough, and Talecris. He has conducted clinical trials for Almirall, Altana, Astellas, Centocor, GlaxoSmithKline, Nabi, Novartis, and Pfizer. He has served as a consultant for Adams, Almirall, Altana, AstraZeneca, Bend, Biolipox, Centocor, Critical Therapeutics, GlaxoSmithKline, ICOS, Johnson & Johnson, Novartis, Ono, Parengenix, Pfizer, Roche, Sankyo, Sanofi, and Schering-Plough. E. K. S. received an honorarium for a talk on COPD genetics in 2006, grant support for 2 studies of COPD genetics (2004–2009), and consulting fees (2005–2009) from GlaxoSmithKline. He received an honorarium from Wyeth for a talk on COPD genetics given in 2004. He received an honorarium from Bayer for a symposium held at the 2005 meeting of the European Respiratory Society. He received honoraria for talks given in 2007 and 2008 and consulting fees from AstraZeneca in 2008 and 2009.

REFERENCES
