Thorax Online First, published on December 16, 2010 as 10.1136/thx.2010.146118

Impact of non-linear smoking effects on the identification of gene-by-smoking interactions in COPD genetics studies

P J Castaldi,[1,2] D L Demeo,[3] C P Hersh,[3] D A Lomas,[4] I C Soerheim,[5] A Gulsvik,[5] P Bakke,[5] S Rennard,[6] P Pare,[7] J Vestbo,[8,9] AATGM Investigators, ICGN Investigators, E K Silverman[2,3]

► Additional data are published online only. To view these files please visit the journal online (http://thorax.bmj.com).

[1]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts, USA
[2]Channing Laboratory and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital
[3]Harvard Medical School, Boston, Massachusetts, USA
[4]Department of Medicine, University of Cambridge, Cambridge Institute for Medical Research, Cambridge, UK
[5]Haukeland University Hospital, Bergen, Norway
[6]University of Nebraska, Omaha, Nebraska, USA
[7]Division of Respiratory Medicine and the James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research, University of British Columbia, St Paul's Hospital, Vancouver, Canada
[8]Department of Cardiology and Respiratory Medicine, Hvidovre Hospital, Copenhagen, Denmark
[9]Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK

Correspondence to
Peter Castaldi, Tufts Medical Center, 800 Washington St, Box 63, Boston, MA 02111, USA; pcastaldi@tuftsmedicalcenter.org

For author footnote see end of the article.

Received 30 June 2010
Accepted 22 October 2010

ABSTRACT

Background The identification of gene-by-environment interactions is important for understanding the genetic basis of chronic obstructive pulmonary disease (COPD). Many COPD genetic association analyses assume a linear relationship between pack-years of smoking exposure and forced expiratory volume in 1 s (FEV_1); however, this assumption has not been evaluated empirically in cohorts with a wide spectrum of COPD severity.

Methods The relationship between FEV_1 and pack-years of smoking exposure was examined in four large cohorts assembled for the purpose of identifying genetic associations with COPD. Using data from the Alpha-1 Antitrypsin Genetic Modifiers Study, the accuracy and power of two different approaches to model smoking were compared by performing a simulation study of a genetic variant with a range of gene-by-smoking interaction effects.

Results Non-linear relationships between smoking and FEV_1 were identified in the four cohorts. It was found that, in most situations where the relationship between pack-years and FEV_1 is non-linear, a piecewise linear approach to model smoking and gene-by-smoking interactions is preferable to the commonly used total pack-years approach. The piecewise linear approach was applied to a genetic association analysis of the PI*Z allele in the Norway Case–Control cohort and a potential PI*Z-by-smoking interaction was identified ($p=0.03$ for FEV_1 analysis, $p=0.01$ for COPD susceptibility analysis).

Conclusion In study samples of subjects with a wide range of COPD severity, a non-linear relationship between pack-years of smoking and FEV_1 is likely. In this setting, approaches that account for this non-linearity can be more powerful and less biased than the more common approach of using total pack-years to model the smoking effect.

Chronic obstructive pulmonary disease (COPD) is well suited to the study of gene-by-environment interactions since the major environmental risk factor for COPD—cigarette smoking—is known and quantifiable. With the advent of large well-powered genome-wide association studies in COPD, the identification of such interactions may be feasible. However, there are a number of challenges to the identification of gene-by-smoking interactions in COPD: (1) the principal genetic risk factors for COPD are still in the process of being identified; (2) a variety of approaches have been used to model smoking effects; and (3) there is no empirical knowledge of the nature, extent or functional form of gene-by-smoking interactions in COPD.

While cigarette smoking is easily quantifiable in terms of pack-years ((average daily number of cigarettes smoked/20 cigarettes per pack) × years of smoking), previous work has shown that pack-years alone may be an overly simplistic means of modelling smoking exposure, and non-linear relations may be present.[1 2] Many COPD genetic association analyses model smoking effects by including a pack-years term in a regression model, which assumes a linear relation between pack-years and forced expiratory volume in 1 s (FEV_1) or, in the case of logistic regression for COPD status, a linear relation between pack-years and the log odds of having COPD. This practice is supported by seminal work on the decline in FEV_1 in general population samples.[3–5] However, it is not clear that these findings apply to the types of study samples typically assembled for COPD genetic association studies—namely, cross-sectional samples that include subjects with a wide range of lung function impairment including severe disease. In this setting, a number of factors may result in a non-linear relation between pack-years and FEV_1. These factors could include survival bias due to the well-demonstrated association between FEV_1 and mortality,[6] and floor effects resulting from a diminished effect of cigarette smoking at very low levels of FEV_1.

We hypothesised that the relation between FEV_1 and pack-years may be non-linear in study samples with a wide range of airway obstruction and that, in this setting, methods of modelling smoking that account for non-linearity may be more accurate and powerful for detecting gene-by-smoking interactions than the traditionally used pack-years approach. We tested this hypothesis in a cohort in which such non-linear effects had been observed, by simulating a genetic variant with known main effects and gene-by-smoking effects. Finally, we assessed the performance of these modelling approaches in a gene-by-smoking analysis of the alpha-1 antitrypsin (AAT) PI MZ genotype in a case–control sample from Norway.

METHODS

Study samples

We examined the relations between FEV_1 (percentage of predicted) and pack-years of cigarette

smoking in four large study samples (the Alpha-1 Antitrypsin Genetic Modifiers Study; the International COPD Genetics Network; the Boston Early-Onset COPD Study; and the Bergen, Norway Case–Control Study). The recruitment and inclusion criteria for these studies have been reported previously.[7–10] In brief, the Alpha-1 Antitrypsin Genetic Modifiers Study is a family-based study of individuals with the PI*ZZ genotype. The International COPD Genetics Network and the Boston Early-Onset COPD Study are family-based studies in which families were identified through a proband affected with COPD. The Bergen, Norway Case–Control Study is a population-based study with a minimum required level of smoking exposure of 2.5 pack-years for both cases and controls. In each of the four studies, subjects underwent spirometric testing in accordance with American Thoracic Society standards.[11]

Relation of FEV_1 to pack-years

For each of the four studies we generated scatterplots of the relation between FEV_1 and pack-years and drew smoothing curves through the data using a cubic spline fitting routine. All analyses were performed using SAS Version 9.2.

Simulation studies

Using data from the Alpha-1 Antitrypsin Genetic Modifiers Study, we simulated a randomly-assigned biallelic genetic variant in accordance with Hardy–Weinberg proportions. We conducted simulations under multiple scenarios, with each scenario characterised by a particular minor allele frequency, genetic main effect and gene-by-smoking effect on FEV_1 percentage predicted. For each scenario we conducted 1000 simulations. The range of allele frequencies was 10–40%. The main effect of the gene was specified such that each copy of the minor allele decreased FEV_1 percentage predicted by 1 unit, and the gene-by-smoking interaction effect ranged from −0.45 to +0.45 units per allele per pack-year. For comparison, the main effect of pack-years in this dataset (after adjusting for age and sex) was approximately −1 unit per pack-year.

In each simulation we calculated an estimated FEV_1 for each individual based on their observed FEV_1, their simulated genotype and the strength of the simulated genetic main effect and gene-by-smoking interaction effect. In our primary analysis we assumed that the gene-by-smoking interaction effect followed the same non-linear form as the smoking main effect. In each of our analyses, non-smokers were included in the analysis with a value of zero for the pack-years variable. A detailed description of the simulation methods used is included in the online supplement.

Using linear regression we estimated the genetic main effect and gene-by-smoking effect in each simulated dataset. We ran two regression models, one in which the smoking main effect and gene-by-smoking interaction were modelled using the pack-years approach (inclusion of a pack-years term in the regression equation) and another in which these effects were modelled with a piecewise linear approach (inclusion of separate variables to represent distinct intervals of smoking exposure). In each model we adjusted for age and sex in addition to the smoking and genetic variables. We recorded the β coefficients from each model in each simulation and calculated the mean and SD of these values. The bias of the two approaches was quantified by comparing the estimated values of the genetic main effects and gene-by-smoking effects with the actual values, and the power was estimated by recording the number of times each β coefficient was associated with a p value of <0.05.

For the piecewise linear approach we determined a cut-off point for the pack-years variable based on the shape of the relation between pack-years and FEV_1. In the Alpha-1 Antitrypsin Genetic Modifiers Study, which was the basis for these simulations, a cut-off point of 20 pack-years was selected based on visual inspection and improvement in model fit. The model fit of the piecewise linear model was compared with the pack-years model using the F-test. This cut-off point was used to code two variables, with one variable representing the first 20 pack-years of exposure and another variable representing all subsequent pack-years. The interaction term in the piecewise linear model included only the 'piece' that was statistically significantly associated with FEV_1 in a multivariate context; thus, the interaction term was of the following form: first 20 pack-years of smoking × copies of minor allele.

Gene-by-smoking analysis of the PI*Z allele in the Norway Case–Control Study

The two approaches to model smoking were compared in a gene-by-smoking analysis of the PI*Z allele in the Norway Case–Control Study data using regression methods to test for genetic associations with the FEV_1 level and COPD susceptibility (ie, presence or absence of COPD). For the FEV_1 analysis we applied sample weights to correct for oversampling of COPD cases, assuming a 10% prevalence of COPD in the general population. One analysis was performed using the traditional approach of modelling smoking with the pack-years approach and a similar analysis was performed using a piecewise linear approach. Based on inspection and overall model fit for the FEV_1 model, we chose a cut-off point of 40 pack-years for the piecewise linear variable. We tested the main effect of the PI*Z allele as well as the Z-by-smoking interaction.

Alpha-1 antitrypsin typing

Phenotyping for the PI*Z allele in the Norway Case–Control Study was performed by isoelectric focusing. Individuals with severe AAT deficiency (PI*Z, null-null, or SZ) were excluded from the Norway Case–Control Study.

RESULTS

The baseline characteristics of the four study samples are shown in table 1. Each study had significant numbers of individuals with severe airflow obstruction, although the median FEV_1 level varied substantially between studies.

The relation between pack-years of smoking and FEV_1 (percentage of predicted) in each of the study samples is shown in figure 1. In each study sample there was a non-linear relation between FEV_1 and pack-years. For the two study samples in which piecewise linear modelling of smoking was performed (the Alpha-1 Antitrypsin Genetic Modifiers Study and the Norway Case–Control Study), the models with the piecewise linear smoking approach fit the data better than the models with the linear approach ($p<0.001$ in both instances). All of the study samples had a similar pattern of an initial strong negative effect of smoking on FEV_1 level with a subsequent decrease in the negative impact of additional pack-years. With the exception of the Norway study, there seemed to be a plateau phase at which additional pack-years were not associated with a further decline in FEV_1. In all four samples the slope of the FEV_1/pack-years relation decreased at an FEV_1 level of approximately 30–50% of predicted. For three of the samples this corresponded to a smoking exposure of 40–60 pack-years; however, in the more genetically susceptible Alpha-1 Antitrypsin Deficiency cohort, the levelling of the FEV_1/pack-years relation occurred at approximately 20 pack-years exposure.

Table 1 Characteristics of study samples

	AAT	EOCOPD	ICGN	Norway
No of subjects	372	972	3058	1909
Mean (SD) age (years)	52 (10)	46 (18)	58 (8)	61 (11)
Female, n (%)	202 (54)	567 (58)	1374 (45)	847 (44)
Ever smokers, n (%)	231 (62)	659 (68)	3058 (100)	1909 (100)
Median (IQR) pack-years of smoking	5 (0–19)	14 (0–35)	39 (25–55)	23 (13–34)
Median (IQR) FEV_1 (% predicted)	58 (33–93)	84 (60–96)	58 (35–87)	76 (48–90)

AAT, Alpha-1 Antitrypsin Genetic Modifiers Study; EOCOPD, Boston Early-Onset COPD Study; FEV_1, forced e xpiratory volume in 1 s; ICGN, International COPD Genetics Network; Norway, the Bergen, Norway Case–Control Study.

The results of the simulation study are shown in table 2 and in figure 1 in the online supplement. Under most of the simulated scenarios, the piecewise linear approach yielded more accurate estimates of genetic main effect size and gene-by-smoking interactions than the pack-years approach. The direction of bias in the estimates generated by the pack-years approach was consistent with that expected from an approach that does not fully account for the strength of the gene-by-smoking interaction. When the genetic main effects and gene-by-smoking interaction effects were in the same direction (ie, both main effect and interaction effect were negative), modelling with pack-years systematically overestimated the magnitude of the genetic main effect and underestimated gene-by-smoking interactions. When the genetic main effects and gene-by-smoking interaction effects were in opposing directions (ie, main effect negative, interaction effect positive), modelling with pack-years underestimated both genetic main effects and gene-by-smoking interactions. Increasing the strength of the gene-by-smoking interaction led to more bias when pack-years was used to model smoking effects. While in some scenarios the piecewise linear approach to smoking yielded biased estimates, in almost all instances the bias was smaller than that of the pack-years approach, and this bias reached statistical significance in only a small number of scenarios.

Graphical depictions of power to detect gene-by-smoking effects are shown in figure 2. In terms of power to detect gene-by-smoking interactions, the piecewise linear approach was more powerful than the pack-years approach.

We conducted two sensitivity analyses to examine the robustness of our results. In one sensitivity analysis we assumed a linear relation between pack-years and the strength of the gene-by-smoking effect (see table 1 in online supplement). In this scenario the piecewise linear approach was often comparable to or superior to the pack-years approach, although there were certain situations in which the pack-years approach performed better. To assess the impact of choice of the cut-off

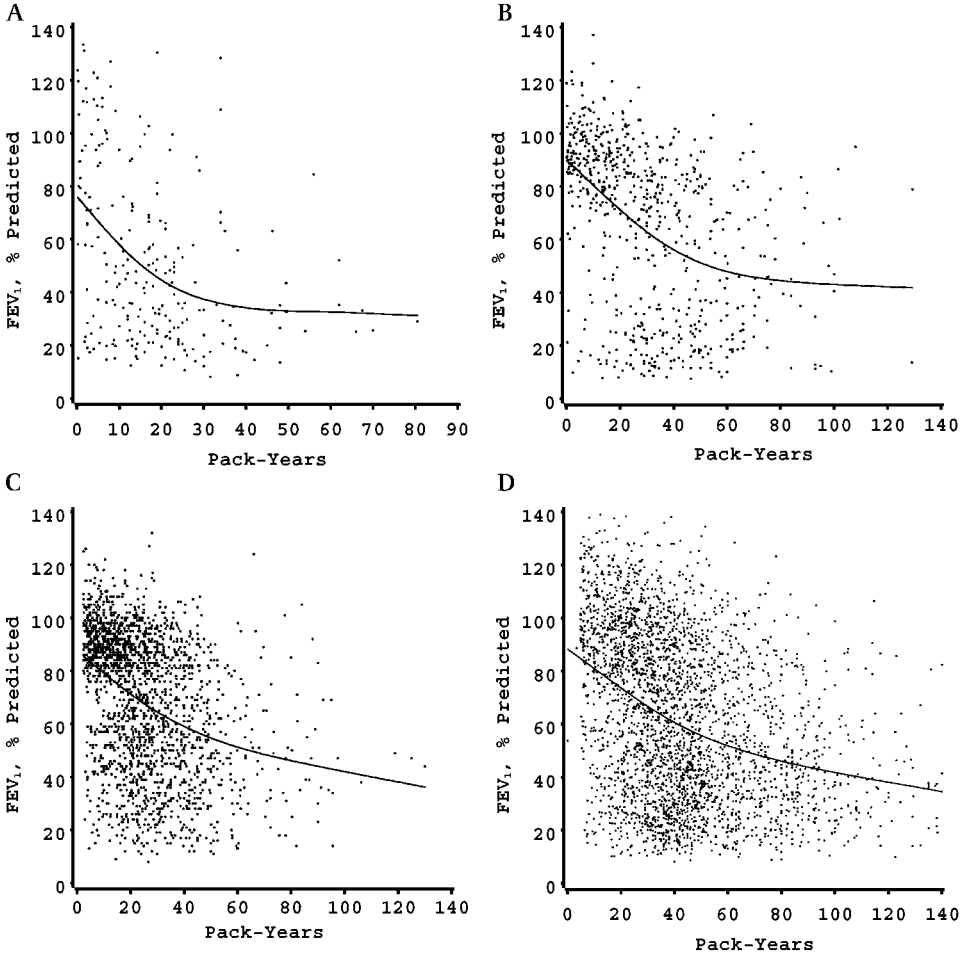

Figure 1 Scatterplots of forced expiratory volume in 1 s (FEV1) percentage predicted by pack-years in four large cohorts with smoothing curve: (A) Alpha-1 Antitrypsin Genetic Modifiers Study; (B) Boston Early-Onset COPD Study; (C) Norway Case–Control Study; (D) International COPD Genetics Network. Flattening of the curve occurs at FEV_1 levels between 30% and 50% of predicted.

Table 2 Simulation study results of the impact of two different methods of coding smoking on the accuracy of estimation for genetic main effect and gene-by-smoking interaction effects in the Alpha-1 Antitrypsin Genetic Modifiers Study*

Simulation parameters			Regression results†							
			Linear pack-years				Piecewise linear			
			Main effects		G×S		Main effects		G×S	
SNP main effect	G×S	MAF‡	Bias†	p Value§	Bias†	p Value§	Bias†	p Value§	Bias†	p Value§
0	0	0.10	0.17	0.024	−0.02	<0.001	0.05	0.512	−0.004	0.523
0	0	0.25	0.04	0.464	−0.01	<0.001	0.03	0.597	−0.01	0.040
0	0	0.40	−0.07	0.265	0.01	0.002	−0.04	0.466	0.01	0.251
−1	0	0.10	−0.04	0.578	−0.01	0.029	−0.19	0.014	0.01	0.027
−1	0	0.25	0.04	0.464	−0.004	0.282	0.08	0.198	−0.01	0.061
−1	0	0.40	0.05	0.382	0.002	0.520	0.08	0.205	−0.003	0.588
−1	−0.07	0.10	−0.09	0.250	0.02	<0.001	−0.05	0.523	0.003	0.664
−1	−0.07	0.25	−0.18	0.002	0.03	<0.001	−0.02	0.770	0.001	0.839
−1	−0.07	0.40	−0.09	0.124	0.04	<0.001	0.11	0.054	0.003	0.503
−1	−0.33	0.10	−0.73	<0.001	0.16	<0.001	−0.11	0.157	0.03	<0.001
−1	−0.33	0.25	−0.78	<0.001	0.18	<0.001	0.03	0.665	0.03	<0.001
−1	−0.33	0.40	−0.83	<0.001	0.18	<0.001	−0.07	0.288	0.03	<0.001
−1	−0.45	0.10	−0.87	<0.001	0.24	<0.001	−0.03	0.651	0.08	<0.001
−1	−0.45	0.25	−1.24	<0.001	0.27	<0.001	−0.23	<0.001	0.08	<0.001
−1	−0.45	0.40	−1.17	<0.001	0.27	<0.001	−0.15	0.015	0.08	<0.001
−1	0.07	0.10	0.23	0.001	−0.04	<0.001	−0.04	0.601	0.01	0.305
−1	0.07	0.25	0.21	<0.001	−0.03	<0.001	0.004	0.943	0.004	0.457
−1	0.07	0.40	0.04	0.507	−0.02	<0.001	−0.05	0.429	<0.001	0.988
−1	0.33	0.10	0.91	<0.001	−0.17	<0.001	−0.06	0.470	0.002	0.775
−1	0.33	0.25	0.94	<0.001	−0.17	<0.001	0.02	0.726	<0.001	0.919
−1	0.33	0.40	0.81	<0.001	−0.15	<0.001	0.01	0.821	<0.001	0.993
−1	0.45	0.10	1.18	<0.001	−0.23	<0.001	−0.08	0.273	−0.004	0.459
−1	0.45	0.25	1.26	<0.001	−0.22	<0.001	0.12	0.043	−0.01	0.059
−1	0.45	0.40	1.16	<0.001	−0.22	<0.001	0.04	0.534	−0.002	0.756

*Simulations were performed for same direction and opposite direction main and interaction effects. Strength of interaction effect is based on the multivariate smoking main effect of −1 unit from FEV_1 percentage predicted per pack-year.
Model is FEV_1 (% predicted)=age + sex + pack-years + gene + G×S, where the pack-years variable and the gene-by-smoking interaction term are calculated using either the total number of pack-years smoked (linear pack-years adjustment) or a piecewise linear representation of pack-years.
†Mean bias of β coefficient for main effect and gene-by-smoking interaction term from 1000 regressions on 1000 simulated datasets (ie, mean of the observed β coefficients − the simulated value of the pertinent effect).
‡Minor allele frequency of the simulated biallelic genetic variant.
§p Value for test of null hypothesis that bias=0.
FEV_1, forced expiratory volume in 1 s; G×S, gene-by-smoking interaction term; MAF, minor allele frequency; SNP, single nucleotide polymorphism.

point, we performed a sensitivity analysis in which we repeated our simulations using a range of cut-off points for the piecewise linear transformation of pack years (see table 2 in online supplement). As in the primary analysis, the underlying functional form of the gene-by-smoking interaction mirrored the form of the pack-years main effect. These results demonstrate that, while the cut-off point of 20 pack-years in this dataset performs better than the extremes, it is difficult to identify a single cut-off point that performs best for genetic main and interaction effects across all scenarios.

We applied these traditional pack-years and piecewise linear methods to case–control candidate gene data, performing genetic association analyses for genetic main effects and gene-by-smoking effects of the PI*Z allele in individuals from the Norway Case–Control COPD Study. We tested for association between PI MZ and two outcomes—FEV_1 percentage predicted and COPD susceptibility. The characteristics of PI MZ and PI MM subjects are shown in table 3. There were no statistically significant differences between the two groups in age, gender, pack-years or FEV_1. A cut-off of 40 pack-years was selected for the piecewise linear approach based on visual inspection and model fit. This cut-off was also supported in an examination of the relation between COPD susceptibility and pack-years (see figure 2 in online supplement). Using linear regression, we tested for an association between PI genotype and pack-years that might confound the association between genotype and FEV_1 or the gene-by-smoking interaction and found no evidence for this association (unadjusted p=0.79, p adjusted for age and sex=0.96).

The results of these analyses are shown in table 4. In both the FEV_1 and COPD susceptibility analyses, the main effect and Z allele-by-smoking effects are in opposite directions. In a manner that is consistent with our simulation results, the analyses using the piecewise linear approach yielded a stronger genetic main effect and Z allele-by-smoking interaction estimates than the pack-years approach. In both the FEV_1 and COPD susceptibility analyses, the piecewise linear approach demonstrates a statistically significant gene-by-smoking effect of the Z allele (p=0.03 and p=0.01, respectively), whereas the pack-years approach did not identify any statistically significant interactions.

DISCUSSION

This study identified a non-linear relation between smoking and FEV_1 in four large study samples. In simulation studies it was found that, in some scenarios, a piecewise linear approach to model smoking is superior to the commonly used pack-years approach in terms of accuracy and power to identify gene-by-smoking interactions. We applied this method in an analysis of the association of the PI MZ genotype with FEV_1 and COPD susceptibility and were able to detect statistically significant

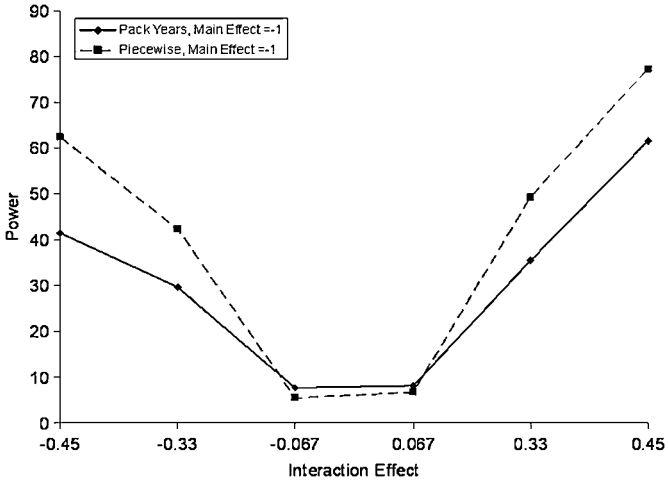

Figure 2 Observed power to detect gene-by-smoking interactions by the pack-years approach and the piecewise linear approach to model smoking. Simulation study based on data from the Genetic Modifiers of Alpha-1 Antitrypsin Disease Study. Simulation parameters are as follows: minor allele frequency 25%, genetic main effect −1 unit from observed FEV_1 percentage predicted per copy of minor allele, gene-by-smoking effect varies as shown. For this power analysis, the threshold for detecting an effect was set at $\alpha<0.05$ for the null hypothesis that the gene-by-smoking effect is equal to zero. For gene-by-smoking effects the piecewise linear model is more powerful.

Table 3 Characteristics of PI MZ and PI MM subjects in the Norway Case–Control Study

	PI MZ	PI MM	p Value
No of subjects	78	1591	–
Mean (SD) age (years)	62 (10)	60 (11)	0.33
Female, n (%)	43 (55)	875 (55)	0.97
Median (IQR) pack-years of smoking	25 (16–31)	22 (13–34)	0.24
Median (IQR) FEV_1 (% predicted)	72 (46–95)	81 (54–94)	0.38

FEV_1, forced expiratory volume in 1 s.

main and gene-by-smoking interaction effects with the piecewise linear modelling approach that would not have been detected with a pack-years approach. This pattern of results is consistent with the results of our simulations.

Previous work demonstrating a linear relation between FEV_1 and pack-years has generally focused on healthy population samples.[3–5 12 13] However, study samples recruited for many genetic association studies are specifically enriched for severe COPD, and our results show that the relation between pack-years and FEV_1 in these samples can be non-linear and should be considered when performing gene-by-smoking interaction analyses. A similar non-linear phenomenon in which risk tapers at higher levels of smoking exposure has been demonstrated with smoking intensity in lung cancer.[1]

The two most likely explanations for the non-linear relations observed are (1) survival bias (ie, differential population sampling at higher levels of cigarette exposure) and (2) a physiological floor in FEV_1 which, once reached, results in a diminished FEV_1 response to additional cigarette exposure. If these two mechanisms are active, the data points of most interest would be those that occur prior to the plateau phase in the relation between FEV_1 and pack-years, since the points on the plateau portion of the curve are likely to be affected either by survival bias or floor effects that may act to dilute the strength of any observed gene-by-smoking interactions. An additional problem with pack-years data is the potential for recall bias, particularly for individuals with extensive smoking histories or for those who have stopped smoking many years before the time of smoking ascertainment. If this bias increases with pack-years exposure, it could dilute the association between pack-years and FEV_1 at the extreme end of the pack-years distribution. In the cross-sectional data used in this study, it is difficult to distinguish between these explanations. Further study of this topic using longitudinal data would be useful, although survival bias can also affect longitudinal analyses of pulmonary function.[5] It should also be noted that a non-linear relation between pack-years and FEV_1 may result from occult interactions of pack-years with other variables. Thus, our proposed modelling approach may not necessarily reflect the true underlying relationship between FEV_1 and other important covariates.

In our analysis of the PI*Z allele-by-smoking interaction, we noted opposing directions of the main effect of the PI*Z allele and the PI*Z-by-smoking interaction. This result suggests that the deleterious effects of the PI*Z allele may become less prominent as smoking exposure increases. These results are consistent with a previously published report noting increased susceptibility to emphysema in PI*MZ individuals compared with PI*MM individuals that was limited to the low-smoking exposure subgroup.[14] It is possible that, for individuals with an increased genetic susceptibility to COPD, this difference is most notable at relatively low levels of smoke exposure and, as the

Table 4 Results from an analysis of main effect and gene-by-smoking interaction effects of the Z allele in PI MZ and PI MM subjects from the Bergen Norway Case–Control Study using two different methods of adjustment for smoking

		Z allele	p Value	Z×S interaction*	p Value	Pack-years†	p Value
Model for FEV_1 (% predicted)	Linear	−1.87	0.62	0.16	0.27	−0.33	<0.001
	Piecewise linear	−5.59	0.14	0.35	0.03	−0.38	<0.001
Model for COPD status	Linear	2.78	0.07	0.97	0.09	1.05	<0.001
	Piecewise linear	4.87	0.01	0.94	0.01	1.08	<0.001

Regression models adjusted for age, sex and smoking (using either a linear or piecewise linear form of the pack-years variable). The FEV_1 model includes a sample weight adjustment to reflect oversampling of COPD cases.
Results reported as β coefficients for FEV_1 model, ORs for COPD model and their respective p values.
*The Z×S interaction term is of the form (number of copies of Z allele × pack-years) for the linear smoking adjustment and (number of copies of Z allele × first 40 pack-years) for the linear and piecewise linear adjustments, respectively.
†Smoking adjustment for the linear pack-years model is done by including a numerical term for total pack-years of smoking. For the piecewise linear model, the smoking main effect is represented by two variables representing the first 40 pack-years of smoking and subsequent exposure. The reported beta coefficient is for the term representing the first 40 pack-years of smoking.
COPD, chronic obstructive pulmonary disease; FEV_1, forced expiratory volume in 1 s.

smoking burden increases, this relative difference becomes more difficult to detect.

Our study has the following strengths. First, we demonstrated the phenomenon of non-linearity between FEV_1 and pack-years in four large study samples. Second, our simulation strategy allowed us to compare the accuracy and power of two different approaches to model smoking in a setting in which the true values of genetic main and gene-by-smoking effects were known. Since our simulations were based on actual data, we preserved the natural noise present in FEV_1 measurements. Third, we were able to take the findings of our simulated studies and test them in a genetic association analysis of candidate gene data. Our findings are in line with previous results.[15] The main effects OR of the PI MZ genotype from the piecewise linear analysis for COPD susceptibility is comparable to a recent cumulative meta-analysis estimate, and the OR obtained using the total pack-years approach to these data is within the 95% CI limits of the meta-analysis estimate, suggesting that our sample is comparable to those of other PI MZ studies. Finally, our sample size compares favourably with most previous genetic association studies of PI MZ individuals.

One of the limitations of our study is that we have taken a simple approach—that is, piecewise linear modelling—to model the observed non-linearity of the smoking main effect, but a number of other modelling options could have been pursued such as multivariate adaptive regression splines (MARS) or generalised additive models. MARS incorporate piecewise linear modelling approaches similar to those used in this study, but it automates the selection of the cut-off point and model building process. MARS is more extensive in its modelling algorithms but can also require more degrees of freedom than our manual piecewise linear approach. Generalised additive models can fit highly non-linear curves to data in a piecewise fashion, but interpretation and hypothesis testing for covariates in these models is not straightforward. We also examined transforming the pack-years variable with packs-squared and inverse transformations, but these did not fit the data as well as the piecewise linear approach. Since our purpose was primarily to explore the implications of non-linearity of smoking main effects on the identification of gene-by-smoking interactions, the simplicity and interpretability of the piecewise linear approach were better suited for these purposes. As such, this method is a useful means of demonstrating the potential importance of non-linear smoking effects for COPD genetic association analyses, but further work is required to identify the optimal approach or set of approaches for handling such non-linear effects in large-scale genetic association analyses.

There are also other sources of complexity to consider regarding the identification of genetic interactions in the setting of non-linear effects that have not been fully explored in this paper. We assume that the functional form of the gene-by-smoking interaction mirrors that of the smoking main effect, but no empirical data are available regarding the true functional form of gene-by-smoking interactions in COPD and it is possible that the functional form may vary across different genetic variants. As more COPD-associated variants are identified, more empirical data regarding the form of gene-by-smoking interactions will become available. In addition, while our results support the concept that better fit for the smoking main effect can reduce bias in the gene-by-smoking interaction term, identification of the optimal method for selecting cut-off points for the piecewise linear variable requires further exploration.

A further limitation of our study is that it used self-reported smoking history. It is likely that this is relatively accurate for the interval of smoke exposure, but it is much less clear how well it serves as a measure of exposure.[16] Smokers vary greatly in their smoking behaviour. The exposure to smoke-derived toxins can therefore vary greatly from one smoker to the next despite similar numbers of cigarettes smoked. In addition, smoke chemistry is exceedingly complex.[17] Changes in smoke topography—that is, the way in which a cigarette is smoked including puff volume, puff time, dwell time and number of puffs per cigarette—all have profound effects on toxin exposure.[18] Even within a single individual, cigarettes are smoked differently and yields of toxin will vary, and it is likely that there will be differential exposure among the many toxins contained in smoke.[19] At present there are limited means of measuring exposure to specific smoke-derived toxins, but methodologies in this regard are advancing.

With the advent of large COPD genome-wide association studies, well-powered examinations for moderate to large gene-by-smoking interactions will be feasible, and gene-by-smoking interaction is likely to be an important aspect of future COPD genetic association analyses. We have shown that, in cross-sectional data of populations with a wide range of airflow obstruction, non-linear relations between FEV_1 and pack-years may be observed. In these situations, a piecewise linear approach to model the smoking main effect and gene-by-smoking interactions is preferable to modelling smoking as total pack-years, since it reduces bias and can be more powerful for detecting gene-by-smoking interactions.

Author footnote

The ICGN (International COPD Genetics Network) investigators are: Alvar Agusti (Hospital Universitari Son Dureta, Mallorca, Spain); Peter Calverley (University of Liverpool, Liverpool, UK); Claudio F Donner (S. Maugeri Foundation, Veruno, Novara, Italy); Robert D Levy (James Hogg iCAPTURE Centre, University of British Columbia, Vancouver, Canada); David Lomas (University of Cambridge, Cambridge, UK); Barry J Make (National Jewish Health, Denver, Colorado, USA); Wayne Anderson (GlaxoSmithKline, Research Triangle Park, North Carolina, USA); Peter Pare (James Hogg iCAPTURE Centre, University of British Columbia, Vancouver, Canada); Sreekumar Pillai (GlaxoSmithKline, Research Triangle Park, North Carolina, USA); Stephen Rennard (University of Nebraska, Omaha, Nebraska, USA); Emiel Wouters (University Hospital Maastricht, Maastricht, The Netherlands); Edwin K Silverman (The Channing Laboratory and Pulmonary and Critical Care Division, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA); and Jørgen Vestbo (Hvidovre Hospital, Copenhagen, Denmark). The AATGM (Alpha-1 Antitrypsin Genetic Modifiers Study) investigators are: Alan Barker (University of Oregon), Mark Brantly (University of Florida), Edward J Campbell (Utah Valley Pulmonary Clinic), Edward Eden (St Luke's/Roosevelt Hospital) N Gerard McElvaney (Beaumont Hospital, Dublin), Stephen Rennard (University of Nebraska), Robert Sandhaus (National Jewish Health), Edwin K Silverman (Brigham and Women's Hospital), James Stocks (University of Texas Health Center at Tyler), James Stoller (Cleveland Clinic), Charlie Strange (Medical University of South Carolina), Gerard Turino (St Luke's/Roosevelt Hospital).

Acknowledgements The authors thank John Ioannidis and David Kent for their discussions and input.

Funding The authors were supported by the following grants: K08HL102265, UL1 RR025752, R01 HL084323, R01 HL075478, U01 089856, and P01 HL083069. The International COPD Genetics Network is funded by a grant from GlaxoSmithKline.

Competing interests PDP served on the Advisory Board for Talecris Biotherapeutics and received grant support from GSK, Merck (≥$100 001), the NIH ($50,001—100,000), CIHR (Canada) and AllerGenNCE (≥$100 001). EKS received grant support and consulting fees from GlaxoSmithKline for studies of COPD genetics and honoraria and consulting fees from AstraZeneca. SR has consulted or participated in advisory boards for Able Associates, Adelphia Research, Almirall/Prescott, APT Pharma/Britnall, Aradigm, AstraZeneca, Boehringer Ingelheim, Chiesi, CommonHealth, Consult Complete, COPDForum, DataMonitor, Decision Resources, Defined Health, Dey, Dunn Group, Eaton Associates, Equinox, Gerson, GlaxoSmithKline, Infomed, KOL Connection, M Pankove, MedaCorp, MDRx Financial, Mpex, Novartis, Nycomed, Oriel Therapeutics, Otsuka, Pennside Partners, Pfizer (Varenicline), PharmaVentures, Pharmaxis, Price Waterhouse, Propagate, Pulmatrix, Reckner Associates, Recruiting Resources, Roche, Schlesinger Medical, Scimed, Sudler and Hennessey, TargeGen,

Theravance, UBC, Uptake Medical and VantagePoint Management; has given lectures for the American Thoracic Society, AstraZeneca, Boehringer Ingelheim, California Allergy Society, Creative Educational Concept, France Foundation, Information TV, Network for Continuing Ed, Novartis, Pfizer and SOMA; and has received industry-sponsored grants from AstraZeneca, Biomarck, Centocor, Mpex, Nabi, Novartis and Otsuka. DAL has received grant support, consultancy fees and honoraria from GlaxoSmithKline, consultancy fees from Talecris Biotherapeutics, Genzyme and Amicus Therapeutics and honoraria from LKB. JV has received honoraria for consulting and presenting for pharmaceutical companies with an interest in COPD, and is an investigator on the ECLIPSE study and the International COPD Genetics Network, both sponsored by GlaxoSmithKline.

Ethics approval This study was conducted with the approval of the Partners IRB.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Rachet B,** Siemiatycki J, Abrahamowicz M, *et al*. A flexible modeling approach to estimating the component effects of smoking behavior on lung cancer. *J Clin Epidemiol* 2004;**57**:1076—85.
2. **Hoffmann K,** Bergmann MM. Modeling smoking history: a comparison of different approaches. *Am J Epidemiol* 2003;**158**:393—4.
3. **Dockery DW,** Speizer FE, Ferris BG Jr, *et al*. Cumulative and reversible effects of lifetime smoking on simple tests of lung function in adults. *Am Rev Respir Dis* 1988;**137**:286—92.
4. **Burrows B,** Knudson RJ, Cline MG, *et al*. Quantitative relationships between cigarette smoking and ventilatory function. *Am Rev Respir Dis* 1977;**115**:195—205.
5. **Xu X,** Dockery DW, Ware JH, *et al*. Effects of cigarette smoking on rate of loss of pulmonary function in adults: a longitudinal assessment. *Am Rev Respir Dis* 1992;**146**:1345—8.
6. **Stavem K,** Aaser E, Sandvik L, *et al*. Lung function, smoking and mortality in a 26-year follow-up of healthy middle-aged males. *Eur Respir J* 2005;**25**:618—25.
7. **DeMeo DL,** Sandhaus RA, Barker AF, *et al*. Determinants of airflow obstruction in severe alpha-1-antitrypsin deficiency. *Thorax* 2007;**62**:806—13.
8. **Patel BD,** Coxson HO, Pillai SG, *et al*. Airway wall thickening and emphysema show independent familial aggregation in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2008;**178**:500—5.
9. **Silverman EK,** Chapman HA, Drazen JM, *et al*. Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am J Respir Crit Care Med* 1998;**157**:1770—8.
10. **Pillai SG,** Ge D, Zhu G, *et al*. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genetics* 2009;**5**:e1000421.
11. **American Thoracic Society**. Standardization of spirometry, 1994 update. *Am J Respir Crit Care Med* 1995;**152**:1107—36.
12. **Fletcher C,** Peto R. The natural history of chronic airflow obstruction. *BMJ* 1977;**1**:1645—8.
13. **Kerstjens HA,** Rijcken B, Schouten JP, *et al*. Decline of FEV_1 by age and smoking status: facts, figures, and fallacies. *Thorax* 1997;**52**:820—7.
14. **Sorheim IC,** Bakke P, Gulsvik A, *et al*. Alpha-1 antitrypsin PI MZ heterozygosity is associated with airflow obstruction in two large cohorts. *Chest* 2010;**138**:1125—32.
15. **Hersh CP,** Dahl M, Ly NP, *et al*. Chronic obstructive pulmonary disease in $alpha_1$-antitrypsin PI MZ heterozygotes: a meta-analysis. *Thorax* 2004;**59**:843—9.
16. **Hatsukami DK,** Benowitz NL, Rennard SI, *et al*. Biomarkers to assess the utility of potential reduced exposure tobacco products. *Nicotine Tob Res* 2006;**8**:600—22.
17. **Borgerding M,** Klus H. Analysis of complex mixtures—cigarette smoke. *Exp Toxicol Pathol* 2005;**57**(Suppl 1):43—73.
18. **O'Connor RJ,** Kozlowski LT, Hammond D, *et al*. Digital image analysis of cigarette filter staining to estimate smoke exposure. *Nicotine Tob Res* 2007;**9**:865—71.
19. **Mooney M,** Green C, Hatsukami D. Nicotine self-administration: cigarette versus nicotine gum diurnal topography. *Hum Psychopharmacol* 2006;**21**:539—48.