LIFETIME SMOKING AND ITS IMPACT ON RISK FOR LUNG CANCER AND CARDIOVASCULAR DISEASE: RESULTS FROM THE FRAMINGHAM HEART STUDY

By

Meredith Stevenson Duncan

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Epidemiology

August 7, 2020

Nashville, Tennessee

Approved:

Matthew S. Freiberg, M.D.

Loren Lipworth, Sc.D.

Robert A. Greevy, Ph.D.

Melinda C. Aldrich, Ph.D.

Hilary A. Tindle, M.D.

To my husband, Benjamin, for supplying me with necessary dissertation fuel: love, laughter, and liquor

ACKNOWLEDGEMENTS

This work would not have been possible without the dedication of the Framingham Heart Study participants and staff. I am also indebted to both the Vanderbilt University Medical Center Division of Cardiovascular Medicine and the Division of Epidemiology, who have actively supported both my professional and educational endeavors for the last two years.

I am grateful to everyone with whom I have had the pleasure to work during this and other projects. Each member of my Dissertation Committee has provided me with extensive personal and professional mentorship over the last year; their words of wisdom will carry me far in my future. I would like to specifically acknowledge Dr. Matthew Freiberg for his guidance, even when I didn’t want to hear it, and to Dr. Loren Lipworth, for listening to me complain about Matt’s good advice that I wasn’t ready to follow. I am also grateful to the members of the V-C^{3}REATE team (both past and present) who encouraged and supported me and became friends as well as colleagues.

No one has shown more patience and care in the pursuit of my education than my family. I would like to thank my husband, who has now moved to three different states in support of my education and career and has put up with an embarrassing amount of whining. Finally, I would like to thank my parents who have been my endless cheerleaders: pushing me forward when all I wanted to do was quit. I made it because of you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

IMPACT OF INCLUDING COMPREHENSIVE SMOKING DATA IN ASCVD RISK ESTIMATION

Introduction

The cornerstone for atherosclerotic cardiovascular disease (ASCVD) prevention is assessing risk prior to the event and implementing lifestyle modifications accordingly; the current gold standard risk assessment tool is the ASCVD Risk Estimator Plus.[1,2] This calculator considers former smokers to be at excess ASCVD risk compared to never smokers for the first 5 years of cessation and does not incorporate pack-years smoked. However, a recent study demonstrated that the excess risk of cardiovascular disease among former heavy ($\geq$20 pack-years) smokers as compared to never smokers can remain for up to 16 years post cessation.[3] Thus, both years since quitting smoking and pack-years smoked may play an important role in ASCVD risk estimation.

The backbone of the ASCVD Risk Estimator Plus is the 2013 American College of Cardiology/American Heart Association (ACC/AHA) Guideline on the Assessment of Cardiovascular Risk and its corresponding sex- and race-specific Pooled Cohort Equations (2013 PCE) for estimation of 10-year ASCVD risk.[4] When developing the ASCVD Risk Estimator Plus, Lloyd-Jones et al. sought to expand the 2013 PCE by incorporating use of ASCVD preventive medications and smoking cessation.[1] While the quality of evidence for the estimates regarding the impact of pharmacotherapies on ASCVD risk reduction was considered "high" by the authors, they noted that the "effects [of smoking cessation] on ASCVD risk [are] poorly reported."[1] For this reason, Lloyd-Jones et al. estimated the effects of smoking cessation on ASCVD risk using pooled results from Lee et al. [5,6] which conclude that the excess risk of coronary heart disease and cerebrovascular disease attributable to smoking has a half-life of 4 to 5 years (relative to continuing smokers) but make no comment about the excess risk in former smokers following cessation compared to never smokers. Thus, although the ASCVD Risk Estimator Plus approximates the ASCVD risk benefit from smoking cessation relative to continuing smokers, an improvement over prior tools, evidence is lacking to support the assumption that ASCVD risk among former smokers – regardless of pack-years smoked – approaches that of never smokers within five years of cessation. There are now more than 55 million US adults who have quit smoking,[7] and this number grows as smoking cessation rates increase.[8] Potential misclassification of ASCVD risk among former smokers could have large implications.

In this context, we fit a series of ASCVD risk prediction models using longitudinal data from the Framingham Heart Study. By assessing a variety of model fit metrics spanning goodness-of-fit, discrimination, calibration, and reclassification, we evaluated the predictive utility of adding years since quitting smoking and pack-years smoked to the 2013 PCE for more accurate ASCVD risk prediction.

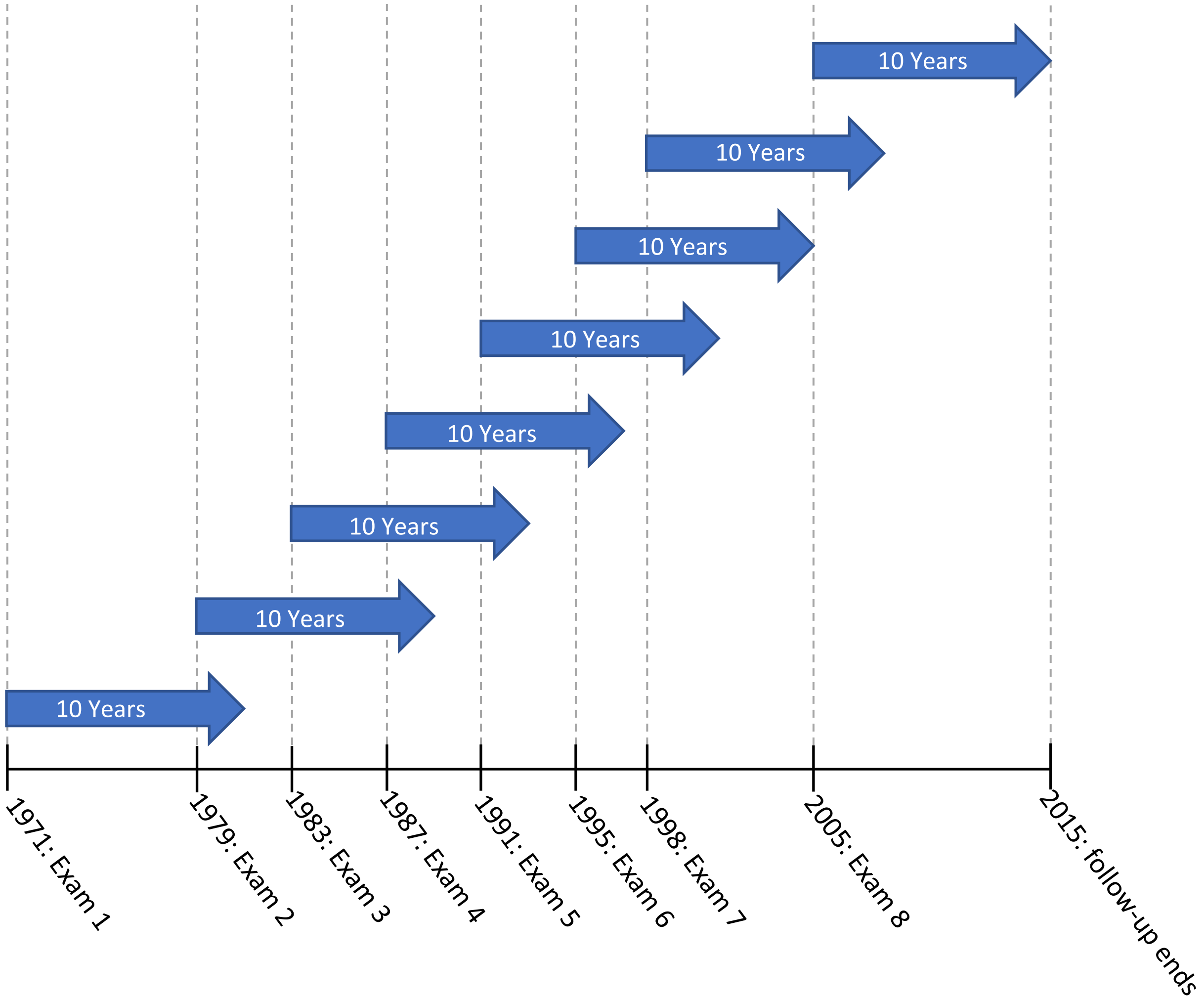
Methods

Study Sample

This investigation included more than 50-years of data on Framingham Heart Study (FHS) Offspring[9] cohort participants who attended their first (1971-1975) examination cycle and had at least one additional contact with study staff following examination one. Participants presented for quadrennial examination cycles beginning in 1979 (up to 8 examination cycles total in this analysis). Each person-examination and the subsequent 10-year follow-up period served as an independent ASCVD risk prediction window, yielding up to 8 data records per individual. **Figure 1** displays examination timing and subsequent follow-up.

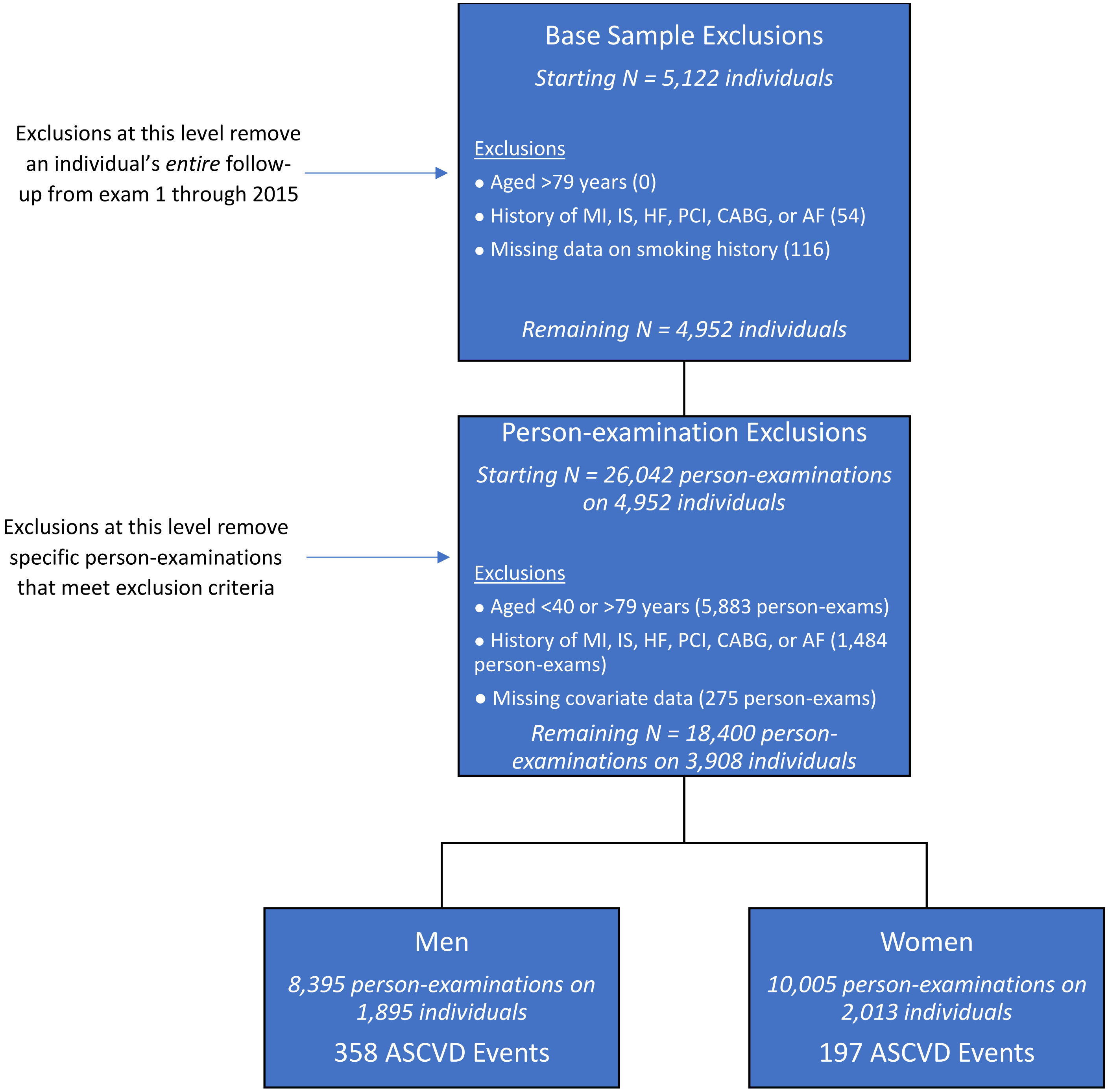
The analysis dataset was constructed in two phases (**Figure 2**). First, we excluded participants from the base sample (n=5,122) if, at examination one, they had a history of myocardial infarction (MI), ischemic stroke (IS), heart failure (HF), coronary artery bypass graft (CABG), percutaneous coronary intervention (PCI), or atrial fibrillation (n=54) or were missing data that would prevent us from quantifying lifetime smoking history (n=116). After these exclusions, 4,952 individuals remained in our sample. In the second phase of dataset construction, we applied exclusion criteria at the person-examination level since these served as individual "baselines" for 10-year ASCVD risk prediction. Exclusion criteria at this level were similar to those described by Goff et al.[4] From a starting sample of 26,042 person-examinations, person-exams were excluded for the following reasons: aged <40 or >79 years (n = 5,883 person-examinations); history of MI, IS, HF, CABG, PCI, or atrial fibrillation (n = 1,484 person-examinations); or missing data on predictors (n = 275 person-examinations). Following exclusions, our final sample for analysis contained 18,400 person-examinations on 3,908 individuals (**Figure 2**).

Figure 1: Timeline for FHS Offspring Examinations



Timing of FHS Offspring examinations and corresponding 10-year follow-up periods.

Figure 2: Sample Flow Diagram



Exclusions at this level remove an individual's *entire* follow-up from exam 1 through 2015

Base Sample Exclusions

Starting N = 5,122 individuals

Exclusions

- Aged >79 years (0)
- History of MI, IS, HF, PCI, CABG, or AF (54)
- Missing data on smoking history (116)

Remaining N = 4,952 individuals

Exclusions at this level remove specific person-examinations that meet exclusion criteria

Person-examination Exclusions

Starting N = 26,042 person-examinations on 4,952 individuals

Exclusions

- Aged <40 or >79 years (5,883 person-exams)
- History of MI, IS, HF, PCI, CABG, or AF (1,484 person-exams)
- Missing covariate data (275 person-exams)

Remaining N = 18,400 person-examinations on 3,908 individuals

Men

8,395 person-examinations on 1,895 individuals

358 ASCVD Events

Women

10,005 person-examinations on 2,013 individuals

197 ASCVD Events

Outcome

FHS participants are under continuous surveillance for the development of new ASCVD events. In this investigation, participants were surveilled until December 31, 2016 for the development of ASCVD events including MI, fatal or non-fatal IS, and coronary heart disease death. For suspected ASCVD events, medical records were obtained with permission; events were adjudicated by three Study physicians as previously detailed.[3,10–12]

ASCVD Risk Factor Definitions

To assess the significance of adding smoking related variables to the ASCVD Risk Estimator Plus, our risk factor definitions were the same as those previously published by Goff et al.[4] and Lloyd-Jones et al.[1] in sex-specific models (age, total cholesterol, use of lipid-lowering medication, high-density lipoprotein [HDL] cholesterol, systolic blood pressure [SBP], use of antihypertensive medication, and diabetes mellitus). Lloyd-Jones et al. also included aspirin therapy in their development of the ASCVD Risk Estimator Plus. Because the safety and efficacy of aspirin is being reconsidered,[13–16] we have excluded it from our models.

Blood was drawn from participants at each examination cycle following an overnight fast of at least 10 hours. Biospecimens were stored at -20 (pre-1990 exams) to -80 C (post-1990 exams) until they were assayed. Total cholesterol and HDL cholesterol were directly measured using standardized assays.[17] Systolic blood pressure was averaged based on two physician readings. Use of lipid-lowering medication or antihypertensive medication was self-reported at examination cycles 2-7 and verified by study staff upon review of provided medication at examination cycle 8. Participants were classified as having diabetes mellitus based fasting blood glucose >126 mg/dL or receipt of medication for the treatment of diabetes mellitus.

Quantification of smoking status and intensity have been previously described.[3,18] Briefly, at the baseline examination, participants were categorized as "current," "former," or "never" smokers based on their responses to questions regarding prior smoking habits. From responses given for age at which the participant starting smoking, usual cigarettes smoked per day in the past, age at quitting (former smokers), and current number of cigarettes smoked per day (current smokers), we calculated ever smokers' pack years and years since quitting (YSQ) for former smokers. Never smokers were assigned a pack-year value of 0. For this analysis our smoking measures included current/former/never smoking status, pack-years, and YSQ.

Years since quitting is a conditionally relevant predictor in that it is intuitive for current smokers to have a value of 0 (i.e., they have not yet quit) and for former smokers' YSQ value to be greater than 0, but there is no relevant value for never smokers. Furthermore, we previously observed that YSQ was associated with ASCVD risk among heavy ever smokers ($\geq$20 pack-years).[3] Thus, we used the two-part predictor method described by Dziak and Henry[19] to adjust for this conditionally relevant predictor in heavy ever smokers only. The two variables that represent the effect of YSQ among heavy ever smokers are an indicator of when YSQ is relevant (heavy ever smokers), and a mean-centered YSQ value (YSQ*), where the mean is calculated in heavy ever smokers only. Never and non-heavy ever smokers' value of YSQ* was then set to 0 so that they would not impact (i.e. exert statistical leverage on) the estimation of the effect of YSQ* among heavy ever smokers.

Statistical Analysis

Summary statistics were stratified by sex; within sex, statistics were pooled over all person-examinations meeting inclusion criteria. We calculated means and standard deviations for normally distributed variables and medians along with the 25th and 75th percentiles for variables with skewed distributions. Counts and percentages were calculated for categorical variables.

We performed a series of sex-specific Cox proportional hazards regressions to predict ASCVD incidence. We began with the 2013 PCE, in which continuous variables were natural-logarithmically transformed. We then modified the 2013 PCE by modeling continuous variables on their natural scale and allowed for nonlinearity in the association between these predictors and ASCVD risk through inclusion of polynomial terms. Finally, we expanded the list of predictors to include a 3-level smoking variable (current/former/never) both alone and in combination with pack-years smoked and years since quitting.

We first fit the sex-specific 2013 PCE in our data, re-estimating both the beta coefficients and the baseline hazard in our sample to allow for fair model comparison. This model (Model 1) included age, total cholesterol, HDL cholesterol, treated SBP, untreated SBP, current smoking status, and diabetes status as predictors. Model 2 used the same predictors as the 2013 PCE with 3 changes: 1) inclusion of continuous variables on their natural scale (i.e., not logarithmically transformed); 2) up to third order polynomials on continuous variables (and their interactions) to account for non-linearity; and 3) adjustment for antihypertensive use rather than classifying SBP as "treated" or "untreated". In Model 3 we replaced the binary indicator of current smoking with a 3-level smoking variable to distinguishing between former and never smokers. In Model 4, we built upon Model 3 and additionally adjusted for pack-years smoked and its interaction with age. Model 5 expanded Model 4 to further adjust for YSQ* and its interaction with age.

To determine whether pack-years smoked and/or years since quitting should be added to the current predictors in the 2013 PCE, we evaluated how well these variables meet the American Heart Association's (AHA) criteria for evaluating an added predictor as described by Hlatky et al.[20] These specific criteria can be summarized as: 1) proof of concept (association); 2) prospective validation; 3) incremental value; 4) clinical utility; 5) clinical outcomes; and 6) cost-effectiveness. Criteria 1 and 2 are satisfied since the cross-sectional and prospective associations of pack-years smoked with ASCVD are well-established and will therefore not be covered in this investigation. Similarly, pack-years and years since quitting are easily calculated from information in a patient's chart or based on a brief battery of questions (when they started, have they quit, how many cigarettes per day, etc.) and require no lab work, making them highly cost-effective to obtain (criterion 6). Criterion 5, "clinical outcomes," refers to whether the use of the risk marker in clinical management improves clinical outcomes, which is typically assessed via a clinical trial.[20] However, clinical trials are unethical in this scenario and once pack-years are accumulated they cannot be reduced – their impact can just be mitigated over several years of cessation. Thus, we refer to our prior work in this cohort which reported lower ASCVD rates among lighter smokers compared to heavier smokers, and diminishing ASCVD risk with greater years since quitting[3] as evidence that criterion 5 is fulfilled. Therefore, we focused on determining whether pack-years and/or years since quitting satisfy the incremental value and clinical utility criteria.

We assessed incremental value via change in Harrell's c-statistic (Δc)[21] and continuous net reclassification improvement (NRI(>0)).[22–25] Statistical significance of both metrics was evaluated via bootstrapped confidence intervals. NRI(>0) values are deemed strong when >0.6, intermediate when >0.4, and weak when <0.2;[25] to our knowledge, there is no recommended threshold of Δc that is uniformly recognized as clinically significant. To visually assess risk reclassification, we created reclassification plots for each model under consideration. In each plot, 10-year predicted probability of ASCVD from the base model appeared on the x-axis, and the corresponding probability from each comparator model appeared on the y-axis.

Clinical utility of a variable refers to its ability to sufficiently move an individual across the risk spectrum when added to a model; this is reflected by the relative integrated discrimination improvement (rIDI).[25] The statistical significance of rIDI can also be assessed using a bootstrapped confidence interval. Regarding clinical significance, if a model including a new predictor has rIDI>1/*p* where *p* is the number of predictors in the base model, this indicates that the added predictor has an effect size greater than the average of that possessed by the predictors in the base model.[25]

Assessment of NRI(>0) and rIDI requires that models are well-calibrated. Thus, we also calculated the D'Agostino and Nam extension of the Hosmer-Lemeshow calibration test[26] using categories of <5%, 5-7.49%, 7.5-19.9%, and ≥20% to define low, borderline, intermediate, and high-risk groups.[1,2] To be thorough in our assessment of pack-years and years since quitting, we evaluated goodness-of-fit via likelihood ratio and Nagelkerke's R^2.[27,28]

Exploratory analyses compared risk category classification under the re-estimated 2013 PCE (Model 1) versus the 2013 PCE on natural scale plus 3-level smoking, pack-years, and years since quitting (Model 5) in events and non-events separately within each sex. Categories were defined as low, borderline, intermediate, and high using the same cutoffs described above to maintain consistency with the ASCVD Risk Estimator Plus.[1,2] Although categorical reclassification has been criticized for poor statistical properties, particularly when it comes to assigning risk category cut-points,[29] reclassification tables are useful to visually assess overall patterns of risk classification under different models; it is in this capacity that they will be utilized here.

As mentioned above in the "Study Sample" section, approximately 1% of person-examinations were missing covariate data. Since the proportion of missingness was small, we assumed that values were missing completely at random. Under this assumption, excluding person-examinations with missing data would not bias our results. All analyses were performed in SAS 9.4 (Cary, NC).

Results

We analyzed data from 18,400 person-examinations on 3,908 individuals (Men: 8,395 person examinations on 1,895 people; Women: 10,005 person-examinations on 2,013 people). Most risk factors were similar between men and women (**Table 1**). Nearly a quarter of men and women were never smokers, but more men were former smokers (45%) compared to women (36%). Correspondingly, a larger proportion of females than males were current smokers (women 42%; men 32%). However, men who had ever smoked tended to smoke more heavily than their female counterparts (**Table 1**). Median pack years differed by sex among current (men 39; women 32) and former (men 22; women 11) smokers.

Table 1: Sample Characteristics by Sex over all Person-Examinations

Characteristic*	**Men: N = 8,395 Person-Examinations**	**Women: N = 10,005 Person-Examinations**
Age, years	54.8 (9.4)	55.5 (9.6)
Systolic Blood Pressure, mmHg	129.0 (16.8)	125.5 (18.6)
Antihypertensive Medication	1864 (22.2)	2068 (20.7)
Diabetes	643 (7.7)	495 (5.0)
Total Cholesterol, mg/dL	206.1 (38.4)	211.2 (38.9)
HDL Cholesterol, mg/dL	44.7 (12.8)	58.4 (16.4)
Lipid Lowering medication use	752 (9.0)	834 (8.3)
Smoking Status		
Current	2684 (32.0)	4187 (41.9)
Former	3790 (45.2)	3573 (35.7)
Never	1921 (22.9)	2245 (22.4)
Cigarettes per day†	20.0 (20.0, 30.0)	20.0 (10.0, 30.0)
Pack-Years		
Current Smokers	39.1 (27.2, 55.3)	32.0 (19.2, 46.6)
Former Smokers	22.0 (10.0, 37.5)	11.0 (4.4, 23.4)
Years Since Quitting‡	16.1 (8.5, 25.4)	15.5 (8.1, 24.5)

* Summary statistics are displayed as: Mean (SD) for age, systolic blood pressure, total and HDL cholesterol; Median (Q1, Q3) for cigarettes per day, pack-years, and years since quitting; and as N (%) for categorical variables.

† Among current smokers only.

‡ Among former smokers only.

Over a total of 169,484 person-years, 555 incident ASCVD events occurred: 358 in men and 197 in women. Predictors in the 5 Cox models in each sex (10 models total) are displayed in **Table 2**. Results of the model fitting procedure in men and women are displayed in **Table 3** and **Table 4**, respectively. All models were well-calibrated (**Table 3** and **Table 4**, **Supplemental Figure 1** and **Supplemental Figure 2**).

Table 2: Model Descriptions

Variable	Men					Women				
	Model Number					Model Number				
	1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
Age	X^1	X^2	X^2	X^2	X^2	X^2	X^2	X^2	X^2	X^2
Total Cholesterol	X^{1*}	X^{2*}	X^{2*}	X^{2*}	X^{2*}	X^{1*}	X^{2*}	X^{2*}	X^{2*}	X^{2*}
HDL Cholesterol	X^{1*}	X^{2*}	X^{2*}	X^{2*}	X^{2*}	X^{1*}	X^{2*}	X^{2*}	X^{2*}	X^{2*}
Non-HDL Cholesterol										
Lipid-Lowering Medication										
Untreated SBP	X^1					X^1				
Treated SBP	X^1					X^1				
SBP		X^3	X^3	X^3	X^3		X^3	X^3	X^3	X^3
Antihypertensives		X†	X†	X†	X†		X	X	X	X
Current Smoking	X*	X*				X*	X*			
Current/Former/Never Smoking			X*	X*	X*			X*	X*	X*
Pack-Years Smoked				X^{2*}	X^{2*}				X^{2*}	X^{2*}
Years Since Quitting					X^{2*}					X^{2*}
Diabetes Mellitus	X	X	X	X	X	X	X	X	X	X
Total Model df	**10**	**23**	**26**	**31**	**36**	**11**	**23**	**26**	**31**	**36**

Abbreviations: degrees of freedom (df); systolic blood pressure (SBP)
An “X” in the cell indicates that the variable was included in the model. Blank cells indicate that the variable was not included in the model.
In model 1, continuous variables were on the natural logarithm scale as in the 2013 PCE.
Superscripts 1-3 indicate the order of the polynomial used to model continuous variables
* Interacted with age
† Interacted with SBP

Model Fitting in Men

We began by fitting the 2013 PCE with re-estimated baseline hazard and beta coefficients in our sample (Model 1, **Table 2**). Since this model has 10 degrees of freedom, rIDI values of comparator models >0.1 are clinically meaningful. When we refit the 2013 PCE with continuous variables on the natural scale and included up to third-order polynomials on these continuous variables to account for

Table 3: Model Summaries in Men

Model No.	Description	-2 Log L	df	Δdf	LR χ^2	LR p-value	R^2	Calibration χ^2	Calibration p-value	c-statistic	Δ Harrell's c [95% CI]	NRI(>0) [95% CI]	Relative IDI [95% CI]
1	Re-estimated 2013 PCE*	6127.79	10	--	--	--	6.40%	6.13	0.11	0.7487	--	--	--
2	M1 on Natural Scale†	6110.53	23	--	--	--	6.77%	6.97	0.07	0.7538	0.005 [0.0005, 0.010]	0.301 [0.193, 0.409]	0.071 [0.001, 0.137]
3	M2 + CFN Smoke	6101.89	26	3	8.64	0.02	6.96%	3.62	0.31	0.7559	0.002 [-0.002, 0.006]	0.145 [0.042, 0.254]	0.047 [0.021, 0.075]
4	M3 + Pack-Years	6086.75	31	8	23.78	<0.01	7.27%	2.72	0.44	0.7596	0.006 [0.001, 0.011]	0.259 [0.156, 0.358]	0.175 [0.092, 0.271]
5	M4 + YSQ	6084.36	36	13	26.17	0.02	7.32%	3.03	0.39	0.7604	0.007 [0.001, 0.120]	0.233 [0.132, 0.332]	0.191 [0.096, 0.295]

Table 4: Model Summaries in Women

Model No.	Description	-2 Log L	df	Δdf	LR χ^2	LR p-value	R^2	Calibration χ^2	Calibration p-value	c-statistic	Δ Harrell's c [95% CI]	NRI(>0) [95% CI]	Relative IDI [95% CI]
1	Re-estimated 2013 PCE*	3339.11	11	--	--	--	8.90%	4.68	0.20	0.8090	--	--	--
2	M1 on Natural Scale†	3326.30	23	--	--	--	9.31%	2.93	0.40	0.8123	0.003 [-0.002, 0.009]	0.040 [-0.109, 0.174]	0.147 [0.044, 0.285]
3	M2 + CFN Smoke	3323.39	26	3	2.91	0.16	9.41%	2.20	0.53	0.8133	0.001 [-0.002, 0.004]	0.163 [0.020, 0.303]	0.026 [-0.001, 0.053]
4	M3 + Pack-Years	3310.79	31	8	15.51	0.02	9.80%	5.74	0.13	0.8179	0.006 [-0.003, 0.014]	0.287 [0.148, 0.424]	0.098 [0.046, 0.165]
5	M4 + YSQ	3306.88	36	13	19.42	0.11	9.92%	3.07	0.38	0.8195	0.007 [-0.002, 0.017]	0.340 [0.204, 0.478]	0.112 [0.052, 0.185]

Abbreviations: Akaike's Information Criterion (AIC); current/former/never (CFN); confidence interval (CI); degrees of freedom (df); integrated discrimination improvement (IDI); likelihood ratio (LR); continuous net reclassification improvement (NRI(>0)); years since quitting smoking (YSQ)

* This is the reference for Model 2 when calculating Δ Harrell's c-statistic, NRI(>0), and relative IDI

† This is the reference for Models 3-5 when calculating the likelihood ratio test, Δ Harrell's c-statistic, NRI(>0), and relative IDI

non-linear associations with ASCVD risk (Model 2), we observed a small but statistically significant increase in Harrell's c-statistic and a moderate NRI(>0) when compared to Model 1; rIDI was statistically significant but not clinically meaningful (**Table 3**). Since these model modifications were statistically significant, Model 2 became our new "base" model to which we compared the added value of a 3-level smoking status variable alone and in conjunction with pack-years and years since quitting. Thus, rIDI values for Model 3-5 >1/23=0.043 were considered clinically meaningful.

When replacing a binary indicator of current smoking with a 3-level smoking variable that further differentiated between former and never smokers (Model 3), this change improved goodness-of-fit (likelihood ratio and R^2), rIDI was both statistically significant and clinically meaningful, NRI(>0) was statistically significant but clinically weak, and Δc was not significant. Addition of pack-years to Model 3 (Model 4) proved significant on all metrics, produced a clinically meaningful rIDI and a moderate NRI(>0). Addition of years since quitting to Model 4 (Model 5) was also significant on all metrics, with rIDI and c-statistic greater than that of Model 4 but a slightly lower NRI(>0) than that of Model 4.

Overall, Model 5 (2013 PCE on natural scale plus 3-level smoking, pack-years, and years since quitting) was the best fit to the data: it had the highest likelihood and R^2, greatest Harrell's c-statistic, moderate NRI(>0) compared to Model 2, and a statistically significant and clinically meaningful rIDI.

Model Fitting in Women

Model fitting in women progressed the same way as in men (**Table 4**). Changing from the re-estimated published 2013 PCE to the 2013 PCE on the natural scale produced a clinically meaningful rIDI value, but Δc and NRI(>0) were not significant. Substituting the current smoking indicator for current/former/never smoking status did not improve model fit. Addition of pack-years smoked improved goodness-of-fit, produced a moderate NRI(>0) and a clinically meaningful rIDI, but did not improve the c-statistic. When years since quitting was added to Model 4, the likelihood ratio test and Δc were not significant, but R^2 increased, the NRI(>0) was moderate, and rIDI was both statistically significant and clinically meaningful (**Table 4**).

Exploratory Analyses: Reclassification Tables

The majority of individuals remained in the same risk category under Models 1 and 5 (**Appendix A, Supplemental Tables 1-4**). Among men, similar proportions of non-events were correctly reclassed at lower risk (7.1%) and incorrectly reclassed as higher risk (6.0%) under Model 5 compared to Model 1 (**Table S1**). However, 14% of men who experienced ASCVD events within 10 years were classified as higher risk under the 2013 PCE on natural scale plus 3-level smoking, pack-years, and years since quitting (Model 5) compared to the re-estimated 2013 PCE (Model 1) while only 6.7% were incorrectly reclassified into a lower risk category (**Appendix A, Supplemental Table 1**). In women, approximately 3% of non-events were correctly reclassified into a lower risk category while a similar proportion was incorrectly reclassed into a higher category under Model 5 compared to Model 1. In contrast, 15% of women with events were assigned a higher risk category under Model 5 versus 1 compared to only 6.1% assigned to a lower risk category (**Appendix A, Supplemental Table 2**). In heavy ever smoking men, 13.6% of non-events and 11.3% of events were correctly reclassed under Model 5 compared to 7.3% of non-events and 9.0% of events who moved in the incorrect direction (**Appendix A, Supplemental Table 3**). Reclassification patterns in heavy ever smoking women were similar to those of the full sample (**Appendix A, Supplemental Table 4**).

Discussion

In the Framingham Offspring cohort, inclusion of pack-years and years since quitting improved ASCVD risk prediction over the 2013 PCE. Specifically, addition of these variables produced moderate NRI(>0) values of 0.23 and 0.34 in men and women, respectively, and meaningful rIDI values of 0.19 (men) and 0.11 (women). Based on these results, pack-years smoked and years since quitting demonstrate incremental value and clinical utility, thus fulfilling the AHA's 2 remaining criteria for evaluating an added predictor.[20]

To our knowledge, this is the first investigation to demonstrate that pack-years and years since quitting improve ASCVD risk prediction compared to current tools. The findings are consistent with our earlier work demonstrating that relative to never smokers, former heavy smokers' (i.e. $\geq$20 pack-years) CVD risk remains significantly elevated beyond 5 years after smoking cessation.[3] These findings also build upon the work by Lloyd-Jones et al. which distinguished risk between former and never smokers during the first 5 years of cessation.[1,30] Here, we extended the time period for which former smokers are at elevated risk relative to never smokers and incorporated a cumulative measure of smoking history: pack-years.

These findings have important implications for patients, health care providers, and health care spending. As of 2018, there were 55 million former smokers and 34.2 million current smokers in the United States.[7,31] Our prior work demonstrates that heavy ever smokers carry excess ASCVD risk that is not currently captured with tools like the ASCVD Risk Estimator Plus. However, as our reclassification tables showed, including pack-years smoked and years since quitting in this model assigned higher risk to a larger proportion of individuals experiencing events than did the 2013 PCE. Using the number of former and current US smokers listed above and the proportions observed in our data, we would estimate that 46% of former (n=25.3 million) and 66% of current smokers (n=22.572 million) have smoked at least 20 pack-years and that 44% of heavy ever smokers are women (n=21,063,680). If 13.4%

of these heavy ever smoking men (n=3,592,315) and 7.9% of heavy smoking women (n=1,664,031) are reclassified in the correct direction (i.e., nonevents into lower risk categories, events into higher risk categories), as we observed in our data, this would equate to approximately 5.2 million Americans with $\geq$20 pack-years smoked who would be correctly reclassified under our models compared to the 2013 PCE. On an individual level, this information may motivate smokers to quit smoking or to see their provider to discuss other ways to lower their ASCVD risk if they have already quit. For health care providers, creating a more accurate risk prediction tool, particularly for heavy ever smokers, may identify high risk patients requiring additional attention such as monitoring or even early detection for CVD, akin to lung cancer screening in high risk current and former smokers.[32] Optimal detection of ASCVD risk in current and former smokers could also result in health care savings. Smoking-related illnesses cost over $300 billion per year in the United States in medical expenses and lost productivity.[33,34] Fortunately, smoking cessation is associated with decreased healthcare expenditure over time.[35–37] Finally, components of the smoking history such as pack years and years since quitting are relatively straightforward to capture at point of care and store within the electronic health record.[38,39]

The incorporation of both pack-years smoked and years since quitting in ASCVD risk estimation is necessary since they contribute different information. Pack-years reflects cumulative exposure to cigarette smoking and distinguishes risk among ever smokers, particularly current smokers, while former smokers' years since quitting quantifies the amount of ASCVD risk attributable to smoking remaining at a given assessment time.

Strengths and Limitations

The current investigation possesses several strengths, including data on the FHS Offspring cohort which spans 45 years from 1971 through December 31, 2016. Furthermore, FHS participants are under continuous surveillance for development of ASCVD events, allowing near complete ascertainment of events. The regular assessment of these individuals also enabled us to update smoking status, pack-years, and years of cessation throughout follow-up to accurately reflect the added value of these predictors on ASCVD risk prediction. Finally, rather than relying on one method, we used several metrics of model performance to assess various facets of model utility including goodness-of-fit, discrimination, and calibration.

With regard to our statistical findings, it is important to keep in mind that determination of clinical importance is not just dependent on change in the c-statistic. There is an upper limit on how well new ASCVD risk estimators can improve c-statistics over existing calculators – particularly when they include the same risk factors[40] and/or the base model already discriminates well between events and nonevents. This is true of the 2013 PCE in our sample which had a Harrell's c-statistic of 0.75 in men and 0.81 in women; as such, our model expansions did not have a marked impact on increasing the c-statistic. However, this is of little concern given that comparison of two c-statistics via Δc is a low-power procedure when many observations are censored.[41] While Δc was modest, we observed increasing Nagelkerke R^2 values as we progressively added smoking variables to the model, and observed moderate NRI(>0) values and rIDI values that were both clinically and statistically significant when compared to a base model without pack-years and years since quitting. Additionally, Nagelkerke's R^2 is scaled to the maximum attainable R^2 so that its range spans 0 to 1.[28] Since our best models only had Nagelkerke R^2 values of 7.32% and 9.92% in men and women, respectively, there still remains a large amount of variability in ASCVD risk that has not yet been explained, highlighting the need for further research in this area.

Although the data source is a strength of this investigation, the FHS Offspring cohort is a community-based cohort and the smoking patterns and sociodemographics of the participants may not reflect those of US population at large. In addition, FHS participants are mostly white individuals of

European ancestry and predominantly middle class, potentially limiting generalizability. Cigarette smoking is more prevalent and cessation rates are lower in low socioeconomic groups compared to higher ones;[42–44] current ASCVD risk assessment tools also tend to underestimate risk in these individuals.[45,46] Thus, including pack-years smoked in ASCVD risk calculators may help minimize risk underestimation. For these reasons, future research should validate these findings in a large, contemporary, and sociodemographically diverse sample. A review of coding manuals for the Jackson Heart Study, Hispanic Community Health Study, Multiethnic Study of Atherosclerosis, Coronary Artery Risk Development in Young Adults Study, and the Omni 2 and Generation 3 FHS cohorts indicates that these NHLBI-funded cohorts have the information needed to harmonize ASCVD risk factor definitions, including pack-years smoked and years since quitting, which would allow data pooling to further investigate this question.

Conclusion

In the FHS Offspring cohort, the addition of pack-years smoked and years since quitting improved ASCVD risk prediction and these variables satisfied all 6 of the AHA's criteria for evaluating a new risk predictor. 16% of women and 14% of men who experienced ASCVD events in our sample were classified into a higher risk group under our model than the 2013 PCE, which serves as the backbone of the ASCVD Risk Estimator Plus. The number of former smokers is growing, and such individuals remain at excess ASCVD risk relative to never smokers for up to 16 years after quitting. If these findings are validated in individuals of other races, ethnicities, and socioeconomic groups, modification of the ASCVD Risk Estimator Plus to include both pack-years smoked and years since quitting could mitigate underestimation of risk among millions of Americans.

CHAPTER 2

INTERACTION BETWEEN SMOKING BEHAVIORS AND POLYGENIC RISK SCORE AND IMPACT ON LUNG CANCER RISK

Introduction

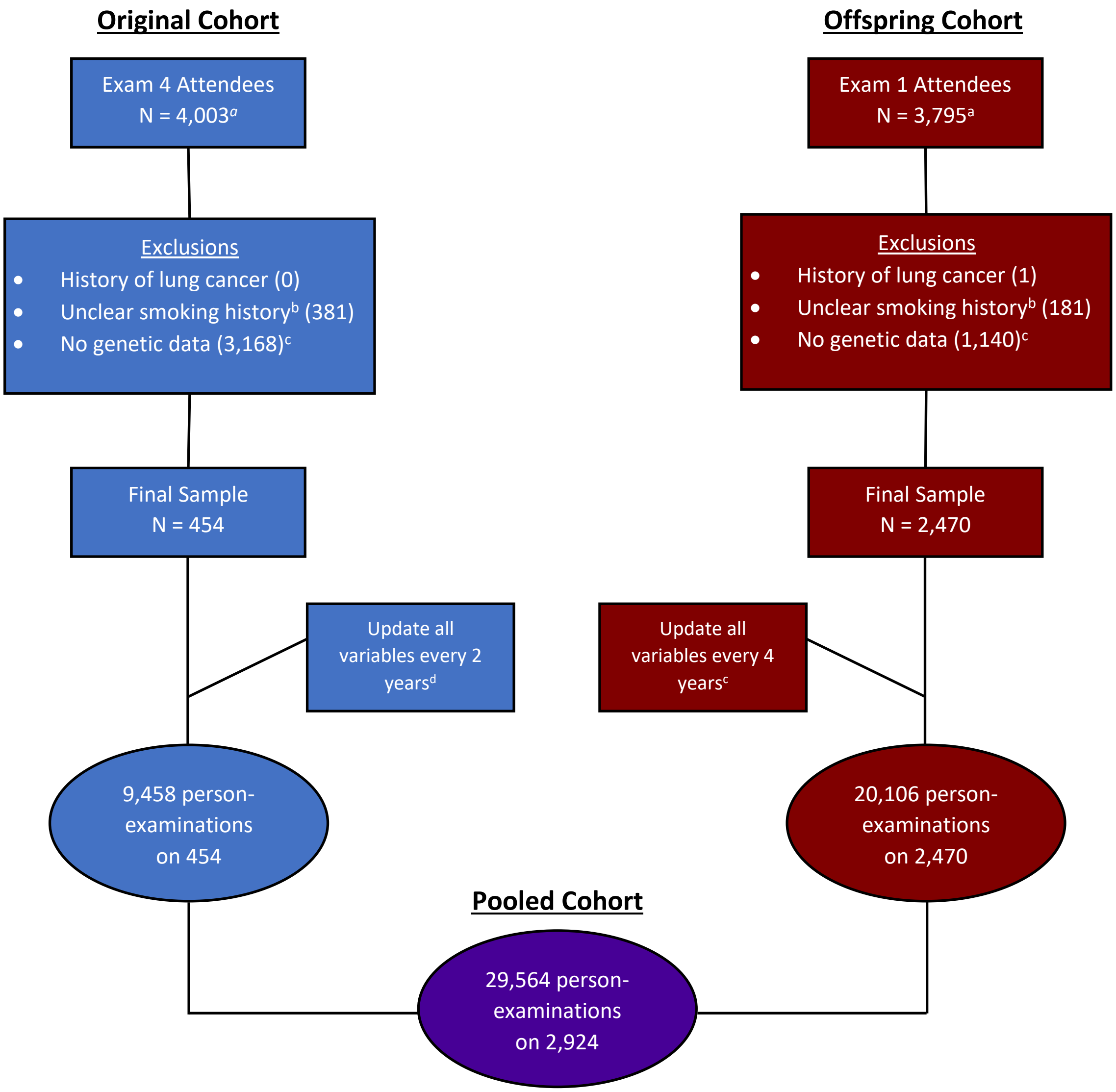
Cigarette smoking is responsible for 80-90% of lung cancer deaths[47] and the Centers for Disease Control and Prevention lists cigarette smoking as the leading cause of preventable death.[33] Smoking cessation is associated with reduced lung cancer[48] risk. In a recent study of Framingham Heart Study (FHS) participants, among those who smoked at least 20 pack-years, former heavy smokers demonstrate roughly 39% lower risk of lung cancer within 5 years of cessation compared to continuing smokers.[18] However, among these same heavier smokers, risk of lung cancer may persist beyond 25 years since quitting compared to those who never smoke.[3,18] These findings have important implications for lung cancer screening[18,49] which excludes former smokers quit for longer than 15 years.[32]

While the United States Preventive Services Task Force (USPSTF) is considering a new set of lung cancer screening guidelines which includes ever smokers with at least 20 cumulative pack-years,[50] the USPSTF currently recommends lung cancer screening via low-dose computed tomography (CT) scan for individuals between the ages of 55 and 80 (inclusive) who have accumulated at least 30 pack-years of smoking and are current smokers or former smokers quit within the preceding 15 years.[49] Tindle et al. reported that 41% of the lung cancers among former smokers in the FHS Original and Offspring cohorts occurred beyond 15 years since quitting, which exceeds the screening eligibility window.[18] It is estimated that removing the 15-year threshold from the lung cancer screening guidelines would add approximately 3 million individuals to the screening pool,[51] with large financial implications given the estimated cost of about $242 per screen.[52] Other concerns about relaxing current lung cancer screening eligibility criteria include attenuation of the anticipated benefits of screening and potential increased harm from screening individuals who are at lower risk than that observed in the National Lung Screening Trial, which informed the USPSTF recommendation.[32,49,53–55] Thus, additional strategies may be needed to identify the subset of individuals at highest risk for developing lung cancer and tailor the guidelines to best achieve this goal.

One potential avenue to identify those at high risk is the assessment of genetic risk of lung cancer. Several genome-wide association studies (GWAS) have identified genes and/or single nucleotide polymorphisms (SNPs) associated with lung cancer.[56–67] Three genes that are strongly associated with lung cancer are nicotinic acetylcholine receptors, *CHRNA3*, *CHRNB4*, and *CHRNA5*,[59,64,67,68] all of which reside on chromosome 15. These same loci are also associated with smoking behaviors.[68–70] Similarly, *CYP2A6* and *CYP2B6* reside on chromosome 19 and encode for enzymes governing nicotine metabolism, which influences cigarette consumption, and therefore, lung cancer risk.[68,71,72] Thus, it is believed that these genes are associated with lung cancer *because* they increase an individual's propensity to smoke both longer and more heavily.[73] Another mechanism of increased lung cancer risk in addition to sheer pack-year accumulation is the P450 system which metabolizes and activates tobacco-specific nitrosamines, the carcinogens in cigarettes.[74,75] Thus, those who metabolize nicotine quickly tend to smoke more and are therefore exposed to more carcinogens. These ideas support a gene-by-smoking interaction such that those with both a greater cumulative smoking history and higher burden of lung cancer risk alleles have a greater risk of lung cancer than those with an increased smoking history or high genetic risk alone. If this is the case, inclusion of a polygenic risk score (PRS) in modeling the risk of lung cancer among former heavy smokers may aid in identifying individuals who should be screened outside the current guidelines.

To this end, we applied a PRS to the FHS cohort that was used by Tindle et al. to assess the association between smoking habits and lung cancer risk.[18] We added this PRS and its interaction with smoking to models predicting lung cancer risk. These analyses allowed us to assess whether a gene-by-smoking interaction is present, and to estimate the added utility of genetic information (i.e., beyond comprehensive smoking history) to identify individuals at high risk for lung cancer development.

Figure 3: Sample Flow Diagram


Original Cohort
Exam 4 Attendees
N = 4,003[a]
Exclusions
History of lung cancer (0)
Unclear smoking history[b] (381)
No genetic data (3,168)[c]
Final Sample
N = 454
Update all variables every 2 years[d]
9,458 person-examinations on 454
Offspring Cohort
Exam 1 Attendees
N = 3,795[a]
Exclusions
History of lung cancer (1)
Unclear smoking history[b] (181)
No genetic data (1,140)[c]
Final Sample
N = 2,470
Update all variables every 4 years[c]
20,106 person-examinations on 2,470
Pooled Cohort
29,564 person-examinations on 2,924

a This is the number of individuals who attended the specified exam *and* provided genetic data

b In order to accurately capture lifetime smoking exposure, it was essential to know smoking history prior to baseline

c While genetic samples were provided by these individuals, they did not provide consent for genetic analysis by non-FHS investigators

d Original cohort participants were seen roughly every 2 years. After 5 years without an update (effectively one missed exam plus an additional year), individuals were censored to avoid carrying values forward for an extended period without reassessment. Similarly, Offspring participants were seen roughly every 4 years and were thus censored after 9 years without an update (also corresponding to a single missed exam plus an additional year).

Methods

Sample Description

This investigation includes FHS participants in the Original[76] and Offspring[9] cohorts who attended their fourth (1954-1958, n=4,541) and first (1971-1975, n=5,122) examination cycles, respectively since smoking data were initially collected from participants at these examinations. Included participants were also free of lung cancer at baseline and possessed complete data on smoking history and genetic information (*Figure 3*). Following exclusions, our analytic sample included 2,924 individuals. Participant characteristics were assessed regularly via in-person clinic examination throughout follow-up: approximately every two years for the Original cohort,[76] and every four years for the Offspring cohort.[9]

Outcome Event

All FHS participants are under continuous surveillance for the development of new cancer events.[77] The outcome of interest in this investigation is lung cancer incidence during a follow-up period from the baseline examination (examination cycle 4 for the Original cohort; examination cycle 1 for the Offspring cohort) through 2013 or 2017 for the Original and Offspring cohorts, respectively. Lung cancer cases in FHS were adjudicated by following standardized protocols which include review of medical records and pathology and laboratory reports.[77]

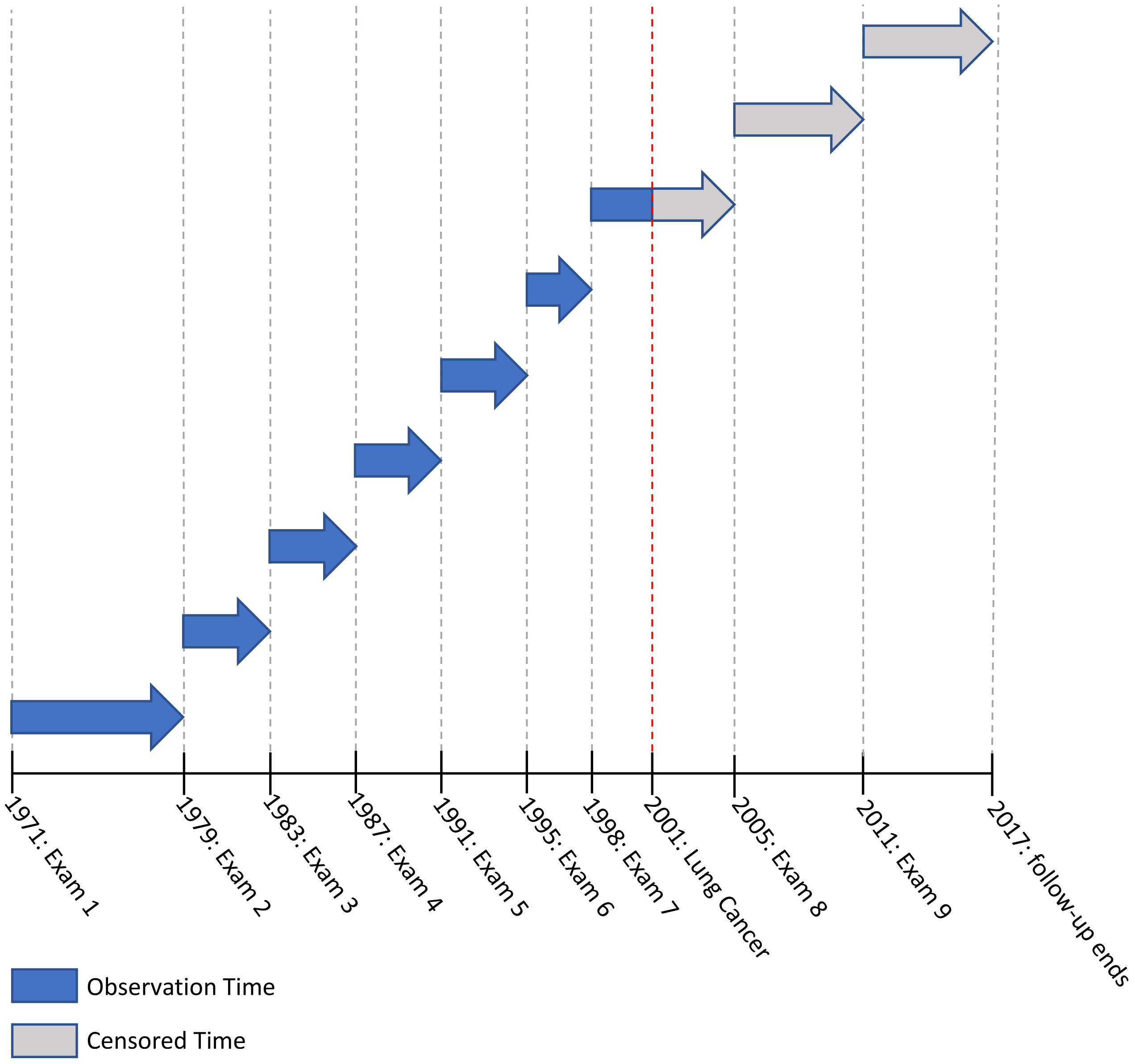
Quantification of Smoking History and Intensity

Collection and construction of smoking variables have been previously described.[3,78] At the baseline examination (cycle 4 for Original cohort, cycle 1 for Offspring cohort), data on current and prior smoking habits were collected so that participants could be categorized as “current,” “former,” or “never” smokers. For current and former smokers, we obtained information on age at which the participant starting smoking, usual number of cigarettes smoked per day in the past, age at quitting smoking (former smokers), and current number of cigarettes smoked per day. From these data, we calculated pack-years at baseline for both current and former smokers, as well as years since quitting (YSQ) for former smokers. Never smokers were assigned a pack-years value of 0, while their years since quitting value was set to missing. Pack-years and YSQ were updated at each examination in the follow-up period as described below.

At post-baseline examinations, data on current smoking status and cigarettes per day were collected, allowing us to calculate cumulative smoking exposure. For a given participant, smoking status (current, former, never) could change over time such that each participant contributed person examinations and person time to the category reflecting his or her status at each assessment. If an individual developed lung cancer, this event counted only in the group to which the individual belonged at the time of the event. The median number of examinations during which smoking was assessed was 22 for the Original cohort and 9 for the Offspring cohort.

In order to avoid carrying smoking information forward for an extended period with no update, we censored participants after a single missed exam plus an additional year without an update (5 years for Original, 9 years for Offspring). Once censored, participants were ineligible to re-enter the sample since their smoking habits in the interim were unknown. An example of timing of examinations is shown in **Figure 4**.

Figure 4: Example Timeline



Sample timeline of data collection, timing of examinations (i.e., when smoking status is updated), and lung cancer incidence in an FHS Offspring participant who attended 9 examinations and was diagnosed with lung cancer in 2001.

Genetic Data and Quality Control

For this analysis, we used genotyping data captured by the Affymetrix GeneChip Human Mapping 500K Array and the 50K Human Gene Focused Panel platforms which are included as part of the FHS SNP Health Association Resource (SHARe) data available on the database of Genotypes and Phenotypes (dbGaP).[79] Biospecimens for DNA extraction were collected from FHS participants between 1971 and 2002 and genotyped on the Affymetrix 500K and MIPS 50K arrays. Genotyping data were mapped to genome build 37.[79] Depending on when an individual's DNA was collected, there is potential for immortal time bias since there are potentially decades between the baseline examination and DNA extraction. A large amount of time between baseline and genetic testing will only bias our results if many people die from lung cancer prior to DNA collection. Thus, we examined the baseline characteristics of individuals excluded due to a lack of genetic information versus those included in the sample. After assessing the possibility of immortal person-time bias but before combining phenotypic and genotypic data, we performed quality control procedures of the genotyping data.[80]

Prior to imputation and uploading genetic data to dbGaP, standard quality control was performed.[79] Genotypes were imputed to the 1000 Genomes using MACH (version 1.00.15) and HapMap (release 22, build 36, CEU) as the imputation backbone. From the imputed data, we excluded SNPs with average call rate <90%, R^2<0.7, or minor allele frequency (<5%).[81]

We then assessed population substructure since lung cancer risk and distribution of alleles differ by genetic ancestry. FHS staff previously conducted a principle component analysis to identify population substructure on these individuals using EIGENSTRAT[82,83] . In regression models, we adjusted for the first principle component, which represents European ancestry.[80]

Polygenic Risk Score

After following the genotyping quality control protocol outlined above, we began the process of building a polygenic risk score (PRS). A PRS is the product of a weight and the corresponding number of risk alleles a person possesses at a given locus, summed over all loci (**Eq. 1**).

$$PRS_i = \sum_{j=1}^{M} W_j G_{ij} \qquad \textbf{Eq. 1}$$

where N is the total sample size; $i = 1, \ldots, N$; M is the number of SNPs contributing to the calculation of the PRS; W_j is the weight associated with the j^{th} SNP; and G_{ij} is the number of risk alleles (0, 1, or 2) that individual i possesses at locus j.

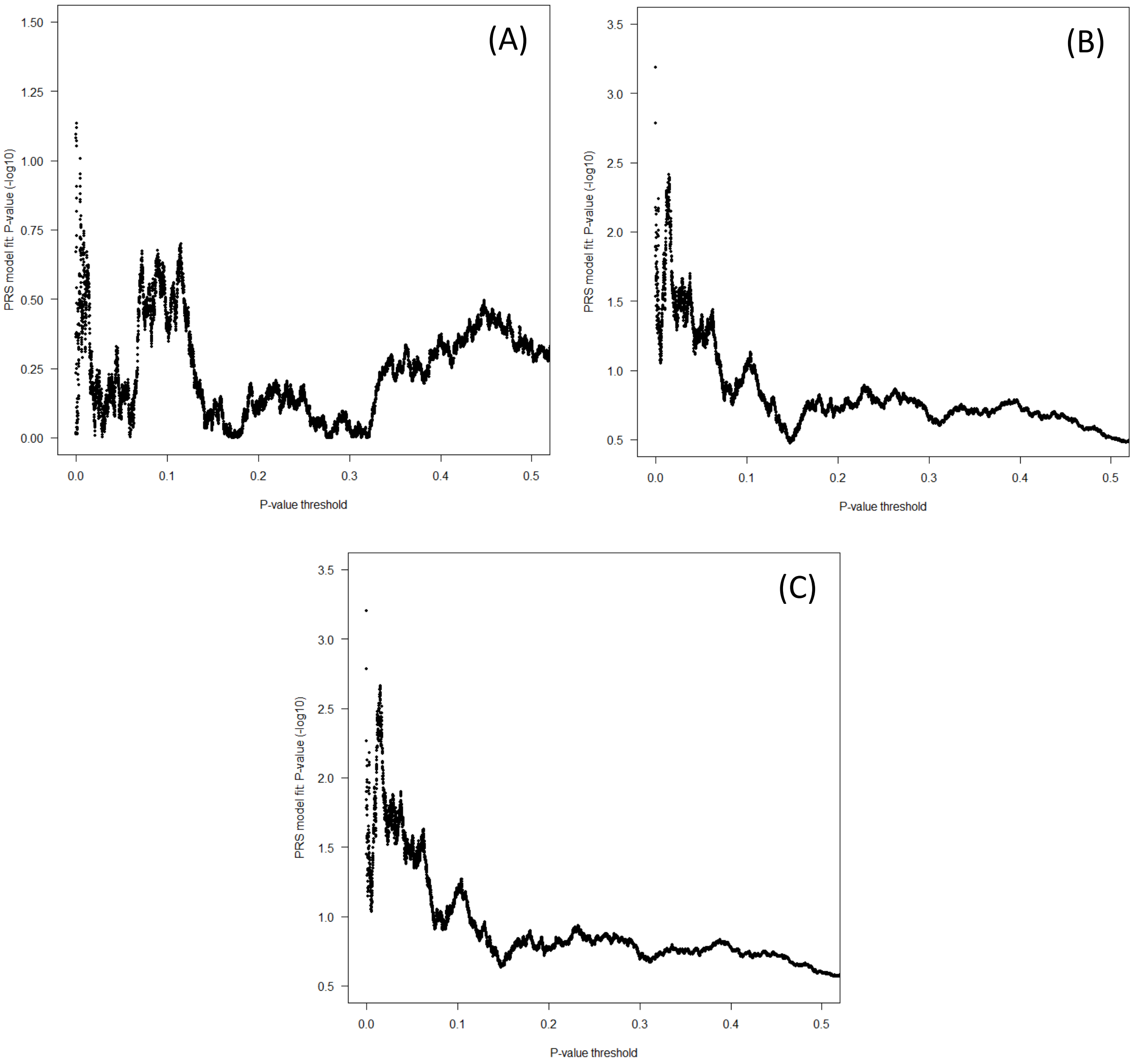
Here, we assume that SNPs do not interact with one another to influence lung cancer risk and that an additive model describes the SNPs' effect on lung cancer risk. The weight (W_j) each SNP is given is derived in a large discovery sample, while the PRS is calculated in an independent target sample. In this analysis, the Framingham Heart Study serves as the target sample and the weights were derived from a lung cancer GWAS of the OncoArray Consortium (14,803 cases; 12,262 controls) performed by McKay et al.[56] The OncoArray genotyping platform covers 533,631 SNPs that passed quality control procedures and were included as valid markers;[84] of these, 517,482 SNPs passed the filtering algorithm described by McKay et al. and were included in their analyses of lung cancer risk.[56] Genotypes were then imputed in the OnocArray Consortium using the 1000 Genomes (Phase 3) as the reference panel.[85] We determined the overlap between the OncoArray and the imputed FHS data (4,565,749 SNPs, merged based on base pair position), and built our PRS from the shared variants.

For our primary analyses, we used the R package PRSice (a p-value thresholding method) to develop our PRS and choose which SNPs to include.[86] Before performing any analyses, PRSice automatically excludes ambiguous (i.e., palindromic) SNPs; here, 708,236 ambiguous SNPs were excluded, leaving a total of 3,857,513 SNPs. PRSice further prunes the number of SNPs under consideration by removing one from each pair of SNPs in linkage disequilibrium (i.e., one from each pair of SNPs that are highly correlated), opting to retain the SNP with the greatest association with the outcome. From the 83,304 SNPs that remained after this pruning, we used PRSice to construct our PRS.

We produced three different types of PRS in PRSice: one unweighted and two weighted. Unweighted PRS may be more robust to errors in estimating weights in the discovery sample but tend to perform worse than weighted scores when the sample size increases since the sampling error does not approach zero.[87] An unweighted approach is equivalent to the sum of risk alleles at each locus included in the PRS, or a weight equal to 1 for all SNPs in **Eq. 1**. We use PRS_{unx} to denote the unweighted PRS. The second PRS (PRS_{β}) was weighted by the regression coefficient (log-odds) between the SNP and lung cancer risk, i.e., β, and the third ($PRS_{\beta/Var(\beta)}$) was weighted by the regression coefficient (log-odds) between the SNP and lung cancer risk divided by its estimated variance, i.e., β/[Var(β)] in order to incorporate a measure of uncertainty. All PRS were centered on their mean for analyses.

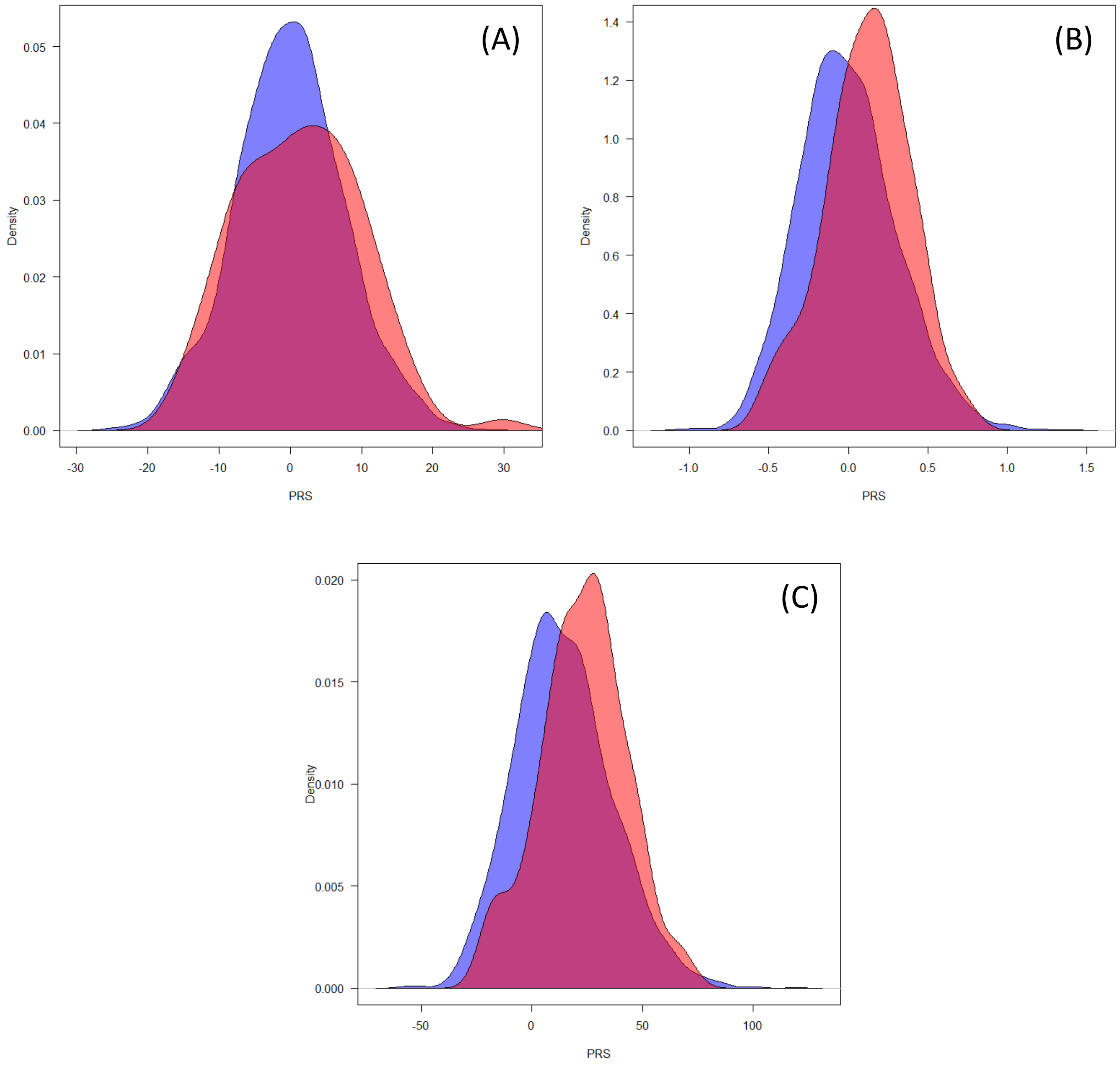
We generated high resolution plots to display the $-\log_{10}$(p-value) of the association between PRS and lung cancer across various thresholds for each of the three weighting schemes (**Figure 5**). To examine the distributions of the three PRS generated by PRSice, we then generated distribution plots stratified by incident lung cancer (**Figure 6**). We then created Manhattan plots for the two weighted PRS with the y-axis of each Manhattan plot representing the weights given to each SNP in the PRS (**Figure 7**).

Figure 5: High Resolution Plots



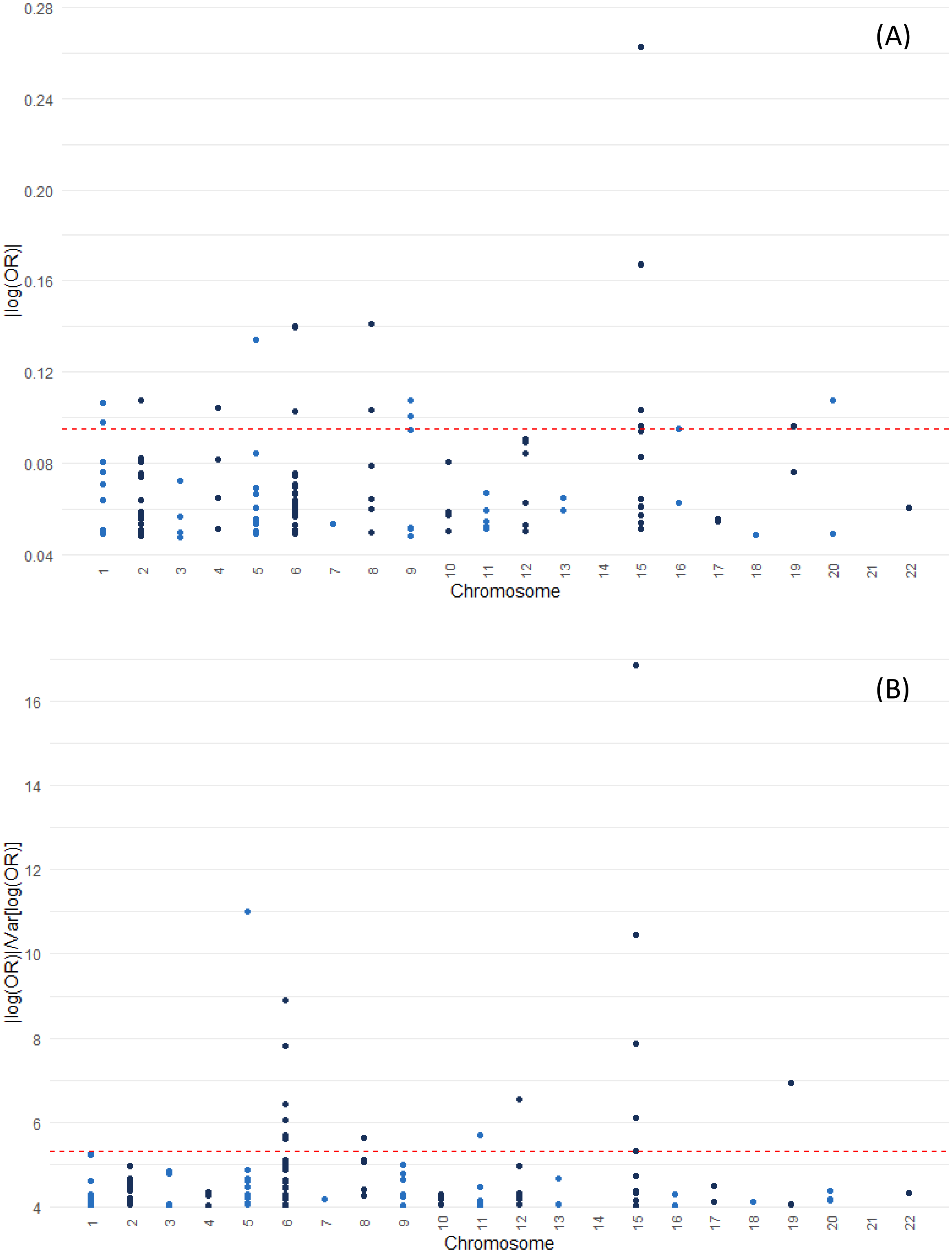
High resolution plots displaying the -log$_{10}$(p-value) of the association between lung cancer and the PRS in the FHS sample (y-axis) versus the threshold below which all SNPs with p-values for their individual association with lung cancer in the base sample (OncoArray Consortium) were included (x-axis). i.e., if 20 SNPs had a p-value<5×10^{-8} for their association with lung cancer in the OncoArray Consortium, and a PRS comprised of those 20 SNPs in FHS had a p-value of 0.2 when assessing it's association with lung cancer, then we would plot a point at (x,y)=(5×10^{-8}, 0.70). Panel A corresponds to PRS$_{unw}$, Panel B correspods to PRS$_{\beta}$, and Panel C corresponds to PRS$_{\beta/Var(\beta)}$

Figure 6: Distribution plots of polygenic risk scores by event status



Blue shading displays the distribution of the PRS among controls; red is among cases. Clockwise, panels display PRS_{unw} (panel A), PRS_{β} (panel B), and $PRS_{\beta/Var(\beta)}$ (panel C)

Figure 7: Manhattan Plots



Panel A shows the weight given to each SNP included in PRS_{β} with a reference corresponding to an odds ratio of 1.1; Panel B shows the weight assigned to each SNP contributing to $PRS_{\beta/Var(\beta)}$ with a reference line corresponding to a genome-wide significance level equal to 5×10^{-8}

In sensitivity analyses, we included all 83,304 SNPs remaining after LD pruning (R^2>0.8) to build PRS rather than using p-value thresholding to choose a subset of SNPs. We again built the three PRS (one unweighted, two weighted) as described above.

Confounders

The directed acyclic graph (DAG) in **Figure 8** displays the relationships we believe exist in these data based on prior research.[56,68,88–94] A confounder is a variable that is associated with both the exposure and outcome and does not lie along the causal pathway. Based on the DAG above, population structure, age, sex, lung cancer genetics, family history of lung cancer, and education (via socioeconomic status) meet this definition of a confounder. Thus, we adjusted for these variables in statistical models to produce an unbiased estimate of the association between smoking and lung cancer. However, family history of lung cancer is unmeasured in FHS data, and therefore this confounding path could not be closed. For our conclusions to be valid, we must assume that the bias incurred by not adjusting for family history of lung cancer is small. This assumption is likely valid given that only 0.2% of the American population has a current or prior diagnosis of lung cancer.[95]

Quantification of population structure and lung cancer genetics (via PRS) were described above. Age was calculated at each FHS examination using a participant's verified date of birth and examination date. Sex and education were self-reported.

Missing Phenotypic Data

In the FHS phenotypic data, missingness was relatively low. In the Original cohort, 87% of person-exams had no missing data while in the Offspring cohort, 98% of person-exams had no missing data. Education level was most frequently missing in the Offspring cohort with a mere 2.1% missingness. Smoking status was not assessed at all Original cohort exams and was thus missing at 13% of person-examinations. However, smoking status was collected at all Offspring examinations and was therefore missing at <1% of person-exams in this cohort. All other variables had <5% missingness. Missing data was handled using multiple imputation by chained equations techniques to produce five complete datasets for analysis. We imputed continuous variables thorough the use of predictive mean matching[96] in order to produce imputed values that are clinically plausible. Categorical variables were imputed using the discriminate function with a non-informative Jeffrey's prior.[97] Results across imputed datasets were combined according to Rubin's rules.[98]

Statistical Analysis

We calculated baseline summary statistics in each FHS cohort separately and pooled. Means and standard deviations (SD) were reported for normally distributed continuous variables, while medians along with the 25th and 75th percentile were reported for continuous variables with skewed distributions. We report counts and percentages for categorical variables. Data from the Original and Offspring cohorts were pooled for all further analyses. Using Poisson regression with an offset term equal to the natural-logarithm of follow-up time, we calculated lung cancer incidence rates per 1000 person-years stratified by smoking status; current and former smokers were further categorized by above/below 20 pack-years.

After asserting that the proportional hazards assumption was not seriously violated via the interaction of pack-years and PRS (separately) with the natural logarithm of follow-up time, we fit Cox proportional hazards regression models with incident lung cancer as the outcome. Because the FHS includes related individuals, we used mixed-effects Cox proportional hazards regression in our analyses that incorporated the kinship matrix – a symmetric matrix of dimension equal to the number of subjects in the sample (2,924) where each entry represents the proportion of genetic information shared between each pair of individuals based on their familial relation – to adjust the variance accordingly.

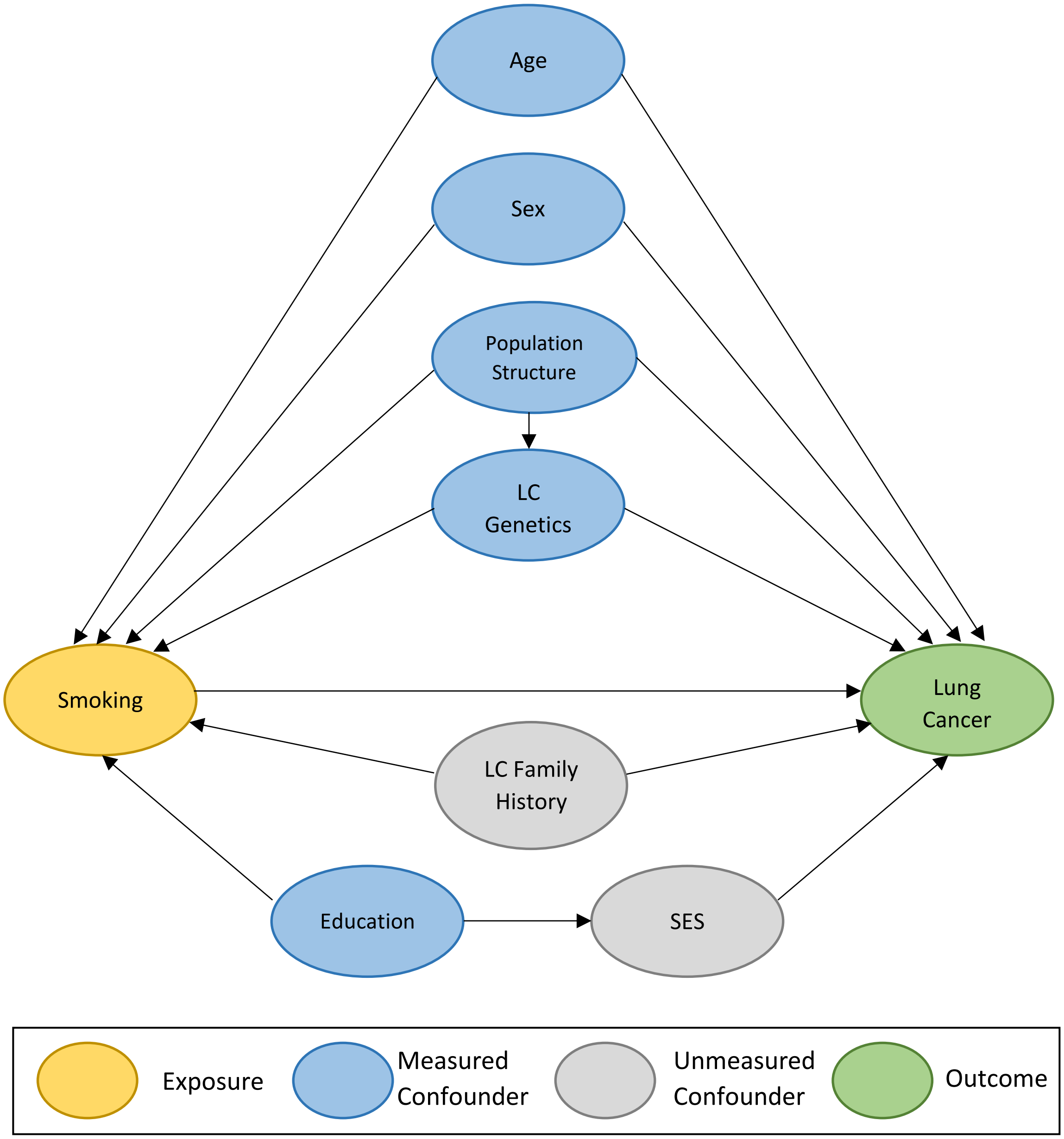
Models included pack-years, an indicator of current smoking status, PRS, population structure, age, sex, and education as independent variables. We first assessed whether PRS modifies the effect of the association between pack-years and lung cancer risk. Along with the interaction between pack-years smoked and PRS, models included the main effect of pack-years, an indicator of current smoking status, PRS, population structure, age, sex, and education as independent variables. This test of heterogeneity was performed for all 3 versions of the PRS: PRS_{unw}, PRS_{β}, and $PRS_{\beta/Var(\beta)}$.

After we determined that there was an interaction between pack-years smoked and PRS, we calculated the effect of pack-years at different PRS values: the mean, mean + 1SD, and mean - 1SD. We then examined the impact of PRS on lung cancer among never smokers, ever smokers with <20 pack-years, and ever smokers with $\geq$20 pack-years; the 20 pack-year cut point was chosen to be consistent with our prior work examining the impact of smoking in this cohort and the new USPSTF guidelines under consideration.[3,18,50] For the full sample and each subset (never smokers, ever smokers <20 pack-years, ever smokers $\geq$20 pack-years), we fit a Cox proportional hazards regression model for lung cancer that included PRS and population structure as predictors to avoid the "Table 2 Fallacy"[99] which occurs when the effect estimates of secondary exposures are presented in the same manner as the primary exposure estimated from the same model.[95] Thus, in these models, we did not adjust for other variables since only population structure can confound the association between the PRS and lung cancer risk (**Figure 8**).

Finally, as an exploratory analysis, we sought to determine whether any of the PRS we developed could identify individuals who developed lung cancer outside of the USPSTF lung cancer screening guidelines under consideration.[50] Here, we limited our models to the 80 ever smokers who developed lung cancer and performed logistic regression analyses with "eligible for lung cancer screening" as our dependent variable predicted from the PRS and population structure. The above analyses were repeated using 3 genome-wide PRS that incorporated all 83,304 SNPs remaining after LD pruning and are included in supplemental results.

A two-sided p-value <0.05 was considered statistically significant except for tests for heterogeneity which considered a two-sided p-value <0.2 significant since such tests are typically underpowered. Analyses were performed in SAS 9.4 (Cary, NC) and R 4.0.1. All Cox regression was performed in R and used the “coxme” function in the “kinship2” package to fit models that accounted for the relatedness among FHS participants.

Figure 8: Directed Acyclic Graph



Displays the presumed relationships among smoking, lung cancer risk, and the confounders and effect modifiers of this association.

LC: Lung Cancer; SES: Socioeconomic status

Results

Sample Characteristics

Our sample included 2,924 members of the FHS – 454 from the Original cohort, and 2,470 Offspring cohort participants. At baseline, Original cohort participants had an average age of 44 years compared to the Offspring cohort, which had an average age of 34 years (**Table 5**). Both cohorts were more than 50% female. 47% of Original cohort members were current smokers at baseline compared to 40% of Offspring cohort members. Only 6% of Original cohort members were former smokers, while 20% of the Offspring cohort had previously smoked. Cigarettes per day and pack-years among former smokers were similar between cohorts; current smokers in the Original cohort had a median of 17 pack-years at baseline compared to 12.5 among Offspring cohort participants. Polygenic risk score distributions were similar between cohorts (**Table 5**).

Polygenic Risk Scores

Using the p-value thresholding method implemented in PRSice, we determined that including the 638 SNPs with a p-value less than 0.00076 produced an unweighted PRS with the highest R^2 (**Figure 5 Panel A**), and that a PRS including the 120 SNPs with a p-value less than $6x10^{-5}$ produced weighted PRS with the greatest correlation with lung cancer (**Figure 5 Panels B** and **C**). When examining the PRS distributions by incident lung cancer status, we found that the distribution of PRS_{unw} was quite similar between events and non-events (**Figure 6 Panel A**), but that PRS_{β} and $PRS_{\beta/Var(\beta)}$ distributions were slightly shifted to the right in events versus non-events indicating a greater burden of risk alleles in events than non-events (PRS_{β}, **Figure 6 Panel B**; $PRS_{\beta/Var(\beta)}$, **Figure 6 Panel C**). Manhattan plots displaying the chromosome locations (x-axis) and weights of the PRS (y-axis) are displayed in **Figure 7**. The top panel displays the weights for PRS_{β}, while the bottom panel corresponds to $PRS_{\beta/Var(\beta)}$. Both weighted PRS include SNPs with large weights on chromosome 15 (rs72738786, rs11072774, rs12907065) ;

$\text{PRS}_{\beta/\text{SE}2}$ also assigned large weights to SNPs on chromosome 5 (rs6554758), 6 (rs116822326, rs115375792), and 19 (rs11667314).

Table 5: Sample Characteristics at Baseline

Characteristic[a]	Pooled Cohort (Total N = 2924)		Original Cohort (Total N = 454)		Offspring Cohort (Total N = 2470)	
	N	Summary	N	Summary	N	Summary
Age, years	2924	35.4 (9.8)	454	43.5 (6.1)	2470	34.0 (9.7)
Sex	2924	--	454	--	2470	--
Male	--	1286 (44.0)	--	172 (37.9)	--	1114 (45.1)
Female	--	1638 (56.0)	--	282 (62.1)	--	1356 (54.9)
Education	2585	--	452	--	2133	--
Less than High School Graduate	--	253 (9.8)	--	141 (31.2)	--	112 (5.3)
High School Graduate	--	865 (33.5)	--	174 (38.5)	--	691 (32.4)
More than High School	--	1467 (56.8)	--	137 (30.3)	--	1330 (62.3)
Systolic Blood Pressure, mmHg	2923	120.0 (14.9)	454	123.5 (16.9)	2469	119.4 (14.5)
Diastolic Blood Pressure, mmHg	2923	77.6 (10.1)	454	79.7 (10.2)	2469	77.3 (10.0)
Antihypertensive Medication	2922	64 (2.2)	454	10 (2.2)	2468	54 (2.2)
Hypertension	2922	456 (15.6)	454	89 (19.6)	2468	367 (14.9)
Body Mass Index, kg/m^2	2923	24.2 (21.9, 27.0)	453	24.7 (22.6, 27.3)	2470	24.1 (21.7, 26.9)
Diabetes	2895	13 (0.5)	446	1 (0.2)		12 (0.5)
Total Cholesterol, mg/dL	2910	198.8 (40.4)	449	228.5 (43.7)	2461	193.4 (37.4)
Smoking Status	2924	--	454	--	2470	--
Current	--	1170 (40.0)	--	211 (46.5)	--	959 (38.8)
Former	--	520 (17.8)	--	25 (5.5)	--	495 (20.0)
Never	--	1234 (42.2)	--	218 (48.0)	--	1016 (41.1)
Cigarettes per day[c]	1234	20.0 (10.0, 30.0)	218	20.0 (9.0, 20.0)	1016	20.0 (10.0, 30.0)
Pack-Years	--	--	--	--	--	--
Current Smokers	1234	13.4 (5.4, 24.0)	218	17.1 (6.9, 24.8)	1016	12.5 (5.0, 24.0)
Former Smokers	520	11.0 (4.0, 22.2)	25	10.2 (3.8, 22.6)	495	11.0 (4.0, 22.0)
Years Since Quitting[d]	520	6.0 (3.0, 10.0)		1.0 (1.0, 1.1)	495	6.0 (3.0, 10.0)
Polygenic Risk Score	--	--	--	--	--	--
Unweighted	2924	165.0 (160.0, 170.0)	454	166.0 (161.0, 170.0)	2470	165.0 (160.0, 170.5)
Weighted by β	2924	0.2 (0.0, 0.4)	454	0.2 (0.0, 0.4)	2470	0.2 (0.0, 0.4)
Weighted by β/Var(β)	2924	13.2 (-0.7, 28.3)	454	14.4 (-1.4, 28.9)	2470	13.1 (-0.6, 28.1)

a Summary statistics are displayed as Mean (SD) for age, systolic blood pressure, diastolic blood pressure, and total cholesterol, as Median (Q1, Q3) for body mass index, cigarettes per day, pack-years, years since quitting, and polygenic risk scores, and as N (%) for categorical variables.

b Self-reported consumption of at least one alcoholic beverage per month.

c Among current smokers only.

d Among former smokers only.

Effect of Pack-Years Smoked and PRS on Lung Cancer Risk

We first assessed the lung cancer incidence rate by smoking status and intensity. Among the 2,924 individuals in our sample, 86 were diagnosed with lung cancer: 6 never smokers, 48 former smokers, and 32 current smokers. We observed that both former and current smokers had higher lung cancer incidence rates than never smokers, with the highest rates in current smokers (**Table 6**). When stratifying smokers' incidence rates by above/below 20 pack-years, point estimates were slightly higher in former smokers in both categories than current smokers, but confidence intervals overlapped, indicating no statistically significant difference.

Table 6: Lung Cancer Incidence by Smoking Status and Intensity

Smoking Status	Person-Examinations	Person-Years	Lung Cancers	Incidence Rate per 1000PY [95% CI]
Never	10,848	44,532	6	0.13 [0.06, 0.30]
Former	9,975	42,904	48	1.12 [0.84, 1.49]
< 20 PKY	*5,696*	*25,537*	*6*	*0.26 [0.12, 0.57]*
≥ 20 PKY	*4,279*	*17,367*	*42*	*2.39 [1.76, 3.24]*
Current	6,114	25,194	32	1.27 [0.90, 1.80]
< 20 PKY	*2,394*	*10,912*	*1*	*0.09 [0.01, 0.65]*
≥ 20 PKY	*3,720*	*14,282*	*31*	*2.17 [1.53, 3.09]*

Cells are time-updated such that as individuals begin and quit smoking, they contribute person-time to various groups. An individual's lung cancer event only contributes to the group he or she was in at the time of diagnosis. Incidence rates and corresponding 95% confidence intervals use data from all 5 multiple imputations. Other columns are based on the first imputation alone.

We confirmed that the proportional hazards assumption was not severely violated by evaluating interactions of pack-years smoked and PRS (individually) with the natural logarithm of follow-up time (all p-values>0.2). We then assessed the presence of a gene-by-smoking interaction between PRS and pack-years smoked on lung cancer risk by performing tests for heterogeneity for each PRS interacted with pack-years. All p-values were <0.2 (PRS_{unw}×PKY p=0.18; PRS_{β}×PKY p=0.09; $\text{PRS}_{\beta/\text{Var}(\beta)}$×PKY p=0.10), indicating presence of heterogeneity. Because of this heterogeneity, we analyzed the effect of pack-years at the Mean, Mean-1SD, and Mean+1SD for each PRS, and assessed the impact of each PRS within strata of pack-years. Each additional 10 pack-years was associated with a 56% increase in the risk of lung cancer at one SD below the mean of each PRS, a 48% increase at the mean PRS, and a 40% increase at one SD above the mean of each PRS (**Table 7**). In other words, as each PRS increased, the effect of an additional pack-year smoked conferred a smaller risk of lung cancer.

Table 7: Effect of Pack-Years at Varying PRS Values

PRS Value	HR [95% CI] per 10 Pack-Years	p-value
PRS$_{unw}$		
Mean – 1SD	1.53 [1.38, 1.70]	<0.0001
Mean	1.46 [1.35, 1.58]	<0.0001
Mean + 1SD	1.39 [1.25, 1.54]	<0.0001
PRS$_{\beta}$		
Mean – 1SD	1.56 [1.41, 1.73]	<0.0001
Mean	1.48 [1.37, 1.60]	<0.0001
Mean + 1SD	1.40 [1.26, 1.55]	<0.0001
PRS$_{\beta/Var(\beta)}$		
Mean – 1SD	1.56 [1.49, 1.63]	<0.0001
Mean	1.48 [1.37, 1.60]	<0.0001
Mean + 1 SD	1.40 [1.27, 1.56]	<0.0001

Hazard Ratios are estimated from mixed-effects Cox proportional hazards regression models adjusting for pack-years, PRS, the interaction between pack-years and PRS, age, sex, current smoking status, education, and population structure. The variance structure accounts for familial relationships.

When examining the association between PRS and lung cancer within strata of pack-years (never smokers, ever smokers<20 pack-years, ever smokers$\geq$20 pack-years) we observed no association between any of the PRS and risk of lung cancer (**Table 8**). However, since number of events were low, we also assessed this relationship in the full cohort and observed that each standard deviation higher in PRS$_{\beta}$ or PRS$_{\beta/Var(\beta)}$ was associated with a 26% higher risk of lung cancer (**Table 8**).

Table 8: Effect of Polygenic Risk Score on Lung Cancer by Smoking Status

Polygenic Risk Score	Hazard Ratio [95% CI] per SD increase in PRS	p-value
PRS_{unw} (SD = 7.65)		
Full Sample	1.10 [0.89, 1.36]	0.39
Never Smokers	*1.55 [0.72, 3.35]*	*0.26*
Ever Smokers <20 PKY	*0.85 [0.38, 1.89]*	*0.69*
Ever Smokers ≥20 PKY	*1.05 [0.83, 1.33]*	*0.67*
PRS_{β} (SD = 0.31)		
Full Sample	1.26 [1.01, 1.56]	0.04
Never Smokers	*1.56 [0.75, 3.25]*	*0.23*
Ever Smokers <20 PKY	*1.47 [0.73, 2.95]*	*0.28*
Ever Smokers ≥20 PKY	*1.13 [0.89, 1.44]*	*0.32*
$PRS_{\beta/Var(\beta)}$ (SD = 21.92)		
Full Sample	1.26 [1.01, 1.56]	0.04
Never Smokers	*1.57 [0.76, 3.28]*	*0.23*
Ever Smokers <20 PKY	*1.37 [0.66, 2.81]*	*0.40*
Ever Smokers ≥20 PKY	*1.14 [0.90, 0.28]*	*0.28*

Estimates are from a mixed-effects Cox proportional hazards regression model adjusted for population structure; the variance structure accounts for familial relationships. Hazard Ratios and corresponding 95% confidence intervals are estimated per standard deviation increase in polygenic risk score.

Exploratory and Sensitivity Analyses

Among the 80 ever smokers who developed lung cancer, 31 individuals developed lung cancer during a period when they met the USPSTF lung cancer screening guidelines under consideration;[50] the remaining 61% did not meet lung cancer screening eligibility criteria at the time of diagnosis. In logistic regression models to determine whether any of the 3 PRS were associated with lung cancer screening eligibility at the time of diagnosis, we observed no significant results (**Table 9**).

Table 9: Association between PRS and Lung Cancer Screening Eligibility

PRS Value	Odds Ratio [95% CI]	p-value
PRS_{unw}	0.96 [0.61, 1.51]	0.85
PRS_{β}	0.94 [0.57, 1.55]	0.81
$PRS_{\beta/Var(\beta)}$	0.87 [0.53, 1.43]	0.59

Estimates are from a logistic regression model with a binary indicator for "met USPSTF lung cancer screening criteria under consideration" as the outcome. PRS and population structure were included as independent variables. Odds Ratio is per SD increase in the PRS.

In sensitivity analyses where we constructed the 3 PRS from the 83,304 SNPs that remained following LD pruning without p-value thresholding, we observed that the distribution of all the PRS were almost completely overlapping between events and non-events (**Appendix B, Supplemental Figure 5**). As such, we did not observe an interaction between any PRS and pack-years smoked on lung cancer risk (PRS_{unw}×PKY p=0.88; PRS_{β}×PKY p=0.70; $PRS_{\beta/Var(\beta)}$×PKY p=0.77), nor did we observe an association between PRS and lung cancer risk in the full sample (**Appendix B, Supplemental Figure 6** and **Supplemental Table 6**). However, consistent with the primary analyses, the association between pack-years smoked and lung cancer risk remained highly significant (**Appendix B, Supplemental Table 5**). There remained no association between the weighted PRS and odds of meeting the USPSTF lung cancer screening guidelines under consideration at the time of diagnosis among ever smokers, but each standard deviation increase in the unweighted PRS was associated with 2.18 times the odds of meeting the USPSTF lung cancer screening guidelines at the time of diagnosis among ever smokers (95% CI: [1.36, 3.49]; **Appendix B, Supplemental Table 7**).

Discussion

In our sample, the overall effect of the PRS was positively associated with lung cancer such that each standard deviation increase in PRS was associated with 26% higher risk of lung cancer. Additionally, we confirmed that pack-years remains a strong risk factor for lung cancer, even after accounting for genetic contribution. We are among the first to identify an interaction between pack-years smoked and PRS on lung cancer risk in a prospective sample, so the effects of these variables should be interpreted in that context. After accounting for the interaction, pack-years remained significantly associated with lung cancer incidence, but its effect decreased with increasing PRS. At the mean value of the PRS, each additional 10 pack-years was associated with a 48% increase in lung cancer risk while at a PRS value 1 standard deviation above the mean, the effect decreased to a 40% increase in lung cancer risk per 10 additional pack-years. When we assessed the effect of PRS on lung cancer risk within strata of pack-years, the association was positive but not statistically significant due to small numbers of events (6 lung cancers among never smokers, 7 in ever smokers with <20 pack-years, and 73 in ever smokers with $\geq$20 pack-years). None of our PRS were associated with development of lung cancer in individuals who were ineligible for lung cancer screening under the current USPSTF recommendations under consideration.[50]

Presence of a gene-by-smoking interaction has been observed before,[100,101] but few have examined the interaction between a lung cancer PRS and smoking.[101] To our knowledge, this is the first of such investigations to observe this interaction in a prospective cohort with multiple assessments of smoking habits over several decades and adjudicated lung cancer incidence. While VanderWeele et al. were among the first to report a significant gene-by-smoking interaction on lung cancer, they examined only 2 SNPs on chromosome 15.[100] Additionally, Qian et al. observed significant effect modification when incorporating principle components to represent the main genetic effect and its interaction with pack-years smoked, but not when incorporating genotypes from the target sample. Our results connect and bolster these prior findings by incorporating genetic information across the genome and using genotypes from the target sample rather than principle components.

Our finding that a PRS is associated with lung cancer risk is also consistent with prior findings. Jia et al. used data from 400,812 participants in the UK Biobank to build a PRS for lung cancer based on 19 SNPs and observed that those with a PRS in the highest quintile had 1.71 times the risk of lung cancer compared to individuals with a PRS in the lowest quintile.[102] However, they did not adjust for smoking status or pack-years smoked and did not assess for effect modification.[102] Dai et al. also constructed a PRS containing 19 SNPs for lung cancer in a prospective sample of Chinese men and women and found it to be associated with diagnosis of lung cancer.[103] However, while they did not directly test for effect modification of this association by pack-years smoked, they did perform stratified analyses in nonsmokers as well as heavy ($\geq$30 pack-years) and light smokers (<30 pack-years), and observed synergistic effect-modification such that those with the high genetic risk and a $\geq$30 pack-year smoking history were at the greatest lung cancer risk.[103] Although our results differed from those of Dai et al. in that we did not observe significant associations of the PRS in strata of smokers, our data did have more statistical support on the side of a direct association between the weighted PRS and lung cancer. Given that their sample included more than 95,000 individuals compared to less than 3,000 in our sample, the lack of a statistically significant association is more likely due to insufficient power than the true absence of an association.

The current results also contribute to the discussion surrounding the construction of PRS. Since the OncoArray Consortium had a large sample for their GWAS, the weights we used in constructing our weighted PRS were well-defined. As such, PRS_{β} and $PRS_{\beta/Var(\beta)}$ were more strongly associated with lung cancer risk than PRS_{unw} and support the use of weighted PRS over unweighted PRS. There is also debate surrounding whether PRS should be constructed from a subset of SNPs via p-value thresholding or other methods, or if the entire genome should be used for building PRS.[104–106] There is a movement toward using the full genome to produce PRS for traits that are truly polygenic like cardiovascular disease,[105] which has thousands of regions across the genome that contribute small amounts of information regarding its development. However, our results demonstrate that while such an approach works for diseases that are highly polygenic, in the case of lung cancer, our weighted PRS using 120 SNPs was more highly associated with lung cancer than our weighted PRS that used all 83,304 SNPs that remained following LD pruning. Although more research is needed, our results caution against thinking that inclusion of more SNPs in PRS construction is always better.

Our results suggest more work is needed to determine whether incorporating a PRS into the lung cancer screening process would be fruitful. It is possible that with a much larger sample, like a reanalysis of the UK Biobank data altered to include a genome-wide lung cancer PRS and its interaction with pack-years, a PRS may aid in identifying individuals diagnosed with lung cancer outside the current screening criteria. However, in our sample of 80 ever smokers who developed lung cancer, 49 of whom were not eligible for screening at the time of diagnosis, not only were the results non-significant, the odds ratio was also close to the null value of 1. Another potential way to incorporate genetics and/or family history into the screening process is to simply ask ever smoking patients if they have a first or second degree relative who has developed lung cancer outside the current guidelines. Among the 80 ever smokers who developed lung cancer in our sample, 21 of them were related to at least one other ever smoker who developed lung cancer. These 21 individuals were spread across 10 families: 9 with 2 members developing lung cancer, 1 with 3 members developing lung cancer. Out of the 10 families, 7 of them had all members either eligible or not eligible for lung cancer screening based on the USPSTF recommendations under consideration at the time of diagnosis. While more research is certainly warranted, this could be one potential option.

Although our results offer suggested next steps in this line of research, their impact is somewhat limited due to lack of power resulting from a relatively small sample size and low number of events.

Existence of relatedness among participants further decreased our effective sample size and power. In addition, we were unable to completely close all confounding pathways between smoking and lung cancer. In particular, although FHS is a study that includes family members, there is no efficient way to adjust for family history of lung cancer since participants are never asked about this specifically and only about half of the parents of Offspring cohort participants are part of the Original cohort. We also assumed that adjusting for highest education achieved prior to the baseline examination was a suitable proxy for socioeconomic status. However, socioeconomic status is multifaceted and adjustment for educational attainment at a single point in time is likely insufficient to completely account for the confounding effect of socioeconomic status in the association between smoking status and lung cancer risk. We were also unable to include important regions of the genome, including the *CYP2A6* nicotine metabolism gene, which is associated with both smoking behavior and lung cancer development, in our PRS since this region is notoriously challenging to genotype and was not covered by our chip or imputation panel.[107] Finally, FHS Original and Offspring cohort participants are mostly white and predominantly of European ancestry so results may not generalize to individuals of other genetic ancestry.

In conclusion, our results support the presence of a gene-by-smoking interaction on the effect of lung cancer incidence and reinforce the negative effect of continued smoking on lung cancer risk even in those with low genetic risk. These results are consistent with prior findings, but it remains unclear whether incorporating genetic information into routine lung cancer risk assessment is of value. Larger studies with both genetic data and longitudinal smoking information should investigate this further.

REFERENCES

1. Lloyd-Jones DM, Huffman MD, Karmali KN, et al. Estimating Longitudinal Risks and Benefits from Cardiovascular Preventive Therapies among Medicare Patients: The Million Hearts Longitudinal ASCVD Risk Assessment Tool: A Special Report from the American Heart Association and American College of Cardiolog. *Circulation*. 2017;135(13):e793-e813. doi:10.1161/CIR.0000000000000467

2. ASCVD Risk Estimator Plus. http://tools.acc.org/ASCVD-Risk-Estimator-Plus/#!/calculate/estimate/.

3. Duncan MS, Freiberg MS, Jr RAG, Kundu S, Vasan RS, Tindle HA. Association of Smoking Cessation With Subsequent Risk of Cardiovascular Disease. *J Am Med Assoc*. 2019;322(7):642-650. doi:10.1001/jama.2019.10298

4. Goff DC, Lloyd-jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *J Am Coll Cardiol*. 2014;63(25):2886. doi:10.1016/j.jacc.2014.02.606

5. Lee PN, Fry JS, Thornton AJ. Estimating the decline in excess risk of cerebrovascular disease following quitting smoking – A systematic review based on the negative exponential model. *Regul Toxicol Pharmacol*. 2014;68(1):85-95. doi:10.1016/J.YRTPH.2013.11.013

6. Lee PN, Fry JS, Hamling JS. Using the negative exponential distribution to quantitatively review the evidence on how rapidly the excess risk of ischaemic heart disease declines following quitting smoking. *Regul Toxicol Pharmacol*. 2012;64(1):51-67. doi:10.1016/J.YRTPH.2012.06.009

7. Creamer MLR, Wang TW, Babb S, et al. Tobacco Product Use and Cessation Indicators Among Adults - United States, 2018. *MMWR Morb Mortal Wkly Rep*. 2019;68(45):1013-1019. doi:10.15585/mmwr.mm6845a2

8. Jamal A, Phillips E, Gentzke AS, et al. Current Cigarette Smoking Among Adults — United States, 2016. *MMWR Morb Mortal Wkly Rep*. 2018;67(2):53-59. doi:10.15585/mmwr.mm6702a1

9. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. *Am J Epidemiol*. 1979;110(3):281-290. doi:10.1093/aje/kwx110

10. Tsao CW, Vasan RS. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *Int J Epidemiol*. 2015;44(6):1800-1813. doi:10.1093/ije/dyv337

11. Lloyd-Jones DM, Martin DO, Larson MG, Levy D. Accuracy of Death Certificates for Coding Coronary Heart Disease as the Cause of Death. *Ann Intern Med*. 1998;129(12):1020-1026.

12. Fox CS, Evans JC, Larson MG, et al. A comparison of death certificate out-of-hospital coronary heart disease death with physician-adjudicated sudden cardiac death. *Am J Cardiol*. 2005;95(7):856-859. doi:10.1016/j.amjcard.2004.12.011

13. Carson E, Hemenway AN. Recent Evidence Examining Efficacy and Safety of Aspirin for Primary Cardiovascular Disease Prevention. *Ann Pharmacother*. 2019;53(7):738-745. doi:10.1177/1060028018825140

14. McNeil JJ, Woods RL, Nelson MR, et al. Effect of Aspirin on Disability-free Survival in the Healthy Elderly. *N Engl J Med*. 2018;379(16):1499-1508. doi:10.1056/NEJMoa1800722

15. McNeil JJ, Wolfe R, Woods RL, et al. Effect of Aspirin on Cardiovascular Events and Bleeding in the Healthy Elderly. *N Engl J Med*. 2018;379(16):1509-1518. doi:10.1056/NEJMoa1805819

16. McNeil JJ, Nelson MR, Woods RL, et al. Effect of Aspirin on All-Cause Mortality in the Healthy Elderly. *N Engl J Med*. 2018;379(16):1519-1528. doi:10.1056/NEJMoa1803955

17. Cobain MR, Pencina MJ, D'Agostino RB, Vasan RS. Lifetime risk for developing dyslipidemia: the Framingham Offspring Study. *Am J Med*. 2007;120(7):623-630. doi:10.1016/j.amjmed.2006.12.015

18. Tindle HA, Stevenson Duncan M, Greevy RA, et al. Lifetime Smoking History and Risk of Lung Cancer: Results From the Framingham Heart Study. *J Natl Cancer Inst*. 2018;110(11). doi:10.1093/jnci/djy041

19. Dziak JJ, Henry KL. Two-Part Predictors in Regression Models. *Multivariate Behav Res*. 2017;52(5):551-561. doi:10.1080/00273171.2017.1333404

20. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: A scientific statement from the American heart association. *Circulation*. 2009;119(17):2408-2416. doi:10.1161/CIRCULATIONAHA.109.192278

21. Harrell FE. The PHGLM Procedure. In: *SUGI Supplemental Library Guide*. 5th ed. Cary, NC; 1986.

22. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2

23. Pencina MJ, D'Agostino RB, Vasan RS. Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med*. 2010;48(12):1703-1711. doi:10.1515/CCLM.2010.340

24. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157-172. doi:10.1002/sim.2929

25. Pencina MJ, D'Agostino RB, Pencina KM, Janssens ACJW, Greenland P. Interpreting Incremental Value of Markers Added to Risk Prediction Models. *Am J Epidemiol*. 2012;176(6):473-481.

doi:10.1093/aje/kws207

26. D'Agostino RB, Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handb Stat*. 2004;23:1-25.

27. Maddala G. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press; 1983.

28. Nagelkerke NJD. *A Note on a General Definition of the Coefficient of Determination*. Vol 78.; 1991.

29. Cook NR. Clinically Relevant Measures of Fit? A Note of Caution. *Am J Epidemiol*. 2012;176(6):488-491. doi:10.1093/aje/kws208

30. ASCVD Risk Estimator +. http://tools.acc.org/ASCVD-Risk-Estimator-Plus/#!/calculate/estimate/. Accessed March 10, 2020.

31. U.S. Department of Health and Human Services. *Smoking Cessation: A Report of the Surgeon General*. Rockville, MD; 2020. https://www.hhs.gov/sites/default/files/2020-cessation-sgr-full-report.pdf. Accessed June 9, 2020.

32. Centers for Medicare and Medicaid Services. *Decision Memo for Screening for Lung Cancer with Low Dose Computed Tomography (LDCT)*. Baltimore, MD; 2015. https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=274.

33. USDHHS. The Health Consequences of Smoking—50 Years of Progress A Report of the Surgeon General. *A Rep Surg Gen*. 2014:1081. doi:NBK179276

34. Xu X, Bishop EE, Kennedy SM, Simpson SA, Pechacek TF. Annual healthcare spending attributable to cigarette smoking: An update. *Am J Prev Med*. 2015;48(3):326-333. doi:10.1016/j.amepre.2014.10.012

35. Lightwood JM, Glantz SA. Short-term Economic and Health Benefits of Smoking Cessation. *Circulation*. 1997;96(4):1089-1096. doi:10.1161/01.CIR.96.4.1089

36. Lightwood J, Glantz SA. Smoking Behavior and Healthcare Expenditure in the United States, 1992–2009: Panel Data Estimates. Hall WD, ed. *PLOS Med*. 2016;13(5):e1002020. doi:10.1371/journal.pmed.1002020

37. Fishman P, Thompson E, Merikle E, Curry S. Changes in health care costs before and after smoking cessation. *Nicotine Tob Res*. 2006;8(3):393-401. doi:10.1080/14622200600670314

38. Wang Y, Chen ES, Pakhomov S, Lindemann E, Melton GB. Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record. *AMIA . Annu Symp proceedings AMIA Symp*. 2016;2016:1209-1218.

39. Milinovich A, Kattan MW. Extracting and utilizing electronic health data from Epic for research. *Ann Transl Med*. 2018;6(3):42-42. doi:10.21037/atm.2018.01.13

40. Amin NP, Martin SS, Blaha MJ, Nasir K, Blumenthal RS, Michos ED. Headed in the right direction but at risk for miscalculation: A critical appraisal of the 2013 ACC/AHA risk assessment guidelines. *J Am Coll Cardiol*. 2014;63(25 PART A):2789-2794. doi:10.1016/j.jacc.2014.04.010

41. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: A one-shot nonparametric approach. *Stat Med*. 2015;34(4):685-703. doi:10.1002/sim.6370

42. Laaksonen M, Rahkonen O, Karvonen S, Lahelma E. Socioeconomic status and smoking Analysing inequalities with multiple indicators. doi:10.1093/eurpub/cki115

43. Garrett BE, Martell BN, Caraballo RS, King BA. Socioeconomic differences in cigarette smoking among sociodemographic groups. *Prev Chronic Dis*. 2019;16(6). doi:10.5888/pcd16.180553

44. Hiscock R, Bauld L, Amos A, Fidler JA, Munafò M, Munafò M. Socioeconomic status and smoking: a review. *Ann NY Acad Sci*. doi:10.1111/j.1749-6632.2011.06202.x

45. Colantonio LD, Richman JS, Carson AP, et al. Performance of the Atherosclerotic Cardiovascular Disease Pooled Cohort Risk Equations by Social Deprivation Status. *J Am Heart Assoc*. 2017;6(3). doi:10.1161/JAHA.117.005676

46. Dalton JE, Perzynski AT, Zidar DA, et al. Accuracy of cardiovascular risk prediction varies by neighborhood socioeconomic position a retrospective cohort study. *Ann Intern Med*. 2017;167(7):456-464. doi:10.7326/M16-2543

47. USDHHS. *The Health Consequences of Smoking: A Report of the Surgeon General*. Atlanta, Georgia; 2004. https://www.cdc.gov/tobacco/data_statistics/sgr/2004/complete_report/index.htm.

48. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ*. 2000;321(7257):323-329. doi:10.1136/bmj.321.7257.323

49. Moyer VA. Screening for Lung Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med*. 2014;160(5):330-338. doi:10.7326/M13-2771

50. United States Preventive Services Taskforce. https://uspreventiveservicestaskforce.org/uspstf/draft-recommendation/lung-cancer-screening-2020#fullrecommendationstart. Accessed July 14, 2020.

51. Pinsky PF, Berg CD. Applying the National Lung Screening Trial eligibility criteria to the US population: what percent of the population and of incident lung cancers would be covered? *J Med Screen*. 2012;19(3):154-156. doi:10.1258/jms.2012.012010

52. Centers for Medicare and Medicaid Services. Physician Fee Schedule Search. https://www.cms.gov/apps/physician-fee-schedule/license-agreement.aspx. Accessed July 2, 2020.

53. Kramer BS, Berg CD, Aberle DR, Prorok PC. Lung cancer screening with low-dose helical CT: results from the National Lung Screening Trial (NLST). *J Med Screen*. 2011;18(3):109-111. doi:10.1258/jms.2011.011055

54. Bach PB, Mirkin JN, Oliver TK, et al. Benefits and Harms of CT Screening for Lung Cancer. *JAMA*. 2012;307(22):2418. doi:10.1001/jama.2012.5521

55. Humphrey LL, Deffebach M, Pappas M, et al. Screening for Lung Cancer With Low-Dose Computed Tomography: A Systematic Review to Update the U.S. Preventive Services Task Force Recommendation. *Ann Intern Med*. 2013;159(6):411. doi:10.7326/0003-4819-159-6-201309170-00690

56. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet*. 2017;49(7):1126-1132. doi:10.1038/ng.3892

57. Bossé Y, Amos CI. A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol Biomarkers Prev*. 2018;27(4):363-379. doi:10.1158/1055-9965.EPI-16-0794

58. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*. 2008;452(7187):638-642. doi:10.1038/nature06846

59. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008;452(7187):633-637.

doi:10.1038/nature06885

60. Wang Y, McKay JD, Rafnar T, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet*. 2014;46(7):736-741. doi:10.1038/ng.3002

61. Timofeeva MN, Hung RJ, Rafnar T, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet*. 2012;21(22):4980-4995. doi:10.1093/hmg/dds334

62. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet*. 2009;85(5):679-691. doi:10.1016/j.ajhg.2009.09.012

63. Broderick P, Wang Y, Vijayakrishnan J, et al. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res*. 2009;69(16):6633-6641. doi:10.1158/0008-5472.CAN-09-0680

64. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*. 2008;40(12):1407-1409. doi:10.1038/ng.273

65. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*. 2008;40(12):1404-1406. doi:10.1038/ng.254

66. Liu P, Vikis HG, Wang D, et al. Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *J Natl Cancer Inst*. 2008;100(18):1326-1330. doi:10.1093/jnci/djn268

67. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*. 2008;40(5):616-622. doi:10.1038/ng.109

68. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*. 2008;452(7187):638-642. doi:10.1038/nature06846

69. Bierut LJ, Stitzel JA, Wang JC, et al. Variants in nicotinic receptors and risk for nicotine dependence. *Am J Psychiatry*. 2008;165(9):1163-1171. doi:10.1176/appi.ajp.2008.07111711

70. Saccone SF, Hinrichs AL, Saccone NL, et al. Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet*. 2007;16(1):36-49. doi:10.1093/hmg/ddl438

71. Mackillop J, Obasi E, Amlung MT, McGeary JE, Knopik VS. The Role of Genetics in Nicotine Dependence: Mapping the Pathways from Genome to Syndrome. *Curr Cardiovasc Risk Rep*. 2010;4(6):446-453. doi:10.1007/s12170-010-0132-6

72. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. 2019;51(2):237-244. doi:10.1038/s41588-018-0307-5

73. Saccone NL, Culverhouse RC, Schwantes-An T-H, et al. Multiple Independent Loci at Chromosome 15q25.1 Affect Smoking Quantity: a Meta-Analysis and Comparison with Lung Cancer and COPD. Gibson G, ed. *PLoS Genet*. 2010;6(8):e1001053. doi:10.1371/journal.pgen.1001053

74. Chiang H chih, Wang CY, Lee HL, Tsou TC. Metabolic effects of CYP2A6 and CYP2A13 on 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK)-induced gene mutation-A mammalian cell-based mutagenesis approach. *Toxicol Appl Pharmacol*. 2011;253(2):145-152. doi:10.1016/j.taap.2011.03.022

75. AM A, P R. The Multifarious Link Between Cytochrome P450s and Cancer. *Oxid Med Cell Longev*.

2020;2020. doi:10.1155/2020/3028387

76. Dawber TR, Kannel WB, Lyell LP. AN APPROACH TO LONGITUDINAL STUDIES IN A COMMUNITY: THE FRAMINGHAM STUDY. *Ann N Y Acad Sci*. 2006;107(2):539-556. doi:10.1111/j.1749-6632.1963.tb13299.x

77. Kreger BE, Splansky GL, Schatzkin A. The cancer experience in the framingham heart study cohort. *Cancer*. 1991;67(1):1-6. doi:10.1002/1097-0142(19910101)67:1<1::AID-CNCR2820670102>3.0.CO;2-W

78. Tindle HA, Stevenson Duncan M, Greevy RA, et al. Lifetime Smoking History and Risk of Lung Cancer: Results From the Framingham Heart Study. *J Natl Cancer Inst*. 2018;110(11):1201-1207. doi:10.1093/jnci/djy041

79. Psaty BM, O'Donnell CJ, Gudnason V, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. *Circ Cardiovasc Genet*. 2009;2(1):73-80. doi:10.1161/CIRCGENETICS.108.829747

80. Turner S, Armstrong LL, Bradford Y, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet*. 2011;Chapter 1:Unit1.19. doi:10.1002/0471142905.hg0119s68

81. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27(2):e1608. doi:10.1002/mpr.1608

82. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-909. doi:10.1038/ng1847

83. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2(12):2074-2093. doi:10.1371/journal.pgen.0020190

84. Amos CI, Dennis J, Wang Z, et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev*. 2017;26(1):126-135. doi:10.1158/1055-9965.EPI-16-0106

85. Amos CI, Dennis J, Wang Z, et al. The oncoarray consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev*. 2017;26(1):126-135. doi:10.1158/1055-9965.EPI-16-0106

86. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics*. 2015;31(9):1466-1468. doi:10.1093/bioinformatics/btu848

87. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. Wray NR, ed. *PLoS Genet*. 2013;9(3):e1003348. doi:10.1371/journal.pgen.1003348

88. Aldrich MC, Kumar R, Colangelo LA, et al. Genetic ancestry-smoking interactions and lung function in African Americans: a cohort study. *PLoS One*. 2012;7(6):e39541. doi:10.1371/journal.pone.0039541

89. Stern MC, Fejerman L, Das R, et al. Variability in Cancer Risk and Outcomes Within US Latinos by National Origin and Genetic Ancestry. *Curr Epidemiol Reports*. 2016;3(3):181-190. doi:10.1007/s40471-016-0083-7

90. Wang TW, Asman K, Gentzke AS, et al. Tobacco Product Use Among Adults — United States, 2017. *MMWR Morb Mortal Wkly Rep*. 2018;67(44):1225-1232. doi:10.15585/mmwr.mm6744a2

91. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. *CA Cancer J Clin*. 2014;64(1):9-29. doi:10.3322/caac.21208

92. HARRIS RE, ZANG EA, ANDERSON JI, WYNDER EL. Race and Sex Differences in Lung Cancer Risk Associated with Cigarette Smoking. *Int J Epidemiol*. 1993;22(4):592-599. doi:10.1093/ije/22.4.592

93. Kanwal M, Ding X-J, Cao Y. Familial risk for lung cancer. *Oncol Lett*. 2017;13(2):535-542. doi:10.3892/ol.2016.5518

94. DEVESA SS, DIAMOND EL. SOCIOECONOMIC AND RACIAL DIFFERENCES IN LUNG CANCER INCIDENCE. *Am J Epidemiol*. 1983;118(6):818-831. doi:10.1093/oxfordjournals.aje.a113700

95. Noone A, Howlader N, Krapcho M, et al. *SEER Cancer Statistics Review, 1975-2015*. Bethesda, MD https://seer.cancer.gov/archive/csr/1975_2015/#contents. Accessed July 2, 2020.

96. Little RJA. Missing-Data Adjustments in Large Surveys. *J Bus Econ Stat*. 1988;6(3):287-296. doi:10.2307/1391878

97. Schafer J. *Analysis of Incomplete Multivariate Data*. 1st ed. Chapman and Hall; 1997. doi:10.1201/9781439821862

98. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons; 1987.

99. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol*. 2013;177(4):292-298. doi:10.1093/aje/kws412

100. VanderWeele TJ, Asomaning K, Tchetgen EJT, et al. Genetic Variants on 15q25.1, Smoking, and Lung Cancer: An Assessment of Mediation and Interaction. *Am J Epidemiol*. 2012;175(10):1013-1020. doi:10.1093/aje/kwr467

101. Qian DC, Han Y, Byun J, et al. A novel pathway-based approach improves lung cancer risk prediction using germline genetic variations. *Cancer Epidemiol Biomarkers Prev*. 2016;25(8):1208-1215. doi:10.1158/1055-9965.EPI-15-1318

102. Jia G, Lu Y, Wen W, et al. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectr*. 2020;4(3). doi:10.1093/jncics/pkaa021

103. Dai J, Lv J, Zhu M, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. *Lancet Respir Med*. 2019;7(10):881-891. doi:10.1016/S2213-2600(19)30144-4

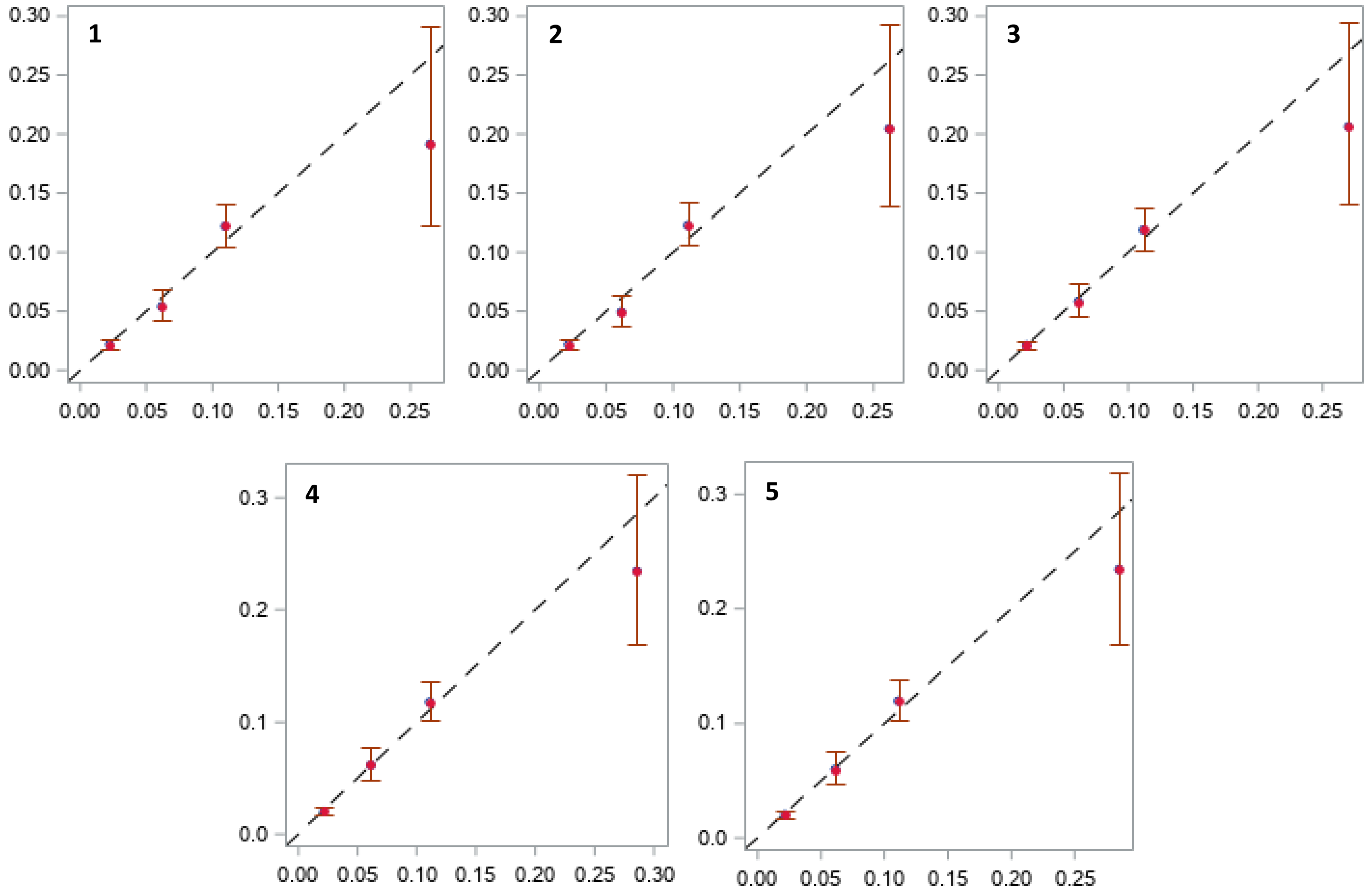
104. Khera A V., Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224. doi:10.1038/s41588-018-0183-z

105. Khera A V., Chaffin M, Zekavat SM, et al. Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation*. 2019;139(13):1593-1602. doi:10.1161/CIRCULATIONAHA.118.035658

106. Claussnitzer M, Cho JH, Collins R, et al. A brief history of human disease genetics. *Nature*. 2020;577(7789):179-189. doi:10.1038/s41586-019-1879-7

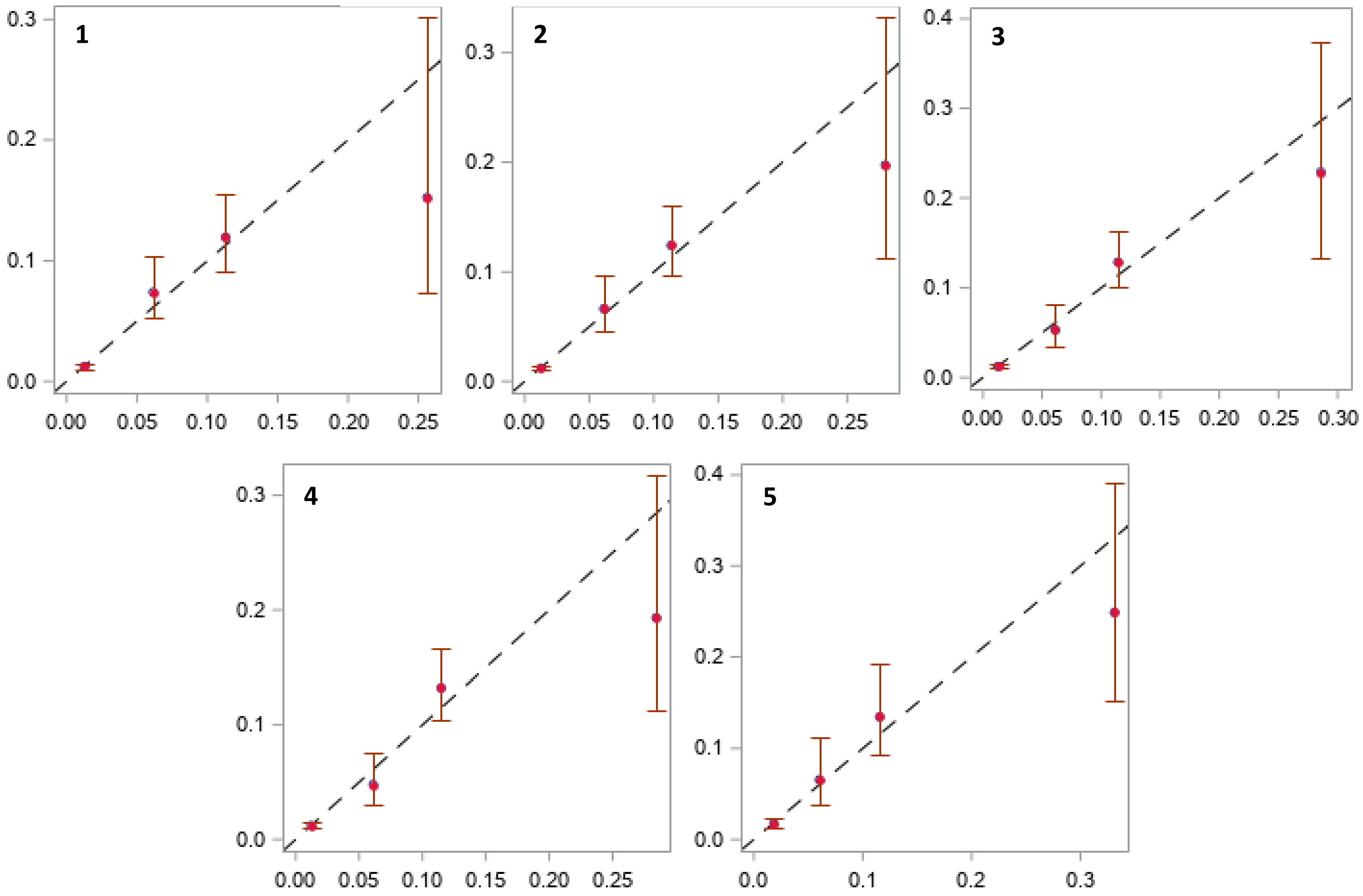
107. Wassenaar CA, Dong Q, Wei Q, Amos CI, Spitz MR, Tyndale RF. Relationship Between CYP2A6 and CHRNA5-CHRNA3-CHRNB4 Variation and Smoking Behaviors and Lung Cancer Risk. *JNCI J Natl Cancer Inst*. 2011;103(17):1342-1346. doi:10.1093/jnci/djr237

Supplemental Figure 1: Calibration Plots in Men



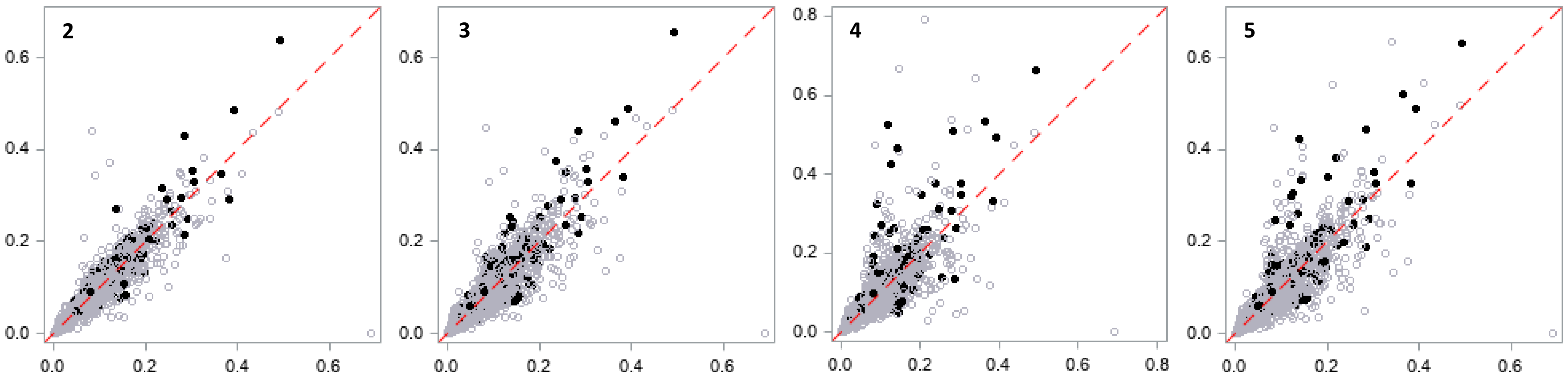
Observed Probability (y-axis) vs. Predicted Probability (x-axis)

Supplemental Figure 2: Calibration Plots in Women



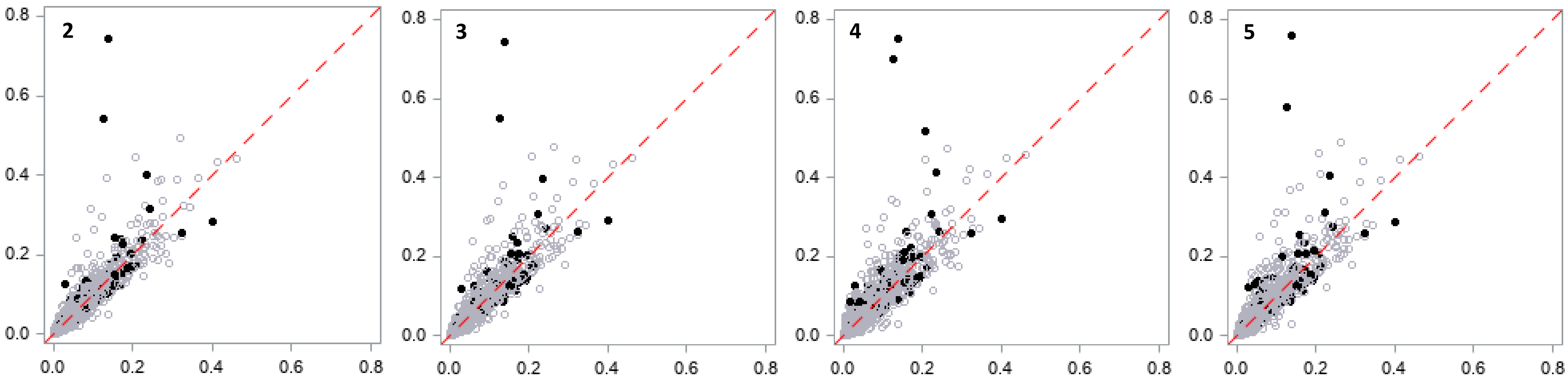
Observed Probability (y-axis) vs. Predicted Probability (x-axis)

Supplemental Figure 3 Predicted Probability of ASCVD in Men: Models 2-5 (y-axis) vs Model 1 (x-axis)



Filled circles indicate individuals who had an ASCVD event, open circles are nonevents

Supplemental Figure 4: Predicted Probability of ASCVD in Women: Models 2-5 (y-axis) vs Model 1 (x-axis)



Filled circles indicate individuals who had an ASCVD event, open circles are nonevents

Supplemental Table 1: Reclassification of Risk under Models 1 and 5 in Men

Model 5

		Non-Events					Events			
		<5%	*5-7.49%*	*7.5-19.9%*	*≥20%*		*<5%*	*5-7.49%*	*7.5-19.9%*	*≥20%*
Model 1	*<5%*	5245	178	4	0	*<5%*	105	10	3	0
	5-7.49%	320	713	238	2	*5-7.49%*	6	39	22	0
	7.5-19.9%	30	181	958	63	*7.5-19.9%*	2	14	124	15
	≥20%	4	3	32	58	*≥20%*	0	0	2	16

Totals:	570	485	6974	50	24	284

Predicted risk under Model 1 in rows; predicted risk under Model 5 in columns

Green fill indicates a movement in the correct direction (lower risk for non-events, higher risk for events); Red fill indicates a move in the incorrect direction (higher risk for non-events, lower risk for events); Gray fill indicates no reclassification

Supplemental Table 2: Reclassification of Risk under Models 1 and 5 in Women

Model 5

		Non-Events					Events			
		<5%	*5-7.49%*	*7.5-19.9%*	*≥20%*		*<5%*	*5-7.49%*	*7.5-19.9%*	*≥20%*
	<5%	8708	146	31	0	*<5%*	97	6	5	0
Model 1	*5-7.49%*	167	174	82	2	*5-7.49%*	7	11	13	0
	7.5-19.9%	9	86	312	25	*7.5-19.9%*	0	4	40	7
	≥20%	0	0	19	39	*≥20%*	0	0	1	6

Totals:	281	286	9233

Totals:	31	12	154

Predicted risk under Model 1 in rows; predicted risk under Model 5 in columns

Green fill indicates a movement in the correct direction (lower risk for non-events, higher risk for events); Red fill indicates a move in the incorrect direction (higher risk for non-events, lower risk for events); Gray fill indicates no reclassification

Supplemental Table 3: Reclassification of Risk under Models 1 and 5 in Heavy Ever Smoking (>20 Pack-Years) Men

Model 5

		Non-Events					Events			
		<5%	*5-7.49%*	*7.5-19.9%*	*≥20%*		*<5%*	*5-7.49%*	*7.5-19.9%*	*≥20%*
Model 1	*<5%*	1557	68	9	0	*<5%*	39	5	0	0
	5-7.49%	258	404	119	1	*5-7.49%*	5	19	10	0
	7.5-19.9%	25	150	705	55	*7.5-19.9%*	2	11	99	10
	≥20%	5	4	25	54	*≥20%*	0	0	2	19

Totals:	467	252	2720	Totals:	25	20	176

Predicted risk under Model 1 in rows; predicted risk under Model 5 in columns

Green fill indicates a movement in the correct direction (lower risk for non-events, higher risk for events); Red fill indicates a move in the incorrect direction (higher risk for non-events, lower risk for events); Gray fill indicates no reclassification

Supplemental Table 4: Reclassification of Risk under Models 1 and 5 in Heavy Ever Smoking (>20 Pack-Years) Women

Model 5

Non-Events

		<5%	*5-7.49%*	*7.5-19.9%*	*≥20%*
Model 1	*<5%*	2094	57	22	0
	5-7.49%	97	70	32	0
	7.5-19.9%	28	59	130	22
	≥20%	0	1	16	24

Totals:	201	133	2318

Events

	<5%	*5-7.49%*	*7.5-19.9%*	*≥20%*
<5%	32	4	3	0
5-7.49%	5	4	4	0
7.5-19.9%	1	4	18	5
≥20%	0	0	1	8

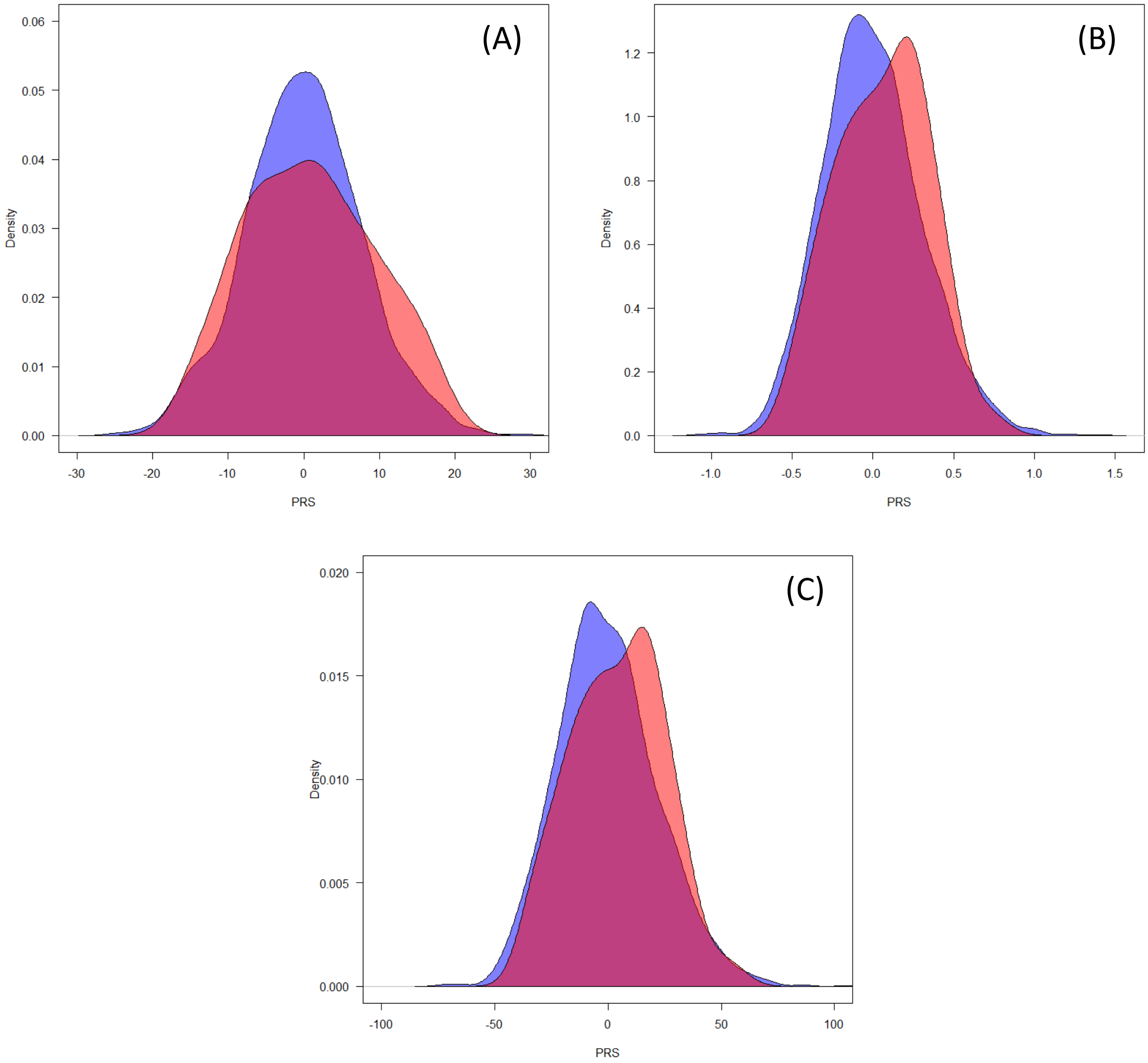
Totals:	16	11	62

Predicted risk under Model 1 in rows; predicted risk under Model 5 in columns

Green fill indicates a movement in the correct direction (lower risk for non-events, higher risk for events); Red fill indicates a move in the incorrect direction (higher risk for non-events, lower risk for events); Gray fill indicates no reclassification

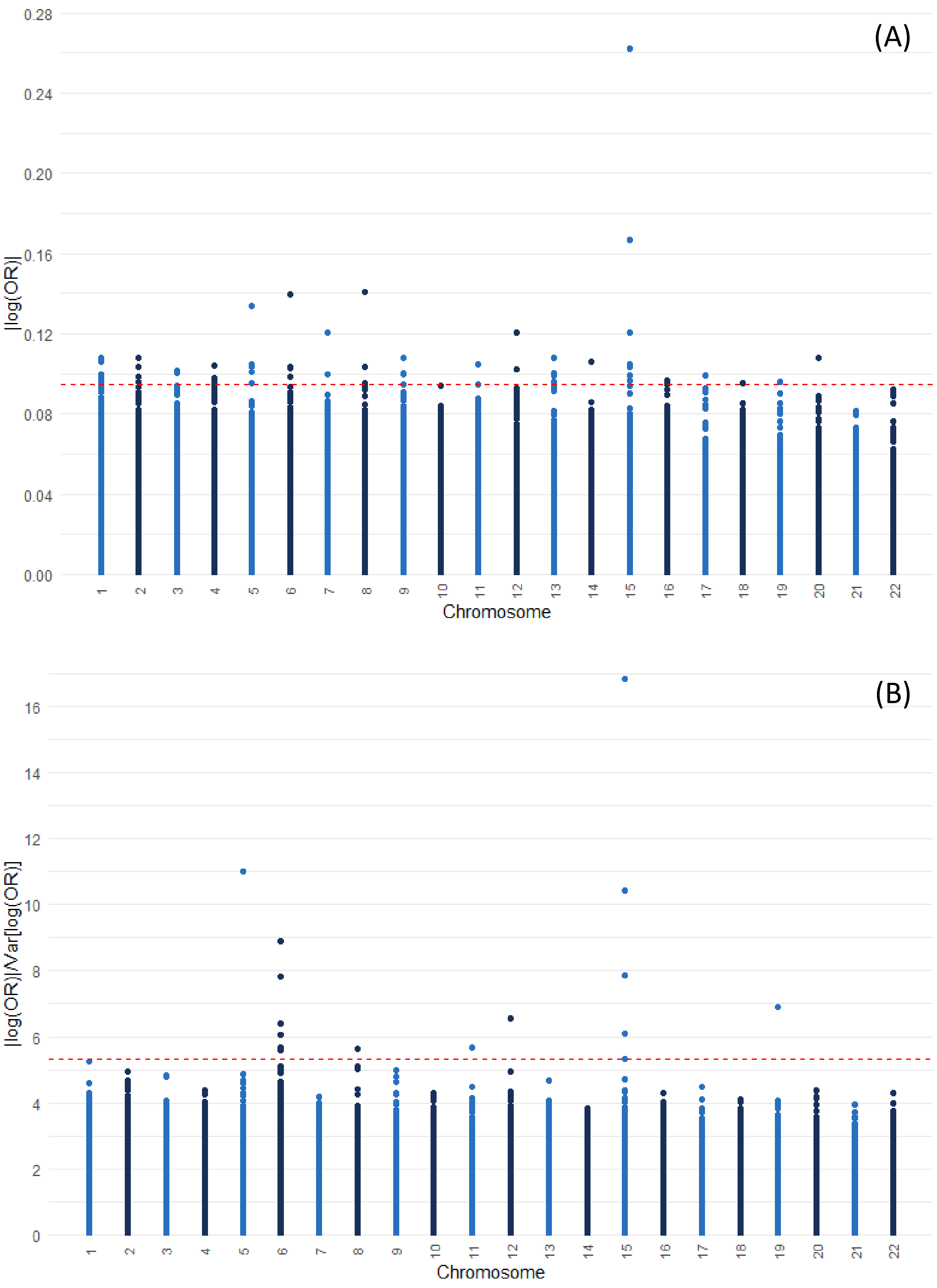
APPENDIX B. SUPPLEMENTAL TABLES AND FIGURES FOR CHAPTER 2

Supplemental Figure 5: Distribution plots of genome-wide polygenic risk scores by event status



Blue shading displays the distribution of the PRS among controls; red is among cases. Clockwise, panels display PRS_{unw} (panel A), PRS_{β} (panel B), and $\text{PRS}_{\beta/\text{Var}(\beta)}$ (panel C)

Supplemental Figure 6: Manhattan Plots for Genome-Wide Polygenic Risk Scores



Panel A shows the weight given to each SNP included in PRS$_{\beta}$ with a reference corresponding to an odds ratio of 1.1; Panel B shows the weight assigned to each SNP contributing to PRS$_{\beta/\text{Var}(\beta)}$ with a reference line corresponding to a genome-wide significance level equal to 5×10^{-8}

Supplemental Table 5: Effect of Pack-Years Adjusting for Genome-Wide PRS

PRS Value	Hazard Ratio [95% CI] per 10 Pack-Years	p-value
PRS_{unw}	1.46 [1.35, 1.58]	<0.0001
PRS_{β}	1.46 [1.35, 1.58]	<0.0001
$PRS_{\beta/Var(\beta)}$	1.46 [1.35, 1.58]	<0.0001

Utilizes all 83,304 SNPs that remained following LD pruning in construction of the PRS. Hazard Ratios are estimated from mixed-effects Cox proportional hazards regression models adjusting for pack-years, PRS, age, sex, current smoking status, education, and population structure. The variance structure accounts for familial relationships.

Supplemental Table 6: Effect of Genome-Wide Polygenic Risk Score on Lung Cancer Risk

Polygenic Risk Score	PRS Standard Deviation	Hazard Ratio [95% CI] per SD increase in PRS	p-value
PRS_{unw}	115.46	1.09 [0.79, 1.51]	0.59
PRS_{β}	2.74	1.05 [0.84, 1.31]	0.65
$PRS_{\beta/Var(\beta)}$	165.28	1.08 [0.86, 1.34]	0.52

Estimates are from a mixed-effects Cox proportional hazards regression model adjusted for population structure; the variance structure accounts for familial relationships. Hazard Ratios and corresponding 95% confidence intervals are estimated per standard deviation increase in polygenic risk score.

Supplemental Table 7: Association between PRS and Fulfillment of USPSTF Screening Criteria

PRS Value	Odds Ratio [95% CI]	p-value
PRS_{unw}	2.18 [1.36, 3.49]	0.001
PRS_{β}	0.91 [0.58, 1.41]	0.67
$PRS_{\beta/Var(\beta)}$	0.89 [0.56, 1.40]	0.60

Estimates are from a logistic regression model with a binary indicator for “eligible for lung cancer screening based on USPSTF guidelines under consideration” as the outcome. PRS and population structure were included as independent variables. Hazard Ratio is per SD increase in the PRS. PRS were built from all 83,304 SNPs remaining after LD pruning