Rare and low-frequency exonic variants and gene-by-smoking interactions in pulmonary function

Yang, T.Z.; Jackson, V.E.; Smith, A.V.; Chen, H.; Bartz, T.M.; Sitlani, C.M.; ... ; Morrison, A.C.

scientific reports

Check for updates

OPEN

Rare and low-frequency exonic variants and gene-by-smoking interactions in pulmonary function

Tianzhong Yang[1,2], Victoria E. Jackson[3], Albert V. Smith[4], Han Chen[5,6], Traci M. Bartz[7,8], Colleen M. Sitlani[8], Bruce M. Psaty[9,10], Sina A. Gharib[11], George T. O'Connor[12,13], Josée Dupuis[14], Jiayi Xu[15,16], Kurt Lohman[17], Yongmei Liu[17], Stephen B. Kritchevsky[18], Patricia A. Cassano[16,19], Claudia Flexeder[20], Christian Gieger[21], Stefan Karrasch[20,22,23], Annette Peters[20,24], Holger Schulz[20,23], Sarah E. Harris[25,26], John M. Starr[26,27], Ian J. Deary[25,26], Ani Manichaikul[28,29], Elizabeth C. Oelsner[30,31], R. G. Barr[31,32], Kent D. Taylor[33], Stephen S. Rich[34], Tobias N. Bonten[35], Dennis O. Mook-Kanamori[36,37], Raymond Noordam[38], Ruifang Li-Gao[36], Marjo-Riitta Jarvelin[39,40,41], Matthias Wielscher[39], Natalie Terzikhan[42,43], Lies Lahousse[42,43], Guy Brusselle[42,43], Stefan Weiss[44,45], Ralf Ewert[46], Sven Gläser[46,47], Georg Homuth[44], Nick Shrine[3], Ian P. Hall[48], Martin Tobin[3,49], Stephanie J. London[50]✉, Peng Wei[51]✉ & Alanna C. Morrison[5]✉

Genome-wide association studies have identified numerous common genetic variants associated with spirometric measures of pulmonary function, including forced expiratory volume in one second (FEV_1), forced vital capacity, and their ratio. However, variants with lower minor allele frequencies are less explored. We conducted a large-scale gene-smoking interaction meta-analysis on exonic rare and low-frequency variants involving 44,429 individuals of European ancestry in the discovery stage and sought replication in the UK BiLEVE study with 45,133 European ancestry samples and UK Biobank study with 59,478 samples. We leveraged data on cigarette smoking, the major environmental risk factor for reduced lung function, by testing gene-by-smoking interaction effects only and simultaneously testing the genetic main effects and interaction effects. The most statistically significant signal that replicated was a previously reported low-frequency signal in *GPR126*, distinct from common variant associations in this gene. Although only nominal replication was obtained for a top rare variant signal rs142935352 in one of the two studies, interaction and joint tests for current smoking and *PDE3B* were significantly associated with FEV_1. This study investigates the utility of assessing gene-by-smoking interactions and underscores their effects on potential pulmonary function.

[1]Department of Biostatistics and Data Science, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [3]Department of Health Sciences, University of Leicester, Leicester, UK. [4]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [5]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [6]Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA. [7]Department of Biostatistics, University of Washington, Seattle, USA. [8]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [9]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology and Health Services, University of Washington, Seattle, WA, USA. [10]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. [11]Division of Pulmonary, Critical Care, and Sleep, Department of Medicine, University of Washington, Seattle, WA, USA. [12]Department of Medicine, Pulmonary Center, Boston University School of Medicine, Boston, MA, USA. [13]National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA. [14]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [15]Pamela Sklar Division of Psychiatric Genomics, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [16]Division of Nutritional Sciences, Cornell University, Ithaca, NY, USA. [17]Division of Cardiology, Department of Medicine, Duke Molecular Physiology Institute, Duke University School of Medicine, Durham, NC, USA. [18]Sticht Center for Healthy Aging and

Alzheimer's Prevention, Wake Forest School of Medicine, Winston-Salem, NC, USA. [19]Department of Healthcare Policy and Research, Weill Cornell Medical College, New York, NY, USA. [20]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [21]Research Unit of Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [22]Institute and Outpatient Clinic for Occupational, Social and Environmental Medicine, Ludwig-Maximilians-Universität, Munich, Germany. [23]Comprehensive Pneumology Center Munich (CPC-M), German Center for Lung Research (DZL), Munich, Germany. [24]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Munich, Germany. [25]Department of Psychology, The University of Edinburgh, Edinburgh, UK. [26]Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK. [27]Alzheimer Scotland Dementia Research Centre, The University of Edinburgh, Edinburgh, UK. [28]Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [29]Division of Biostatistics and Epidemiology, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA. [30]Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY, USA. [31]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA. [32]Department of Medicine, Department of Epidemiology, Columbia University Medical Center, New York, NY, USA. [33]David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [34]Department of Public Health Sciences, Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [35]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands. [36]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. [37]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands. [38]Division of Gerontology and Geriatrics, Department of Internal Medicine, Leiden University Medical Center, Leiden, The Netherlands. [39]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [40]Faculty of Medicine, Center for Life Course Health Research, University of Oulu, Oulu, Finland. [41]Biocenter of Oulu, University of Oulu, Oulu, Finland. [42]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands. [43]Department of Respiratory Medicine, Ghent University Hospital, Ghent, Belgium. [44]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. [45]DZHK (German Centre for Cardiovascular Research), Partner Site Greifswald, Greifswald, Germany. [46]Division of Cardiology, Pneumology, Infectious Diseases, Intensive Care Medicine, Department of Internal Medicine B, University Medicine Greifswald, Greifswald, Germany. [47]Department of Internal Medicine, Vivantes Hospital Berlin Spandau, Berlin, Germany. [48]Division of Respiratory Medicine, University of Nottingham, Nottingham, UK. [49]Leicester Respiratory Biomedical Research Unit, National Institute for Health Research, Glenfield Hospital, Leicester, UK. [50]Department of Health and Human Services, Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA. [51]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ✉email: london2@niehs.nih.gov; Pwei2@mdanderson.org; Alanna.C.Morrison@uth.tmc.edu

Measures of pulmonary function provide important clinical and research tools for evaluating lung disease and other morbidities[1]. Pulmonary function is used to diagnose chronic obstructive pulmonary disease and to monitor the severity and progression of many lung conditions. Furthermore, even within the normal range, pulmonary function is a risk factor for mortality and morbidity from various conditions[2–5]. Genome-wide association studies (GWASs) have identified many single nucleotide variants (SNVs) for pulmonary function after adjusting for cigarette smoking, a well-established risk factor for reduced pulmonary function, as a confounder and potential effect modifier[6–14]. The interplay between genetic and environmental factors likely plays a role in complex traits, including chronic lung diseases. A large GWAS meta-analysis has shown the existence of SNV-by-smoking interaction in relation to two measures of pulmonary function (forced expiratory volume in one second, FEV_1, and the ratio of FEV_1 to the forced vital capacity, FEV_1/FVC) by performing joint analyses of genetic main effects and interaction effects of single SNVs and their interaction with smoking[14]. However, there is a lack of gene-by-environment (GxE) interaction analysis involving low-frequency (minor allele frequency [MAF] 1% to 5%) or rare variants (MAF less than 1%)[15], which are typically not well captured or imputed by previous GWAS meta-analyses of pulmonary function. The more recent use of genotyping platforms including lower frequency and rare variants enables better assessment of the role of rare genetic variation in disease.

Due to the low power to detect association with rare variants, a commonly adopted strategy in analyzing rare variant-phenotype associations is to group variants according to genes or genomic regions and test whether the grouped variants are associated with the phenotype[16]. While single-variant-based interaction tests for common variants are well-established[17,18], methods for detecting rare variant GxE interactions are relatively new. Recently developed novel approaches for testing rare variant GxE interaction effects include a joint test that allows for simultaneous testing of the genetic main and interaction effects as well as the ability to assess gene-based GxE interactions only for both related and unrelated individuals[19–22]. The joint test and the test for the interaction terms only are both important for different scientific hypotheses. The joint test assesses genetic associations with pulmonary function, accounting for heterogeneity of genetic effects in individuals with different environmental exposures (in our context, smoking behaviors). The test of the interaction terms only can identify genes that modify the effect of smoking on pulmonary function. For single variant analysis, an interaction requires at least four fold larger sample size than a genetic main effect of comparable magnitude[23]. Similarly, for gene-based interaction tests, larger sample sizes are required for detection, especially for rare variants. This study is the first to incorporate GxE interactions in modeling rare and low-frequency genetic variants and smoking effects on

Study	Sample size	Male (%)	Age (sd)	Ever smokers (%)	Current smoker (%)	FEV_1 (sd)	FVC (sd)	Ratio (sd)
Discovery stage								
ARIC	10,874	46.8	54.3 (5.7)	40.3	24.3	2935.4 (768.1)	3977.2 (978.0)	73.8 (7.6)
FHS	6172	46.6	44.9 (10.8)	47.8	19.1	3325.6 (844)	4321.9 (1043.2)	77.0 (7.0)
1958BC	4594	55.7	42 (0)	53.4	24.9	3350.0 (782.74)	4275.0 (1024.6)	78.9 (8.4)
RS	567	54.7	79.9 (5.0)	30.0	9.9	2265.68 (679.79)	3025.1 (860.3)	75.0 (8.0)
SHIP	4694	49.3	49.9 (14.4)	39.3	24.6	3260.9 (905.6)	4021.7 (1087.5)	81.2 (6.9)
NFBC1966	1431	45.6	31 (0)	53.0	25.6	3925.1 (778.1)	4685.2 (989.1)	84.2 (6.1)
NEO	6203	47.5	55.7 (6.0)	35.6	16.2	3239.8 (795.4)	4222.8 (1025.1)	77.0 (6.7)
CHS	3496	43.3	72.6 (5.5)	43.5	10.1	2114.8 (653.3)	3026.2 (842.0)	69.8 (10.1)
AGES	1459	40.9	76.3 (8.8)	45.1	11.9	2146.1 (676.0)	2883.8 (845.1)	74.1 (8.8)
LBC1936	970	50.4	69.5 (0.8)	47.2	10.7	2370.7 (676.1)	3040.6 (861.8)	78.5 (10.0)
KORA	1217	46.8	51.6 (5.7)	60.5	24.2	3341.7 (822.0)	4307.8 (1014.2)	77.6(6.2)
MESA	1298	49.4	66.0 (9.8)	43.6	7.6	2565.5 (763.59)	3510.9 (993.34)	73.3 (8.7)
HABC	1454	53.2	73.7 (2.8)	56.5	6.5	2313.7 (653.5)	3114.5 (812.4)	73.7 (2.8)
Replication stage								
UK BiLEVE	45,133	49.1	57.0 (7.9)	49.3	18.2	3565.0 (851.7)	3573.0 (1040.1)	74.0 (7.3)
UK Biobank	59,478	43.2	56.0 (8.0)	45.0	8.5	2874.0 (725)	3754 (932.1)	76.7 (5.6)

Table 1. Descriptive statistics for each participating cohort. FEV_1 and FVC are in unit of mL.

Gene	Chr	Smoking status and phenotype Combination	Average # variants across studies (range)	Discovery Stage		Replication Stage (UK BiLEVE)		Replication Stage (UK Biobank)		GTEx
				P-value (interaction)	*P*-value (joint)	*P*-value (interaction)	*P*-value (joint)	*P*-value (interaction)	*P*-value (joint)	Expressed in lung
GPR126	6	Current & Ratio	18.8 (5–25)	0.40	1.9E−09	0.61	1.1E−16	0.54	1.4E−30	Yes
		Ever & Ratio		0.91	2.8E−08	0.98	1.7E−16	0.39	5.2E−29	Yes
PDE3B	11	Current & FEV_1	7.2 (4–9)	7.1E−06	2.9E−07	0.94	0.37	0.25	0.44	Yes

Table 2. Significant genes in the meta-analyses (significant threshold = 5.7E−07).

pulmonary function based on large samples. In discovery, we used genetic and phenotypic information from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE)[24] and SpiroMeta Consortia, and we sought replication of findings in the UK BiLEVE and UK Biobank studies.

Results

Characteristics and sample sizes of the study cohorts are presented in Table 1. We focused on two smoking categories throughout: ever-smokers vs never-smokers and current smokers vs never-smokers and past-smokers combined. The percentage of ever-smokers per study ranged from 30 to 61% in 44,429 European Ancestry participants and the percentage of current smokers per study ranged from 6 to 25%.

Quantile–quantile plots of the meta-analysis results for the interaction and joint tests from the six combinations of smoking variables status (ever smoking or current smoking) and pulmonary function measures (FEV_1, FVC, and FEV_1/FVC ratio) of discovery stage are presented in Figure S1. Some quantile–quantile plots suggest inflation in the gene-by-current smoking interaction analyses, which could be due to the low frequency of current smokers. The few studies with larger inflation factors ($\lambda > 1.5$) in testing interactions between current smoking status and a specific lung function measure were not included in the corresponding meta-analysis of this smoking-trait combination (Table S1).

In the discovery-stage meta-analyses, *GPR126* was significantly associated with FEV_1/FVC in the gene-by-current-smoking (*p*-value = 1.9E−09) and gene-by-ever-smoking analyses (*p*-value = 2.8E−08) based on joint tests. In both replication studies, *GPR126* was also statistically significant in the joint tests for gene-by-current-smoking (*p*-value = 1.1E−16 in UK BiLEVE, 1.4E−30 in UK Biobank) and gene-by-ever-smoking (*p*-value = 1.7E−16 in UK BiLEVE, 5.2E−29 in UK Biobank, Table 2). The average number of rare variants in *GPR126* across studies was 18.8 (range from 5 to 25). The *p*-value of the interaction test was not significant for either smoking exposure evaluated (*p*-value > 0.5), suggesting that these joint test results were largely due to the genetic main effects.

Gene /variant	Smoking status and phenotype Combination	Discovery Stage		Replication Stage (UK BiLEVE)		Replication Stage (UK Biobank)	
		P-value (interaction)	*P*-value (joint)	*P*-value (interaction)	*P*-value (joint)	*P*-value (interaction)	*P*-value (joint)
Conditional gene-based analysis							
GPR126	Current & Ratio	0.19	2.44E−03	0.52	3.90E−09	0.77	9.58E−17
	Ever & Ratio	0.89	1.13E−02	0.97	4.26E−09	0.47	2.47E−16
Conditional single-variant analysis on the top signal in *GPR126*							
rs17280293	Current & Ratio	0.66	3.93E−03	0.46	3.94E−09	0.68	9.31E−17
	Ever & Ratio	0.98	6.10E−03	0.97	3.97E−09	0.89	2.44E−16

Table 3. Conditional analysis of *GPR126* (adjusting for a common variant rs3817928).

Variants	Gene	Chr	INFO	MAF (%)	Combination	Discovery Stage		Replication Stage (UK BiLEVE)		Replication Stage (UK Biobank)	
						P-value (interaction)	*P*-value (joint)	*P*-value (interaction)	*P*-value (joint)	*P*-value (interaction)	*P*-value (joint)
rs17280293	*GPR126*	6	1	2.8	Current & Ratio	0.74	8.0E−06	0.52	1.2E−16	0.50	1.5E−30
					Ever & Ratio	0.98	2.8E−05	0.99	1.6E−16	0.97	5.3E−29
rs142935352	*PDE3B*	11	0.98	0.6	Current & FEV_1	5.8E−03	0.011	0.01	0.03	0.88	0.99
rs61736639	*PDE3B*	11	0.96	0.7	Current & FEV_1	3.5E−03	1.7E−05	0.71	0.10	0.84	0.56

Table 4. Single variant analysis for the identified significant genes.

A common intronic variant rs3817928 (MAF = 22%) in *GPR126* was previously found to be significantly associated with FEV_1/FVC among European ancestry[25]. Conditioning on this common variant rs3817928, the joint test of gene-by-current-smoking analysis for *GPR126* remained nominally significant in the discovery (*p*-value = 2.4E−03) and genome-wide significant in the replication studies (*p*-value = 3.9E−09 in UK BiLEVE and *p*-value = 9.6E−17 in UK Biobank, Table 3). The joint test of gene-by-ever-smoking analysis conditioning on rs3817928 was also nominally significant in the discovery stage and genome-wide significant in the replication stage (*p*-value = 0.01 in discovery stage, *p*-value = 4.3E−09 in UK BiLEVE, and *p*-value = 2.5E−16 in UK Biobank, Table 3).

To further determine whether any single low-frequency or rare variant may be driving the aggregate variant test results, we conducted single variant interaction only and joint test analyses for all the SNVs in *GPR126* and report the most significant SNV, rs17280293 (MAF = 3%), in Table 4. Figures S2 and S3 show the regional association plots for *GPR126*. The significant joint test results for rs17280293 were replicated in the UK BiLEVE and UK Biobank studies (Table 4). This SNV is predicted to be deleterious by SIFT score (score = 0.001) in dbNSFP[26].

SNV rs17280293 is in low linkage disequilibrium (LD) with the previously reported common variant rs3817928 in *GPR126* ($R^2 = 0.13$ based on 1000 Genomes database Phase 3[27]), thus rs17280293 remained highly significant in the single-variant-based analysis when conditioning on rs3817928 (Table 3). Recently, it was shown that the rs17280293 was the leading signal that drove a gene-based association between *GPR126* and FEV_1/FVC when studying genetic main effects[15]. It was also found to be the most significant signal in the *GPR126* region associated with diffusing capacity of the lung[28]. Our analyses on the interaction-only and joint analyses further highlighted the importance of this low-frequency variant.

PDE3B was significantly associated with FEV_1 in the interaction-only test (*p*-value = 1.9E−06) and joint test (*p*-value = 2.9E−07) in the gene-by-current-smoking analyses (Table 2). The *p*-value of the interaction test was highly significant, suggesting that the joint test result was at least partially driven by the interaction effects. However, the interaction effect was not replicated in either UK BiLEVE or UK Biobank. Single-variant-based interaction and joint analyses were performed on the variants within *PDE3B* (see the regional association plots in Figures S4 and S5). The top two variants are present in Table 4, among which variant rs142935352 (MAF = 0.6%) in *PDE3B* had a significant SNV-by-current smoking interaction in relation to FEV_1 in the discovery studies and was nominally significant in the UK BiLEVE replication study for both the interaction-only (*p*-value = 0.01) and joint tests (*p*-value = 0.02). The other top variant rs61736639 (MAF = 0.7%) was not significant in the replication datasets. According to the Genotype-Tissue Expression (GTEx) project ([29], accessed on 07/18/2021) both *GPR126* (median transcripts per million, TPM = 20.2) and *PDE3B* (median TPM = 8.8) were expressed in lung tissue, relative to a minimal expression threshold of TPM > 0.1[30].

Discussion

Most genome-wide analyses of pulmonary function have focused on common variants or genetic main effects[25,31]. However, the gene-level association could be missed if the signal is driven by rare variants or there is interaction with environmental exposures. Therefore, we focused on rare and low-frequency exonic genetic variants and their interaction with cigarette smoking, the major environmental risk factor for reduced lung function. Specifically, we performed gene-based tests of GxE interaction only and joint tests assessing genetic main effects and interaction effects. The joint analysis has been shown to be nearly optimal across a variety of models for

common variants[17]. In some scenarios, the GxE interaction only tests could have higher statistical power than the joint tests[19,21]. To boost the statistical power for discovery, we combined data across thirteen studies with 44,429 European ancestry individuals through meta-analyses. Significant genes were examined in two replication studies with a total of 104,611 European ancestry individuals. This study is the first and the largest to incorporate GxE interactions in modeling rare and low-frequency variant genetic and smoking effects on pulmonary function. However, statistical power would continue to be improved with increased sample size. Table S3 and S4 provide the top 5 genes for the interaction only and joint tests, which may include the putative genes that did not reach genome-wide significance in the study.

Our results for the GxE joint analysis support the involvement of rare variants in a locus, *GPR126*. We identified one low-frequency variant rs17280293 which was successfully replicated in our study and was previously reported by Jackson et al[15] in a genetic main effect analysis. The SNV is in low LD with the known common variant, suggesting a distinct role of the rare variants. Neither gene-based nor single variant tests identified interaction with smoking for this SNV or this gene. Additionally, we found that *PDE3B* was significantly associated with FEV_1 in the GxE interaction and joint association analysis in the discovery studies, however neither the interaction or joint effect gene-based tests replicated, although one leading rare variant rs142935352 reached nominal significance in UK BiLEVE for the interaction and joint tests. *PDE3B* had not been reported to be associated with lung function but is believed to be an important regulatory factor in modulating the inflammatory response and expressed in the lung[32]. One possible explanation is that smoking increases the inflammatory response in the lung, leading to the gene-by-smoking interaction. The signal of *PDE3B* in the discovery dataset was mainly driven by two SNVs (rs142935352 and rs61736639), both quite rare (MAF < 0.5% in EA), thus difficult to replicate even with large samples. We found that it was challenging to replicate rare signals in general, as rare variants varied across different studies. A variant present in one study was likely not included in the other study (for example, being monomorphic or having high missing rates).

Our study has limitations. First, the results were obtained from European ancestry populations, although we recognized the importance of conducting multi-ancestry studies to identify novel association signals[31]. Second, a few quantile–quantile plots suggested inflation in the gene-by-current-smoking interaction and joint test statistics (Figure S1). GxE analysis is well-known to be prone to systematic inflation for common variants and there could be various factors driving the inflation[33–35]. It is not surprising to see such inflation in rare-variant-based GxE analysis. We speculate that the inflation occurred in some studies due to a combination of a low proportion of current smokers and relatively small sample size, where the asymptotic property of the test statistics may not hold. For example, LBC1936 with a sample size of 970 and 11% of current smokers had an inflation factor > 1.5 for gene-by-current smoking interaction tests but not gene-by-ever smoking interaction (Table S1). Although we took a number of steps to reduce the inflation (adjusting for principal components, excluding studies with large inflation factors, excluding phenotypic outliers followed by inverse normal transformation of the spirometric measures of pulmonary function, excluding a study for a particular gene if it had low minor allele count (< 20), and using dichotomized smoking risk factors[21]), there could still be some inflation. Third, we examined exonic rare variants, but we note that signals from rare variants in the regulatory regions are likely to interact with smoking. Inclusion of rare regulatory variants could improve our understanding of disease mechanisms and warrants further investigation. Fourth, our analysis did not take into account the type of cigarette smoking, which could increase the difficulty of replication due to the potential heterogeneity of GxE interaction effects. Fifth, rare and low-frequency variants in the two replication datasets UK BiLEVE and Biobank Study were imputed using the 1000 Genomes and UK10K haplotype reference panel[36]. Although the imputation quality had been examined through a pseudo-GWAS simulation[36], it may still be suboptimal comparing with sequencing data, thus increasing the replication difficulty. In our case, SNVs included in *GPR126* and *PDE3B* all had INFO score > 0.8 and the leading SNVs had INFO score > 0.95 (Table 4), thus less likely to be affected by the genotyping array platform.

The strengths of the study include the relatively large sample size with genotyping of rare and low-frequency variants. In addition, we employed recently developed methods for incorporating environmental interactions in the study of the influence of rare variants in disease to decrease the multiple testing burden, potentially increase the statistical power. We found that one of the first discovered genes associated with pulmonary function in GWAS of common variants, *GPR126*, likely has low-frequency variants contributing to its role in this phenotype and other putative genes. The gene was successfully replicated but the signal may be driven by the genetic main effect.

Materials and methods

Ethics statement. All the studies contributed to our analyses have their protocols approved by the respective local Institutional Review Board with the details in the Supplementary Materials. All participants wrote informed consent for the genetic studies and all experiments were performed in accordance with relevant guidelines and regulations.

Cohort studies. We combined data from 13 studies for the discovery stage, either from the CHARGE Consortium or the SpiroMeta Consortium: the Age, Environment, Susceptibility (AGES) study, Atherosclerosis Risk in Communities (ARIC) study[37], British 1958 Birth Cohort (1958BC)[38], Cardiovascular Health Study (CHS)[39], Framingham Heart Study (FHS)[40,41], Health, Aging, and Body Composition (HABC) study[42], Northern Finland Birth Cohort of 1966 (NFBC1966)[43], Multi-Ethnic Study of Atherosclerosis (MESA)[44], Rotterdam Study (RS)[45], Study of Health in Pomerania (SHIP)[46], the Netherlands Epidemiology of Obesity Study (NEO)[47], Lothian in Birth Cohort 1936 (LBC1936) study[48], and Cooperative Health Research in the Region of Augsburg (KORA) study[49]. The discovery studies consisted of 44,429 individuals. Replication was conducted in two studies: UK

BiLEVE Study (n = 45,133) and UK Biobank (n = 59,478)[50]. Study descriptions are provided in the Supplementary Note of Jackson et al., 2018[15] and Artigas et al., 2015[51].

Pulmonary function measurements and smoking information. Spirometric measures included FEV_1, FVC, and their ratio (FEV_1/FVC). Details of the spirometry have been previously reported for the discovery[15] and replication studies[52]. Because outlier phenotype values can be influential in rare variant analyses, pulmonary function values with residuals after regressing out covariates (listed below) on each lung function trait larger than four standard deviations from the mean were examined by investigators from each cohort expert in the collection of pulmonary function data. Data points were evaluated regarding whether they were biologically plausible for that individual as opposed to more consistent with data errors. To ensure the normality of the FEV_1, FVC, and FEV_1/FVC, a ranked-based inverse-normal transformation was performed on each pulmonary function trait by the individual participating studies. For FEV_1/FVC, the ratio was first calculated prior to transformation.

Cigarette smoking can have important adverse effects on pulmonary function. Smoking exposure was ascertained by questionnaire at the same time as the pulmonary function tests used in the analyses. Study participants were classified in two different ways: ever smokers vs never-smokers and current smokers vs never-smokers and past-smokers combined. Current smokers were defined as individuals who smoked at least one cigarette per day within the prior 12 months, past smokers were defined as smoking at least one cigarette per day but had stopped at least 12 months, and never smokers reported never having smoked. Ever smokers were a combination of current and past smokers.

Genotyping and quality control. Genotyping of the discovery stage studies was performed mainly using the Illumina HumanExome BeadChip (Table S2). To improve accurate calling of rare variants, genotyped data from nine studies (ARIC, FHS, RS, CHS, AGES, LBC1936, and MESA) were called using GenCall in Illumina's Genome Studio and the curated clustering files from the CHARGE joint calling effort[53]. The remaining studies (1958BC, KORA, and SHIP) called their genotypes in accordance to the UK exome chip consortium best practices[36] using zCall[54].

All studies performed comparable sample-level quality control steps and removed individuals according to the standard quality control metrics, i.e., sex mismatch, duplicate pairs, unexpected relatives, missing pulmonary function measurement, or missing covariate. Other than the standard variant-level quality control step[53], we retained only nonsynonymous variants annotated by dbNSFP[26] with MAF less than 5% among EAs and excluded monomorphic variants, variants with missing rates larger than 5%, and variants on the sex chromosomes. These low-frequency and rare variants were grouped by gene region in the aggregate variant tests[26]. Missing genotypes were imputed using a random draw from the binomial distribution with two trials and success probability equal to the estimated MAF. Genes with one or zero variants or overall cumulative minor allele count less than 20 were excluded.

The replication samples were genotyped using the Affymetrix UK BiLEVE or UK Biobank arrays, which both include substantial overlap with the Illumina Human Exome BeadChip[55]. Thorough sample and genotype quality control steps were undertaken before imputation to a combined 1000 Genomes[56] and UK10K Project reference panel[57]. Following imputation, SNVs were excluded if they had imputation INFO score ≤ 0.5 or minor allele count < 3. Full details of the quality control and imputation procedure of the UK BiLEVE/UK Biobank genotype data were described elsewhere[52].

Statistical analysis. Since single-variant tests are usually underpowered for analyzing rare genetic variants, we used an aggregate variant test that analyzes multiple rare variants in a gene. For studies other than FHS, we utilized the R package rareGE[58] to conduct gene-based interaction and joint tests that simultaneously test the genetic main effects as well as GxE interaction effects. As a set-based variance component test, rareGE has been shown to be powerful across different underlying GxE association patterns[19]. For FHS, we used a modified rareGE joint test which incorporates correlation among family members by including a random intercept with covariance structure proportional to the kinship matrix[19,59]. The rareGE interaction, joint and modified joint tests are briefly described in the Supplementary Materials. Both smoking variables were included in the linear regression model as covariates, while one of them was modeled as the primary exposure variable of interest and was tested for interactions. This resulted in a total of six pairs of pulmonary function traits (transformed FEV_1, FVC, and ratio) and smoking status (ever vs never smoking or current vs past plus never smoking). Additional covariates included age, age-squared, sex, height, height-squared, recruitment sites, and the first 10 principal components accounting for any underlying population substructure. Principal components were constructed by each study using GWAS SNVs or exome-chip ancestry informative markers if GWAS not available[55]. Weight was only included as an adjustment variable for FVC as it is more strongly influenced by adiposity. Study-specific results were meta-analyzed by combining the p-values of the gene-based tests from each discovery study using Stouffer's weighted Z-score method[60]: (1) transform p-values to z-values by the inverse CDF of a Gaussian distribution, i.e., $Z_i = \Phi^{-1}(P_i)$, (2) weight the z-values of each study by the sample size, i.e., $Z = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$, where $w_i = \sqrt{N_i}$ and N_i was the ith study's sample size. The final meta-analysis p-value was calculated as $P = \phi(Z)$. Studies with genomic control inflation factor $\lambda > 1.5$ were excluded.

An a priori significance threshold was defined using the Bonferroni correction for the number of tests evaluated. Specifically, we used a significant threshold at 5.7E−07 based on 14,591 genes for 6 smoking status-lung outcome combinations. Significant genes were evaluated for replication in the UK BiLEVE and UK Biobank studies. To identify specific low-frequency or rare variants driving the signals from the aggregate variant tests,

we conducted single variant analyses for the significant genes in both the discovery and replication datasets. Additionally, conditional analyses were performed for significant replicated genes (*GPR126* herein) with the previously reported leading common SNV, where the reported signal drivers were added in the gene-based test as covariates. For conditional analyses, *p*-values less than 0.05 were considered as nominally significant. We further investigated the tissue-specific gene expression of the significant genes on the GTEx Project portal (https://gtexportal.org/).

Received: 3 October 2020; Accepted: 2 September 2021
Published online: 29 September 2021

References

1. Myint, P. K. *et al.* Respiratory function and self-reported functional health: EPIC-Norfolk population study. *Eur. Respir. J.* **26**(3), 494–502 (2005).
2. Burney, P. G. J. & Hooper, R. Forced vital capacity, airway obstruction and survival in a general population sample from the USA. *Thorax* **66**(1), 49–54 (2011).
3. Young, R. P., Hopkins, R. & Eaton, T. E. Forced expiratory volume in one second: Not just a lung function test but a marker of premature death from all causes. *Eur. Respir. J.* **30**, 616–622 (2007).
4. Schünemann, H. J., Dorn, J., Grant, B. J. B., Winkelstein, W. & Trevisan, M. Pulmonary function is a long-term predictor of mortality in the general population: 29-Year follow-up of the Buffalo Health Study. *Chest* **118**(3), 656–664 (2000).
5. Mannino, D. M., Buist, A. S., Petty, T. L., Enright, P. L. & Redd, S. C. Lung function and mortality in the United States: Data from the First National Health and Nutrition Examination Survey follow up study. *Thorax* **58**(5), 388–393 (2003).
6. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**(D1), D1005–D1012 (2019).
7. Shrine, N. *et al.* Moderate-to-severe asthma in individuals of European ancestry: A genome-wide association study. *Lancet Respir. Med.* **7**(1), 20–34 (2019).
8. Lutz, S. M. *et al.* A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet.* **16**(1), 138 (2015).
9. Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): A genetic association study in UK Biobank. *Lancet Respir. Med.* **3**(10), 769–781 (2015).
10. Loth, D. W. *et al.* Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat. Genet.* **46**(7), 669–677 (2014).
11. Artigas, M. S. *et al.* Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat. Genet.* **43**(11), 1082–1090 (2011).
12. Repapi, E. *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat. Genet.* **42**(1), 36–44 (2010).
13. Wilk, J. B. *et al.* A genome-wide association study of pulmonary function measures in the framingham heart study. *PLoS Genet.* **5**(3), e1000429 (2009).
14. Hancock, D. B. *et al.* Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS Genet.* **8**(12), e1003098 (2012).
15. Jackson, V. E. *et al.* Meta-analysis of exome array data identifies six novel genetic loci for lung function. *Wellcome Open Res.* **3**, 4 (2018).
16. Wei, P., Liu, X. & Fu, Y. X. Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: A comparative study. *BMC Proc.* **5**(SUPPL. 9), S20 (2011).
17. Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* **63**(2), 111–119 (2007).
18. Manning, A. K. *et al.* Meta-analysis of gene-environment interaction: Joint estimation of SNP and SNP × environment regression coefficients. *Genet. Epidemiol.* **35**(1), 11–18 (2011).
19. Chen, H., Meigs, J. B. & Dupuis, J. Incorporating gene-environment interaction in testing for association with rare genetic variants. *Hum. Hered.* **78**(2), 81–90 (2014).
20. Lim, E., Chen, H., Dupuis, J. & Liu, C.-T. A unified method for rare variant analysis of gene-environment interactions. *Stat Med.* **39**(6), 801–813 (2020).
21. Yang, T., Chen, H., Tang, H., Li, D. & Wei, P. A powerful and data-adaptive test for rare-variant–based gene-environment interaction analysis. *Stat. Med.* **38**(7), 1230–1244 (2019).
22. Wang, Z. *et al.* Role of rare and low-frequency variants in gene-alcohol interactions on plasma lipid levels. *Circ. Genomic Precis. Med.* **13**, e002772 (2020).
23. Smith, P. G. & Day, N. E. The design of case-control studies: The influence of confounding and interaction effects. *Int. J. Epidemiol.* **13**(3), 356–365 (1984).
24. Psaty, B. M. *et al.* Cohorts for heart and aging research in genomic epidemiology (CHARGE) Consortium design of prospective meta-analyses of genome-wide association studies from 5 Cohorts. *Circ. Cardiovasc. Genet.* **2**, 73–80 (2009).
25. Hancock, D. B. *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42**(1), 45–52 (2010).
26. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **32**(8), 894–899 (2011).
27. Machiela, M. J. & Chanock, S. J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**(21), 3555–3557 (2015).
28. Joehanes, R. *et al.* Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18**(1), 16 (2017).
29. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**(6509), 1318–1330 (2020).
30. Everaert, C. *et al.* Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.* **7**(1), 1 (2017).
31. Wyss, A. B. *et al.* Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function. *Nat. Commun.* **9**(1), 1–5 (2018).
32. Ahmad, F. *et al.* Phosphodiesterase 3B (PDE3B) regulates NLRP3 inflammasome in adipose tissue. *Sci. Rep.* **6**, 1–3 (2016).
33. Almli, L. M. *et al.* Correcting systematic inflation in genetic association tests that consider interaction effects application to a genome-wide association study of posttraumatic stress disorder. *JAMA Psychiat.* **71**(12), 1392–1399 (2014).
34. Sul, J. H. *et al.* Accounting for population structure in gene-by-environment interactions in genome-wide association studies using mixed models. *PLoS Genet.* **12**(3), e1005849 (2016).

35. Wu, C. & Cui, Y. A novel method for identifying nonlinear gene-environment interactions in case-control association studies. *Hum. Genet.* **132**(12), 1413–1425 (2013).
36. Mahajan, A., Neil, R. & Will, R. Exome-Chip Quality Control SOP (2012).
37. The ARIC investiagors. The atherosclerosis risk in communit (ARIC) study: Design and objectives. *Am. J. Epidemiol.* **129**(4), 687–702 (1989).
38. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int. J. Epidemiol.* **35**(1), 34–41 (2006).
39. Fried, L. P. *et al.* The cardiovascular health study: Design and rationale. *Ann. Epidemiol.* **1**(3), 263–276 (1991).
40. Feinleib, M., Kannel, W. B., Garrison, R. J., McNamara, P. M. & Castelli, W. P. The framingham offspring study. Design and preliminary data. *Prev. Med. (Baltim)* **4**(4), 518–525 (1975).
41. Splansky, G. L. *et al.* The third generation cohort of the national heart, lung, and blood institute's framingham heart study: Design, recruitment, and initial examination. *Am. J. Epidemiol.* **165**(11), 1328–1335 (2007).
42. Georgiopoulou, V. V. *et al.* Lung function and risk for heart failure among older adults: The health ABC study. *Am. J. Med.* **124**(4), 334–341 (2011).
43. Rantakallio, P. The longitudinal study of the Northern Finland birth cohort of 1966. *Paediatr. Perinat. Epidemiol.* **2**(1), 59–88 (1988).
44. Bild, D. E. *et al.* Multi-ethnic study of atherosclerosis: Objectives and design. *Am. J. Epidemiol.* **156**(9), 871–881 (2002).
45. Hofman, A. *et al.* The Rotterdam Study: Objectives and design update. *Eur. J. Epidemiol.* **22**(11), 819–829 (2007).
46. Völzke, H. *et al.* Cohort profile: The study of health in Pomerania. *Int. J. Epidemiol.* **40**(2), 294–307 (2011).
47. De Mutsert, R. *et al.* The Netherlands epidemiology of obesity (NEO) study: Study design and data collection. *Eur. J. Epidemiol.* **28**(6), 513–523 (2013).
48. Deary, I. J., Gow, A. J., Pattie, A. & Starr, J. M. Cohort profile: The lothian birth cohorts of 1921 and 1936. *Int. J. Epidemiol.* **41**(6), 1576–1584 (2012).
49. Holle, R., Happich, M., Löwel, H. & Wichmann, H. E. KORA: A research platform for population based health research. *Gesundheitswesen* **67**, 19–25 (2005).
50. Sudlow, C. *et al.* UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**(3), e1001779 (2015).
51. Artigas, M. S. *et al.* Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat. Commun.* **6**, 8658 (2015).
52. Wain, L. V. *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat. Genet.* **49**(3), 416–425 (2017).
53. Grove, M. L. *et al.* Best practices and joint calling of the HumanExome BeadChip: The CHARGE Consortium. *PLoS ONE* **8**(7), e68095 (2013).
54. Goldstein, J. I. *et al.* Zcall: A rare variant caller for array-based genotyping. *Bioinformatics* **28**(19), 2543–2545 (2012).
55. Jackson, V. E. *et al.* Exome-wide analysis of rare coding variation identifies novel associations with COPD and airflow limitation in MOCS3, IFIT3 and SERPINA12. *Thorax* **71**(6), 501–509 (2016).
56. Altshuler, D. M. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65 (2012).
57. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**(1), 1–9 (2015).
58. Chen, H. rareGE: Testing Gene-Environment Interaction for Rare Genetic Variants. R package version 0.1 (2014). https://CRAN.R-project.org/package=rareGE.
59. Chen, H., Meigs, J. B. & Dupuis, J. Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.* **37**(2), 196–204 (2014).
60. Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams, Jr. R. M. The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1. The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1 (1949).

Acknowledgements

We gratefully acknowledge the contribution of LBC1936 co-author Professor John M. Starr, who died prior to the publication of this manuscript. Drs. Yang, Morrison and Wei were supported by the NIH Grant R21HL126032; Dr. London was supported by the Intramural Research Program of NIH, National Institute of Environmental Health Sciences (Z01 ES43012); Dr. Manichaikul was supported by NIH Grant R01 HL131565. MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176, and the National Center for Advancing Translational Sciences, Grant UL1TR001881. The MESA Lung study was supported by grants R01 HL077612, R01 HL093081 and RC1 HL100543 from the NHLBI. This publication was developed under a STAR research assistance agreement, No. RD831697 (MESA Air), awarded by the U.S Environmental protection Agency. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and the EPA does not endorse any products or commercial services mentioned in this publication. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Health, Aging, and Body Composition (Health ABC) was supported by NIA contracts N01AG62101, N01AG2103, and N01AG62106, and in part by the Intramural Research Program of NIA. This work was also supported, in part, by Intramural Research Programs of the NHGRI. The genome-wide association study in Health ABC was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences, and genotyping services were provided by the Center for Inherited Disease Research, which is fully funded through an NIH contract to The Johns Hopkins University (HHSN268200782096C). This research was further supported by RC1AG035835. Phenotype collection in the LBC1936 was supported by Age UK (The Disconnected Mind project). Genotyping was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, and the Royal Society of Edinburgh. The work was undertaken by The University of Edinburgh Centre for Cognitive Ageing and Cognitive

Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (MR/K026992/1). Funding from the BBSRC and Medical Research Council (MRC) is gratefully acknowledged. M.D. Tobin is supported by a Wellcome Trust Investigator Award (WT202849/Z/16/Z). M.D. Tobin has been supported by the MRC (MR/N011317/1). The research was partially supported by the NIHR Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 07/18/21. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195, HHSN268201500001I and 75N92019D00031). Genotyping, quality control and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study was supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research (Daniel Levy and Christopher J. O'Donnell, Principle Investigators). The Rotterdam Study is funded by Erasmus MC and Erasmus University Rotterdam; the Netherlands Organisation for the Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly (RIDE); the Ministry of Education, Culture and Science; the Ministry for Health, Welfare and Sports; the European Commission (DG XII); and the Municipality of Rotterdam. This work was supported by the Fund for Scientific Research Flanders (FWO) project (3G037618). The Atherosclerosis Risk in Communities study (ARIC) has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). For acknowledgements to other participating cohorts (AGES, 1985BC, NFBC1966, SHIP, CHS, NEO, KORA, UK BioBank and UK Believe), please refer to [15].

Author contributions

T.Y. wrote the main manuscript, performed, and performed the meta-analysis, while T.Y., A.V.S., J.X., T.M.B., C.M.S., S.A.G., J.D., A.M., N.T., V.E.J., C.F., R.N., S.W., S.E.H., and N.W. performed the study-specific analysis; P.W., A.C.M., S.J.L. from the CHARGE. consortium supervised the study; all authors reviewed the manuscript.

Competing interests

The authors of this manuscript have the following potential competing interests: BMP serves on the DSMB of a clinical trial funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale Open Data Access Committee funded by Johnson & Johnson. DOMK is a part-time research consultant at Metabolon, Inc. All other authors do not have any competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-98120-7.

Correspondence and requests for materials should be addressed to S.J.L., P.W. or A.C.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.