Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-24-4

Reviewer A

Major comments:

C1: The study is based on publicly available data. There is a lack of details in the Methods section to reproduce the results.

R1: We thank the reviewer for raising this point. We have added details to the methods section in line with the following issues you raised.

1. We added the data field and ICD-10 codes of respiratory diseases included in the study.
2. We specified the study population and made a detailed description and presentation of the population.
3. We added details on the construction and use of PGS used in mediation analysis. Shared variants and weights used for each PGS also be provided in result section.
4. We added description of call rate to the quality control section, all the variants with call rate < 90% and minor allele count (MAC) ≤ 1 were filtered out.
5. We explained the covariates included in the model by association test.
6. We had corrected some unclear content in the method section.

C2: Specific data fields to define all respiratory diseases in UKB must be provided, at least in supplementary data.

R2: We apologize for not providing all the data fields for respiratory diseases and we have added clarification to this part.

Methods section

ICD-10 codes for other diseases included in the study is shown in Table S1. (Page 8, Line 132-133)

Table S1. Lung cancer-related diseases definition and extraction in UK Biobank

Respiratory diseases	ICD-10(data fields:41270)
Asthma	J45
COPD	J44
Emphysema	J43
Fibrosis	J84.1
Pneumonia	J18
Bronchiectasis	J47
Acute Bronchitis	J20, J21, J22
Chronic Bronchitis	J40, J41, J42, Data-Field 22129
Tuberculosis	A15

C3: It is unclear whether the authors have removed non-White British ancestry? This is unlikely based on the UKB sample size of 427,934.

R3: Indeed, we did not specify the criteria for sample size determination in detail. We selected 472,038 European individuals of White British (Data-Coding 1001), White Irish (Data-Coding 1002) and any other white background (Data-Coding 1003) by Ethnic background (Data-Field 21000). Of them, 427,934 individuals have available whole-exome sequencing data.

Methods section

The UK Biobank (UKB) provides detailed diseases follow-up information linked to whole-exome sequencing (WES) for approximately 450,000 participants (data field: 23148). We included 427,934 white European participants in this research, and detailed inclusion information is presented in Table S2. (Page 8, Line 134-137)

Table S2. Study population included in the study

Ethnic background	Total number of UKB participants	Number of WES participants included
White British	442,510	401,277
White Irish	13,201	11,916
Any other white background	16,327	14,741
Total	472,038	427,934

C4: Shared variants and weights used for each PRS must be provided. The performance of these PRSs in estimating the risk of paired diseases is not reported. In addition, the cumulative proportion of mediation (page 13, line 402) is a misinterpretation.

R4(part1): Thanks for your insightful comments!

According to your comments, we provide a supplement to demonstrate that these PGSs were statistically significant, and details of the shared variants and their weights are provided in the Supplementary Material. It is worth noting that PGS is not applied here for the purpose of disease risk stratification or prediction, but for the purpose of using the idea of PGS to comprehensively measure the impact of all shared variants and to calculate the mediation effect using a unified indicator. We supplemented the area under the receiver operator characteristic curves (AUC) of all PGSs used for mediation analyses because only shared variants intersecting with lung cancer were selected and only PGS variables were included in the model. The AUCs are moderate but statistically significant. It has also been demonstrated that PGS does not enhance the model AUC significantly: “The overall AUC did not substantially change when adding PRS for overall population with AUC of 0.832 (from AUC of 0.828 without PRS)” (Hung, R. J, et al. (2021). Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model. *Cancer research*, *81*(6), 1607–1615.)

Methods section

PGS is not applied here for the purpose of disease risk prediction, but for the purpose of using the idea of PGS to comprehensively measure the impact of all shared variants and calculate the mediation effect using a unified indicator…. We calculated the area under the receiver operator characteristic curves (AUC) of all PGSs used for mediation analyses using Bootstrap. (Page 13, Line 227-233)

Result section

Based on the identified pleiotropic variants, we screened for shared genetic variants for the respiratory diseases and lung cancer, and the shared variants and their weights are provided in Table S6. Then the polygenic score (PGS) was constructed for these five respiratory diseases. The AUCs (95% CI) of PGS_AS, PGS_COPD, PGS_EM, PGS_FI, and PGS_PN are shown in Table S7. Because only shared variants with lung cancer were included in the PGS models, the AUCs performed moderately, but they were all statistically significant. (Page 18, Line 331-338)

Table S6. Shared variants and weights included in polygenic scores applied to mediation analyses

	MarkerID	Weight
PGS_AS (Shared variants of lung cancer and asthma)	1:12115601:G>A	-0.211
	1:152312600:CACTG>C	0.243
	1:153390079:C>A	2.404
	1:156126783:C>T	2.412
	1:171783869:C>T	0.856
	1:179468521:A>T	8.518
	1:181052180:A>ACC	3.338
	1:210674411:G>A	6.702
	1:228280287:C>T	7.565
	1:23969084:CTTCA>C	6.108
	1:38017675:C>T	5.541
	1:53264370:G>A	3.200
	1:59321597:C>T	2.214
	1:77979080:C>CGGCCG	0.045
	10:132909243:C>T	3.768
	11:61783884:T>C	-0.046
	12:71137883:A>T	-0.044
	13:99367958:G>A	-0.070
	14:94055016:T>C	3.106
	15:36657835:G>A	4.312
	16:21069624:A>G	4.804
	16:27345038:C>T	0.059
	17:49406866:G>C	0.043
	17:74843606:C>G	1.397
	17:75506666:C>A	1.316
	19:36394165:A>T	0.027
	19:6418576:G>A	3.995
	2:46378030:A>G	5.495
	2:69437519:C>T	8.045
	3:10289884:A>G	1.174
	3:122635777:A>G	3.957
	3:49860397:C>T	0.034
	4:102267552:C>T	0.076
	4:122177976:G>A	0.084
	5:157494643:G>A	0.073
	6:31862816:A>C	-0.158
	6:32061449:A>G	0.069
	6:32222629:A>G	0.048

	6:32581599:C>G	0.129
	6:32641280:G>A	0.062
	6:32642188:C>A	0.151
	6:32666596:C>T	0.065
	6:32759225:A>G	0.066
	6:32837693:C>G	0.055
	6:33406176:C>A	0.113
	9:137169832:G>C	1.862
PGS_COPD (Shared variants of lung cancer and COPD)	1:153390079:C>A	2.524
	1:156126783:C>T	5.275
	1:16208697:G>A	2.112
	1:179468521:A>T	4.761
	1:181052180:A>ACC	2.782
	1:186357567:A>T	7.112
	1:197735372:A>G	2.145
	1:210674411:G>A	5.148
	1:22569245:A>C	6.347
	1:23969084:CTTCA>C	4.566
	1:38017675:C>T	10.158
	1:53264370:G>A	4.364
	1:59321597:C>T	3.595
	1:6633018:C>T	2.588
	1:77092478:T>C	4.221
	1:77979080:C>CGGCCG	0.073
	10:125836188:A>G	8.813
	10:132909243:C>T	12.403
	11:10760381:T>C	5.386
	11:18245924:A>G	-0.074
	11:61783884:T>C	-0.029
	12:55935829:C>T	3.802
	12:71137883:A>T	-0.024
	13:109127272:TC>T	1.218
	14:94055016:T>C	10.449
	15:36657835:G>A	8.755
	15:40611486:A>G	0.049
	15:78596058:G>A	-0.081
	15:78618839:T>C	0.121
	15:88872824:C>G	3.635
	16:21069624:A>G	10.619
	17:10494304:C>T	11.169
	17:41727795:C>T	7.513
	17:49406866:G>C	0.033
	17:74843606:C>G	2.830

	17:75506666:C>A	3.364
	19:12828670:G>A	5.313
	19:31277740:C>T	6.367
	19:32383003:G>A	5.512
	19:36394165:A>T	0.047
	19:45406523:C>T	4.642
	19:6418576:G>A	2.173
	2:181892321:G>A	7.431
	2:241899324:C>T	0.912
	2:26471077:A>G	6.805
	2:65308295:G>A	10.118
	2:69437519:C>T	3.964
	2:71364133:A>C	2.005
	2:80313425:G>A	1.468
	2:9356285:A>G	9.770
	20:32309937:C>T	2.985
	20:63355597:T>C	0.078
	3:10289884:A>G	1.958
	3:122635777:A>G	5.337
	3:49860397:C>T	0.033
	4:102267552:C>T	0.087
	4:122152554:G>A	6.013
	4:122177976:G>A	0.075
	5:157494643:G>A	0.077
	6:32061449:A>G	0.040
	6:32666596:C>T	0.037
	6:32759225:A>G	0.038
	6:32837693:C>G	0.047
	7:6643851:G>C	14.161
	8:12434171:C>A	5.365
	8:23295975:C>T	2.549
	9:137169832:G>C	3.477
	9:92517677:A>C	5.715
PGS_EM (Shared variants of lung cancer and emphysema)	1:12115601:G>A	-0.195
	1:152312600:CACTG>C	0.145
	1:153390079:C>A	2.131
	1:156126783:C>T	2.345
	1:16208697:G>A	3.927
	1:171783869:C>T	0.881
	1:179468521:A>T	10.408
	1:181052180:A>ACC	1.110
	1:186357567:A>T	9.926
	1:210674411:G>A	6.769

	1:22569245:A>C	4.441
	1:228280287:C>T	21.730
	1:23969084:CTTCA>C	6.129
	1:38017675:C>T	15.774
	1:53264370:G>A	3.954
	1:59321597:C>T	4.852
	1:6633018:C>T	3.950
	1:77092478:T>C	6.686
	1:77979080:C>CGGCCG	0.032
	10:125836188:A>G	8.104
	10:132909243:C>T	43.452
	11:10760381:T>C	5.886
	11:18245924:A>G	-0.038
	12:55935829:C>T	21.029
	13:109127272:TC>T	0.836
	14:94055016:T>C	28.505
	15:36657835:G>A	34.718
	15:40611486:A>G	0.046
	15:78596058:G>A	-0.162
	15:78618839:T>C	0.174
	15:88872824:C>G	14.268
	16:21069624:A>G	26.152
	17:10494304:C>T	29.321
	17:41727795:C>T	27.914
	17:49406866:G>C	0.060
	17:74843606:C>G	11.418
	19:12828670:G>A	8.369
	19:31277740:C>T	20.260
	19:32383003:G>A	3.502
	19:36394165:A>T	0.065
	19:45406523:C>T	7.743
	19:6418576:G>A	8.592
	2:181892321:G>A	8.655
	2:241899324:C>T	0.616
	2:26471077:A>G	4.327
	2:46378030:A>G	3.987
	2:65308295:G>A	9.197
	2:69437519:C>T	7.580
	2:71364133:A>C	5.793
	2:80313425:G>A	1.220
	2:9356285:A>G	26.627
	20:32309937:C>T	12.488
	20:63355597:T>C	0.116

	22:20146081:C>T	7.744
	3:10289884:A>G	3.360
	3:122261452:A>G	27.197
	3:122635777:A>G	24.794
	3:49860397:C>T	0.021
	4:122152554:G>A	3.213
	4:122177976:G>A	0.053
	6:32061449:A>G	0.030
	6:32222629:A>G	0.042
	6:32581599:C>G	0.163
	6:32642188:C>A	0.076
	6:32759225:A>G	0.030
	6:32837693:C>G	0.045
	6:33406176:C>A	0.073
	7:6643851:G>C	12.032
	8:12434171:C>A	11.929
	8:23295975:C>T	6.673
	9:137169832:G>C	3.319
	9:92517677:A>C	8.150
PGS_FI (Shared variants of lung cancer and fibrosis)	1:12115601:G>A	-0.266
	1:16208697:G>A	9.055
	1:186357567:A>T	20.272
	1:197735372:A>G	2.843
	1:22569245:A>C	28.831
	1:23969084:CTTCA>C	10.680
	1:53264370:G>A	14.821
	1:59321597:C>T	2.371
	1:6633018:C>T	7.547
	1:77092478:T>C	3.132
	1:77979080:C>CGGCCG	0.111
	10:125836188:A>G	15.239
	11:18245924:A>G	-0.132
	11:61783884:T>C	-0.041
	12:55935829:C>T	20.359
	15:40611486:A>G	0.092
	15:78596058:G>A	-0.047
	17:75506666:C>A	7.151
	19:31277740:C>T	17.323
	19:32383003:G>A	14.913
	19:45406523:C>T	8.655
	2:181892321:G>A	32.317
	2:241899324:C>T	1.007
	2:26471077:A>G	15.400

	2:46378030:A>G	6.833
	2:65308295:G>A	10.994
	2:69437519:C>T	12.795
	2:80313425:G>A	1.514
	2:9356285:A>G	15.111
	20:32309937:C>T	10.559
	20:63355597:T>C	0.079
	22:20146081:C>T	6.023
	3:122261452:A>G	12.404
	3:122635777:A>G	14.835
	3:49860397:C>T	0.029
	4:122177976:G>A	0.047
	5:157494643:G>A	0.064
	6:32666596:C>T	0.032
	6:32837693:C>G	0.031
	6:33406176:C>A	0.057
	8:12434171:C>A	13.409
	9:137169832:G>C	2.896
PGS_PN (Shared variants of lung cancer and pneumonia)	1:156126783:C>T	2.849
	1:16208697:G>A	1.737
	1:171783869:C>T	0.754
	1:181052180:A>ACC	1.134
	1:186357567:A>T	8.067
	1:197735372:A>G	1.319
	1:22569245:A>C	4.761
	1:228280287:C>T	5.174
	1:38017675:C>T	3.115
	1:53264370:G>A	5.416
	1:77092478:T>C	3.174
	1:77979080:C>CGGCCG	0.084
	10:132909243:C>T	4.245
	11:10760381:T>C	2.704
	11:18245924:A>G	-0.057
	12:55935829:C>T	6.230
	13:109127272:TC>T	0.789
	14:94055016:T>C	4.984
	15:40611486:A>G	0.030
	15:78596058:G>A	-0.048
	15:88872824:C>G	4.347
	17:10494304:C>T	6.748
	17:49406866:G>C	0.039
	17:74843606:C>G	2.919
	19:12828670:G>A	4.505

	19:32383003:G>A	3.568
	19:36394165:A>T	0.027
	19:45406523:C>T	4.196
	2:181892321:G>A	4.711
	2:26471077:A>G	2.835
	2:46378030:A>G	2.205
	2:65308295:G>A	8.223
	2:69437519:C>T	5.353
	20:32309937:C>T	2.170
	22:20146081:C>T	5.938
	3:10289884:A>G	1.896
	3:122635777:A>G	2.512
	3:49860397:C>T	0.029
	4:122152554:G>A	6.425
	7:6643851:G>C	5.607
	8:12434171:C>A	3.318
	8:23295975:C>T	1.205
	9:92517677:A>C	9.555

Table S7. AUC of PGSs used for mediation analyses

PGS based on shared variants	AUC	95% CI	*P*
PGS_AS	0.536	0.533-0.539	4.70E-207
PGS_COPD	0.543	0.539-0.547	1.46E-278
PGS_EM	0.562	0.553-0.571	2.76E-318
PGS_FI	0.544	0.533-0.556	2.89E-130
PGS_PN	0.524	0.520-0.528	2.83E-149

R4(part2): Thanks for pointing out our misrepresentation. We have modified this section.

Discussion section

The mediating effect was significant for all five-lung cancer-related diseases, and the proportions of the mediating effect for COPD, emphysema, and pneumonia all exceeded 20%, … (Page 24, Line 447-450)

C5: Figure 4C (UpSet plot) must be clarified. The legend indicates the numbers of pleiotropic variants and genes. You would need one UpSet plot for variants and another one for genes.

R5: Thanks for pointing out the errors.

We misrepresented variants as genes, and apologize for our carelessness. We have fixed the error.

Figure legends section

(C) UpSet plot to illustrate the numbers (N > 5) and distribution of pleiotropic variants shared across Lung cancer-related respiratory diseases and the number of pleiotropic variants in each lung cancer-related respiratory diseases. (Page 31, Line 621)

C6: Pathway analyses are exploratory in nature and not very informative. Genes associated with immune system function and cancer development were expected.
R6: Thank you for the insightful comment!
We highly agree with your viewpoint, and have moved the presentation of results from this section to the supplementary material. This section is partly to enrich the article with exploratory analysis, and partly to confirm the reliability of these variants by exploring their function. These genes are associated with the immune system and cancer development also explain the reasonableness for the presence of pleiotropic genes among respiratory diseases.
Result section
Figure 6 => Figure S3 (Page 19, Line 360)

C7: In the Discussion, the authors claimed a few novel genes, namely HSD3B7 for lung cancer and SRSF2 and JAK2 for pneumonia. This is relatively disappointing from whole-exome sequencing of more than 400,000 individuals. Globally, the results seem to mitigate the statement in the introduction about the power of large-scale exome sequencing to identify rare coding variants.
R7: We thank the reviewer for pointing this out, and added to the discussion section.
We did not identify abundant novel genes due to the insufficient incident cases and lack of statistical power. This study was conducted focusing on lung cancer. However, the number of incident lung cancer cases in the UKB is low (n ≈ 4,000) and provides insufficient power to assess the effects of rare variants compared with the existing case-control studies based on SNP array. We believe that more novel genes will be found in the future as the sample size increases with additional cases.

Discussion section
Second, we focused on individuals of European ancestry only, and the number of incident lung cancer cases in the UKB is low (n ≈ 4,000) and provides insufficient power to assess the effects of rare variants. (Page 25, Line 468-470)

Minor comments:
C8: Abstract, line 45: you mean 102 variants for the six lung cancer-related diseases?
R8: We thank the reviewer for pointing this out. We have revised.

Abstract section
We identified 102 significant independent variants at single-variant levels for lung cancer and five lung cancer-related diseases. (Page 4, Line 62-63)
C9: Abstract, line 53: “6 to 23”. This is not clear from reading just the abstract?
R9: I'm sorry for not explaining clearly, and I have made some revisions.

Abstract section
Meanwhile, the proportion of mediation effects of these variants ranged from 6 to 23 (emphysema:23%; COPD:20%; pneumonia:20%; fibrosis:7%; asthma:6%) through these five respiratory diseases to the incidence of lung cancer. (Page 5, Line 71-74)

C10: Page 4, lines 124-125: Unclear if the authors have conducted the GWAS themselves. What do you mean by "GWAS data downloaded"? Genotyping data to carry out the GWAS or GWAS summary statistics already available?
R10: Thanks for pointing this out. We apologize for not explaining it clearly. We conducted analysis using the genotype data mentioned in the preceding sentence. (the imputed genetic variants from the UK Biobank (data field: 22828))

Methods section
We conducted a genetic correlation analysis on ten respiratory diseases using the imputed genotype data from the Haplotype Reference Consortium (HRC) and UK10K haplotype resource (Data Field 22828), … (Page 9, Line 147-150)

C11: Page 5, line 148: How smoking status was defined?
R11: Derived using variables "Current tobacco smoking" (Field 1239) and "Past tobacco smoking" (Field 1249).
Individual classed as Ever-smoker if Current tobacco smoking= most days (1) or occasionally (2) OR Past tobacco smoking= most days (1) or occasionally (2) or tried once or twice (3).
Individual were classed as Never-smoker if Current tobacco smoking= no (0) AND Past tobacco smoking= never (4).
Individuals who answered to either question "Do not know" (-1) "Prefer not to answer" (-3) and "None of the above" (-7) were not coded.

C12: Page 6, line 178: "six diseases" is unclear at this point of the manuscript.
R12: I'm sorry for not explaining clearly, and I have made some revisions.

Methods section
Briefly, ASSET explored all possible subsets of all six diseases (five lung cancer-related diseases and lung cancer.) (Page 12, Line 207-208)

C13: Page 6, line 195: Why citing Hung et al. for this statement?
R13: Dr. Rayjean Hung is a core member of the largest international lung cancer GWAS consortium (ILCCO/TRICL). She develops lung cancer polygenic risk score (PRS) that has been demonstrated excellent efficacy in the UK Biobank (UKB). Here, we employed the same construction formula $PRS = \sum \beta_i SNV_i$ to develop PRS for mediation analysis.

C14: Figure S2: Are you presenting the adjusted inflation factors in this figure (as explained in the Methods section)?
R14: Yes, we provided adjusted inflation factor in the top left corner of each small figure.

C15: Excel files should be provided for supplementary tables.
R15: Thank you for your suggestion. We provided an Excel file for all the supplementary tables, when we first submitted the manuscript. I'm not sure if the journal has provided you with a download

option.

Reviewer B

C16: SNPs are called in WES data using a minimum DP of 7 and AB>0.15. With this setting, it appears that one read supporting the alternative base is sufficient to call a heterozygous SNP. Do you find this approach reasonable?

R16: Thanks for your insightful comments!

I apologize for the difficulty in explaining the rationale behind this approach from our professional perspective. However, all our quality control procedures are conducted in accordance with the standards provided on the UKB official website. "*These gVCFs were joint genotyped using GLnexus (https://www.biorxiv.org/content/10.1101/572347v1) to create a single, unfiltered project-level VCF (pVCF). Genotype depth filters (SNV DP≥7, indel DP≥10) were applied prior to variant site filters requiring at least one variant genotype passing an allele balance filter (heterozygous SNV AB>0.15, heterozygous indel<0.20), resulting in a second 'filtered' pVCF.*" (Category 170)

Additionally, there are several published articles that have performing quality control in this way.

1. *"SNV genotypes with read depth (DP) less than seven and indel genotypes with read depth less than ten are changed to no-call genotypes. After the application of the DP genotype filter, a variant-level allele balance filter is applied, retaining only variants that meet either of the following criteria: (i) at least one homozygous variant carrier or (ii) at least one heterozygous variant carrier with an allele balance (AB) greater than the cutoff (AB >= 0.15 for SNVs and AB >= 0.20 for indels)"* . [Backman, J. D., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature, 599(7886), 628–634.]

2. *"In the filtered GL PVCF, any SNV genotype with read depth less than seven reads (DP < 7) was changed to a no-call. After the application of the DP genotype filter, only SNV variant sites that met at least one of the following two criteria were retained: 1) at least one heterozygous variant genotype with allele balance ratio greater than or equal to 15% (AB >= 0.15); 2) at least one homozygous variant genotype "*.

[Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature 586, 749-756, doi:10.1038/s41586-020-2853-0 (2020).]

3. *'The OQFE protocol maps to a full GRCh38 reference version including all alternative contigs in an alt-aware manner. Genotype depth filters (SNV sequencing depth (DP) ≥ 7, indel DP ≥ 10) were applied prior to variant site filters requiring at least one variant genotype passing an allele balance filter (heterozygous SNV allelic balance (AB)>0.15, heterozygous indel."* [Shen, S., et al. (2023). A Large-Scale Exome-Wide Association Study Identifies Novel Germline Mutations in Lung Cancer. *American journal of respiratory and critical care medicine*, *208*(3), 280–289.]

C17: No SNP call rates are used to filter problematic variant positions.

R17: Thank you for your reminder, and we apologize for omitting this important detail. We have made the necessary revisions. Prior to conducting the association analysis, we excluded SNVs with a missing rate ≥ 10%, retaining only those with a call rate ≥90%.

Methods section

In addition, all the variants with call rate < 90% and minor allele count (MAC) ⩽ 1 were filtered out. (Page 10, Line 167-168)

C18: Association analysis is performed without considering batches as a covariate. UK Biobank best practices suggest the inclusion of a batch covariate when using WES for association tests. Additionally, the best practices recommend excluding SNPs with a call rate <90% or variant positions with DP<10.

R18: Thanks for your insightful comments！

We have only identified variable Genotype measurement batch (Data-Field 22000) regarding batches in the UKB dataset, which measured from two closely related SNP arrays (*~50,000 UK BiLEVE array and ~450,000 UK Biobank Axiom array*) but has no relationship with sequencing data (WES). Currently, no batch information is available for sequencing data because they are joint-calling simultaneously from raw sequencing reads.

C19: The authors should better justify the use of mixed models with PCs included as fixed-effect covariates. Why is there a need to add PCs as covariates in a logistic mixed model, and what is the rationale for selecting the first 5 PCs?

R19: We thank the reviewer for pointing this out. We have revised.

In genome-wide association studies, when employing mixed models for analysis, the majority of studies include population principal components as covariates in the model. The SAIGE method, which we utilized, also incorporates population principal components as covariates in the analysis. "*The non-genetic covariates of sex, birth year and principal components 1–4 were adjusted in all tests.*" [Zhou, W., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics, 50(9), 1335–1341.] In mixed models, the genetic relationship matrix (GRM) is included as a covariate in the model, "*while the PCR approach can be regarded as an approximation to a LMM; such an approximation depends on the number of the top principal components (PCs) used, the choice of which is often difficult in practice. Hence, in the presence of population structure, the LMM appears to outperform the PCR method. However, due to the different treatments of fixed versus random effects in the two approaches, we show an advantage of PCR over LMM: in the presence of an unknown but spatially confined environmental confounder (e.g. environmental pollution or life style), the PCs may be able to implicitly and effectively adjust for the confounder while the LMM cannot."* [Zhang, Y., & Pan, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genetic epidemiology*, *39*(3), 149–155.] Therefore, we included PCs as covariates in the logistic mixed model.

In selecting the number of PCs, past studies have included first 3-20 PCs in variety. *"We analyzed 103,796 longitudinal observations from 23,066 members of community-based (FHS, ACT, and ROSMAP) and clinic-based (ADRCs and ADNI) cohorts using generalized linear mixed models including terms for SNP, age, SNP × age interaction, sex, education, and five ancestry principal components."* [Kang, M., et al. (2023). A genome-wide search for pleiotropy in more than 100,000 harmonized longitudinal cognitive domain scores. *Molecular neurodegeneration*, *18*(1), 40.] We referred to this research and included first 5 PCs.

Methods section

A genetic relationship matrix (GRM) was created to fit the model to eliminate the effect of kinship. We also included five principal components in mixed model to adjust for both population structures and non-genetic confounders(19). (Page 10, Line 175-178)

C20: Why was the ASSET meta-analysis performed, selecting only SNPs with a p-value<1e-4? Why didn't the authors use the entire GWAS summary statistics for the five traits? Is this a valid approach to using the ASSET method?

R20: Thank you for raising this question. We have supplemented and clarified in the method section. ASSET explores all possible subsets of studies and evaluates fixed-effect meta-analysis-type test-statistics for each subset. [Bhattacharjee S, et al. Am J Hum Genet. 2012 May 4;90(5):821-35].

When conducting ASSET analysis, it is a typical way to screen variants firstly. For example, the following two studies [Jiang Y, et al. A cross-disorder study to identify causal relationships, shared genetic variants, and genes across 21 digestive disorders. iScience. 2023 Oct 16;26(11):108238] & [Guo P,et al.. Pinpointing novel risk loci for Lewy body dementia and the shared genetic etiology with Alzheimer's disease and Parkinson's disease: a large-scale multi-trait association analysis. BMC Med. 2022 Jun 22;20(1):214.]

The ASSET method considers correlations induced by overlapping participants across different studies (e.g., shared controls). If we incorporate all GWAS summary results into ASSET analysis using an exhaustive approach, it would result in an extremely large number of analyses to be performed (4398417*(6+15+20+15+6) = 272,701,854 times), which is infeasible to complete such high-dimensional computations.

Methods section

Because the method explores all possible subsets of studies and evaluates fixed-effect meta-analysis-type test-statistics for each subset, to avoid excessive computational effort, we used a relatively lenient p-value to comprehensively consider all suggestive association variants across the six respiratory diseases. (Page 12, Line 212-216)

C21: It is not clear what the performance metrics are for the PGS constructed for the considered traits.

C22: The mediation analysis is not well-described, and it is unclear whether the performance of PGS has an impact on it.

R21&R22: We respond C21&C22 together regarding to the PGS problem.

According to your comments, we provide a supplement to demonstrate that these PGSs were statistically significant, and details of the shared variants and their weights are provided in the Supplementary Material. It is worth noting that PGS is not applied here for the purpose of disease risk stratification or prediction, but for the purpose of using the idea of PGS to comprehensively measure the impact of all shared variants and to calculate the mediation effect using a unified indicator. We supplemented the area under the receiver operator characteristic curves (AUC) of all PGSs used for mediation analyses because only shared variants intersecting with lung cancer were selected and only PGS variables were included in the model. The AUCs are moderate but statistically significant. It has also been demonstrated that PGS does not enhance the model AUC significantly: “The overall AUC did not substantially change when adding PRS for overall population with AUC of 0.832 (from AUC of 0.828 without PRS)” (Hung, R. J, et al. (2021). Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model. *Cancer*

research, *81*(6), 1607–1615.) We applied PGS here to unify the effects of the shared variants under the hypothesis that PGS does not have an impact on the mediation analysis.
The PGS details in revised paper have been provided in response R4 or Table S6.