A Phenome-wide association study of a genetic score for the nicotine metabolite ratio

Jadwiga Buchwald[1*], Terho Lehtimäki[2,3], Olli Raitakari[4,5,6], Veikko Salomaa[7], Jaakko Kaprio[1], Matti Pirinen[1,8,9]

1 Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Science (HiLIFE), University of Helsinki, Helsinki, Finland
2 Department of Clinical Chemistry, Fimlab Laboratories, and Finnish Cardiovascular Research Center, Tampere, Finland
3 Faculty of Medicine and Health Technology, Tampere, Finland
4 Centre for Population Health Research, University of Turku and Turku University Hospital, Turku, Finland
5 Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland
6 Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland
7 Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland
8 Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
9 Department of Public Health, University of Helsinki, Helsinki, Finland

* jadwiga.buchwald@helsinki.fi

Abstract

Background

Faster nicotine metabolism associates with heavier smoking and challenges in cessation. Understanding which variables and diseases associate with the rate of nicotine metabolism, defined as the nicotine metabolite ratio (NMR), the 3-hydroxycotinine-to-cotinine-ratio, is crucial for drug development and personalized interventions for nicotine addiction.

Methods

We performed a hypothesis-free phenome-wide association study (PheWAS) of over 21 000 outcome variables from UK Biobank (UKB) to explore how the NMR associates with the phenome. As the exposure variable, we used a genetic score for faster nicotine metabolism based on 10 putatively causal genetic variants, explaining 33.8 % of the variance in the NMR. We analyzed ever and never smokers separately to assess whether the association had a causal pathway through smoking. Additionally, we performed complementary PheWASs in FinnGen and MRBase.

Results

Faster nicotine metabolism was associated with a greater number of cigarettes smoked per day, increased smoking compared to ten years ago, and greater smoking in the past.

We found, inconsistently with previous literature, faster nicotine metabolism to be associated with higher odds of quitting. Notably, we also found associations that did not appear to differ between ever and never smokers, suggesting the association pathway may not go through smoking: faster nicotine metabolism was associated with worse liver enzyme and lipid values with respect to associated diseases, as well as increased coffee and tea consumption.

Conclusions

We observed expected associations with several smoking and nicotine -related phenotypes, but could not replicate prior findings on cessation. Our other findings support a possibility that a future therapy converting fast metabolizers to slower ones could work without adverse side effects and potentially even provide other health-related benefits. Further research using different data and study designs is warranted to validate and extend these discoveries.

Keywords: Nicotine metabolism, Nicotine Metabolite Ratio, Nicotine, PheWAS, phenome, Smoking, FINEMAP, PHESANT, GxE-MR-pheWAS, UK Biobank

Introduction

Smoking remains a leading cause of global mortality [1]. While global smoking prevalence has declined due to effective tobacco control policies, the prevalence has risen or remained stagnant in many countries [2]. Furthermore, the emergence of new nicotine products, such as e-cigarettes, which have aggressively been targeted at the youth, has created a whole new generation of addicted individuals in some countries [2, 3]. Nicotine, a toxic substance in itself [4, 5], is as addictive as cocaine and heroine [6].

The Nicotine Metabolite Ratio, NMR, measured as the 3-hydoxycotinine-to-cotinine ratio, is an established biomarker for the rate of nicotine metabolism [7]. Individuals with higher NMR values, reflecting faster nicotine metabolism, typically smoke more [8, 9] and find quitting more challenging [10–12]. Personalizing cessation treatments based on an individual's NMR could improve cessation rates [13]. Another approach would be to develop smoking cessation drugs tailored to modulate metabolization rates.

The NMR is highly heritable [14], and understanding which genetic loci it associates with, can illuminate potential drug targets. The largest genome-wide association study (GWAS) on the NMR to date identified two independent association loci on chromosomes 4 and 19, explaining 1.8 % and 36.4 % of the variation in the NMR, respectively [15]. Notably, the chromosome 19 locus contains *CYP2A6*, which codes for CYP2A6, the main metabolic enzyme for both nicotine and cotinine, the metabolite to which most of nicotine is metabolized [16]. A smaller fraction of nicotine is metabolized through other enzymatic pathways [16].

Beyond genetics, it is also essential to understand what traits or diseases the NMR is associated with. This knowledge is critical for assessing possible side-effects and other opportunities for drug development. A previous study by Chenowth et al. showed the NMR to be associated with ethnicity, gender, hormonal replacement therapy, BMI, cigarettes smoked per day (CPD), and number of alcohol drinks/week [17]. In that hypothesis-driven study, the investigated variables had been carefully chosen based on previous literature. Therefore, unknown and potentially important associations may have been missed.

Phenome-wide association studies (PheWAS) present a hypothesis-free approach to discover novel associations [18]. PheWAS aim to identify associations with a genetic

instrument (e.g. a single nucleotide polymorphism (SNP) or genetic risk score), proxying a given variable of interest, across an array of phenotypes (the phenome). Two PheWAS of CYP2A6 activity have been published to date. The first, assessed 358 traits for nine *CYP2A6* SNPs, and identified an association between one of the SNPs and hearing loss among the nicotine-exposed subgroup but not among the nicotine-unexposed subgroup [19]. The second, was a PheWAS of a genetic score for CYP2A6 activity in UK Biobank [20]. However, they limited their analyses to a total of 1,029 disease endpoints, based on ICD-9 and ICD-10 diagnostic rubrics. They found associations with lung cancer and other known smoking related diseases; no associations were seen among their subsets of former or never smokers.

Our aim was to use the UK Biobank data (N = 343,662), without limiting ourselves to any specific category of phenotypes, to assess how the NMR is associated with the phenome. This is the largest PheWAS on the NMR to date, encompassing over 21,000 outcome variables. Our study is also the first to explicitly focus on the NMR, rather than solely on CYP2A6 activity.

We created a genetic score for the NMR from ten putatively causal SNPs, explaining 33.8 % of the variance in the NMR. We used the sofware package PHESANT [21] which enabled us to incorporate all variable types (continuous, binary, categorical, ordinal) in the PheWAS, and thus uncover novel associations. Importantly, we used the GxE MR-pheWAS approach [22], meaning that we ran the PheWAS also separately for ever and never smokers. The approach permitted us to distinguish whether the associations reflected a causal pathway through smoking (effect only seen in ever smokers), or some other pathways (effect also/only seen in never smokers). Our findings contribute valuable information for drug development and personalized interventions for smoking cessation and nicotine addiction, as well as for example for cancer treatments, given CYP2A6's role in metabolizing various drugs, including the chemotherapeutic agents letrozole and tegafur [23, 24].

Materials and methods

Study samples

UK Biobank

UK Biobank (UKB) is a population-based prospective study on genetic and non-genetic determinants of diseases of middle and old age. The cohort consists of over 500,000 participants from the UK, aged between 37–73 at recruitment (2006–2010) [25]. The resource comprises imputed genome-wide genotype data from all participants [26], along with a comprehensive range of phenotypic data. This includes data from clinical assessments, questionnaires, sample assays, and health record linkage —many of which are available from all of the participants (see Sudlow et al. [27]). Of note, the nicotine metabolite ratio is not available in the UKB data.

We had access to genetic data from 487,235 individuals after excluding withdrawals. We restricted our sample to individuals of self-reported White British ancestry, and performed further quality control measures, resulting in a final sample size of 343,662 individuals. During the quality control process, we ensured the genetic sex corresponded to the reported sex, and that there were no instances of sex aneuploidy. We also verified that our subset did not include any samples that were outliers with respect to genotype heterozygosity or missingness. Furthemore, we restricted the dataset to a subset that did not include any pairs of third-degree or closer relatives.

FINRISK

The National FINRISK Study consists of cross-sectional population-based data on chronic non-communicable diseases in Finland [28]. Data collection was conducted every five years between 1972–2012. Our previous GWAS of the NMR (n = 5,185 current smokers from 5 cohorts) included 1,405 current smokers (cotinine $\geq$ 10ng/ml) from the 2007 and 2012 FINRISK data collections [15]. These two cross-sectional studies included independent samples from 25–74 year old Finns. Cotinine and 3'hydroxycotinine concentrations were acquired from blood plasma samples using gas chromatography-mass spectrometry. Self-reported variables on smoking behaviour, including the number of factory and self-rolled cigarettes smoked per day (CPD), were obtained from surveys. Genome-wide genotype data were imputed using a Finnish reference panel (see Buchwald et al. [15], Table S2).

Young Finns Study

The Young Finns Study (YFS) is a prospective population-based study of cardiovascular risk factors from childhood to adulthood. The initial cross-sectional sample from 1980 was selected so as to be representative of Finnish children aged 3, 6, 9, 12, 15 and 18. The individuals have been followed up at regular intervals (see Raitakari et al. [29]). Our previous GWAS of the NMR included 714 current smokers from the YFS [15]. For each individual we had chosen the time point for which the NMR, sex, age and BMI were available, cotinine was $\geq$ 10ng/ml, and for which we had the least amount of missing values concerning the other GWAS variables (CPD, Pack years, alcohol use). Whenever there were ties, the most recent time point had been chosen. Cotinine and 3'hydroxycotinine concentrations were acquired from blood plasma samples using liquid chromatography-tandem mass spectrometry, and self-reported smoking behaviour variables including CPD were obtained from surveys. Genome-wide genotype data was imputed using the Haplotype Reference Consortium reference panel (see Buchwald et al. [15], Table S2).

Measures

The Genetic score for the nicotine metabolite ratio

Using the UKB data we created a genetic score (GS) for the NMR. We created it so that higher values reflect faster nicotine metabolism, and thus, also refer to it in this paper as the genetic score for faster nicotine metabolism. The GS served as the independent variable (exposure) in our PheWAS. We constructed the GS as the weighted sum of the 10 putatively causal SNPs highlighted by our FINEMAP analyses (see below). These SNPs together explained 33.8 % of the variance in the NMR. Prior to running the PheWAS, we standardized the GS (zGS) by subtracting the mean and dividing by the standard deviation for a simpler interpretation of the results.

Outcome variables

Of the 5,559 UKB phenotype fields available to us, we included 4,546 in our PheWAS. The excluded phenotypes were either not listed in the variable information file of the PheWAS software package PHESANT version 1.1 [21] (252 dropped), were categorized as auxiliary variables in the UKB data (STRATA=Auxiliary) (further 398 dropped) or based on the grouping done by Gibson et al. in their insomnia PheWAS [30] (further 141 dropped), or were marked for exclusion in the PHESANT variable information file (further 60 dropped). When a phenotype was measured at multiple time points, PHESANT automatically considered only the first occurrence. For phenotypes

measured multiple times at the first occurrence to improve accuracy, PHESANT used the mean value of those measurements. For multiple choice questions where individuals were able to tick all relevant options, PHESANT formed multiple binary variables. Out of the 4,546 phenotype fields, PHESANT created 21,094 outcome variables in total, meaning that our initial PheWAS consisted of 21,094 regression analyses. Each outcome variable served as the dependent variable in its respective regression model.

Covariates

We adjusted our PheWAS for sex and age to increase statistical power and to adjust for potential confounding. We also adjusted for the first ten principal components of genetic structure to control for confounding due to population stratification. We did not include any additional covariates to avoid inducing collider bias.

Ever-Never status

We wanted to perform the PheWAS separately for ever and never smokers in order to distinguish whether the associations were independent of smoking or indicative of a possible causal pathway through smoking, an approach described by Millard et al. [22]. To reduce noise, we excluded experimenters and occasional smokers from our Ever and Never subsets. We created a new Ever-Never variable based on two existing variables:

1. Current tobacco smoking ("Do you smoke tobacco now?", UKB field 1239)
2. Past tobacco smoking ("In the past, how often have you smoked tobacco?", UKB field 1249)

The second question had been asked from all except for those who indicated they currently smoke on "most or all days" in response to the first question. We classified all who answered "most or all days" to either of the two questions as ever smokers. Those who answered "no" to the first question and "never" to the second question, were classified as never smokers. All others were assigned a missing value in our Ever–Never variable, and excluded from the Ever and Never subsets. This included individuals who answered "Occasionally" or "Prefer not to answer" to either question, as well as those who responded with "Tried once or twice" to the second question.

Out of our full sample (n = 343,662), we classified 135,890 as never smokers and 110,348 as ever smokers. Of those who dropped out, 99 % had answered "Occasionally" or "Tried once or twice". This group had marginally higher GS values as compared to the Never group (Table S1 & Fig. S1). The GS distribution of the Ever group did not differ from the Never group (Table S1 & Fig. S1).

Statistical analyses

Forming the GS

We previously performed a GWAS meta-analysis of the NMR in European ancestry current smokers (n = 5,185) [15]. Our main GWAS model included sex, age, BMI, alcohol use (g/week) and birthyear as covariates. We identified two NMR loci, one on chromosome 4 and one on chromosome 19. We then fine-mapped these loci using FINEMAP [31] and the datasets YFS and FINRISK (n = 2,119) to identify the putatively causal SNPs. Our analyses suggested that there is one causal SNP on chromosome 4, explaining 1.7 % of the variance in the NMR, and 13 putatively causal SNPs on chromosome 19 explaining 36.7 % of the variance in the NMR.

In the current study, we wanted to use the SNPs and weights obtained by the earlier FINEMAP analyses to construct the GS for the NMR in the UKB data set.

Unfortunately, two of the 13 SNPs from chromosome 19 did not pass our quality control in the UKB dataset (both had a Hardy-Weinberg Equilibrium $p < 10^{-6}$). Excluding these two SNPs would have dropped the variance explained in the NMR by the chromosome 19 SNPs from 36.7 % to 23.7 %. Therefore, in order to gain maximal power, we reran the FINEMAP (version 1.4) analyses for the chromosome 19 locus as described previously (see Buchwald et al. [15]) using only SNPs that were available and passed quality control in UKB.

In short, we used a subset of the GWAS summary statistics and SNP correlation data from our previous FINEMAP analyses which included SNPs within the ±2.5Mb flanking region of the top associating SNP (rs56113850). We included the SNPs passing the following quality control criteria in both the Finnish and the UKB data: imputation info score > 0.7, call rate > 0.9, and Hardy-Weinberg Equilibrium $p > 10^{-6}$. Additionally, only the SNPs with minor allele frequency (MAF) > 1 % in the Finnish data were included and the multiallelic SNPs were excluded. Altogether, 10,133 SNPs were included in the FINEMAP analysis as opposed to the 12,060 SNPs included previously. The new FINEMAP analysis of the chromosome 19 locus resulted in nine SNPs in the top configuration of putatively causal SNPs, explaining 32.1 % of the variance in the NMR (Table S2). All the aforementioned estimates of the variance of the NMR explained, are the configuration-specific heritability estimates obtained by FINEMAP.

We then calculated the GS as a weighted sum of these nine chromosome 19 SNPs and a single chromosome 4 SNP, resulting in a GS explaining 33.8 % of the variance in the NMR (Table 1). We used the joint effect size estimates from FINEMAP as the weights for the chromosome 19 SNPs. FINEMAP was not able to distinguish which of the three top SNPs in the chromosome 4 locus was most likely to be the causal one [15]. All three SNPs had the same p-value and were perfectly correlated. For the GS of the NMR, we chose to use the SNP located in the middle based on base pair position, and used the effect size from our earlier GWAS meta-analysis as its weight.

To assess the performance and reliability of our GS, we plotted it against the NMR top SNP (rs56113850) in the UKB data. In our previous study, the top SNP alone explained 23 % of the variance in the NMR (see Buchwald et al [15], Table S5). Among the current smokers, we also plotted our GS against cigarettes smoked per day (CPD) which has been shown to have a positive association with the NMR [8, 9]. Additionally, we tested the association of the GS with CPD with a linear model. We did this by first regressing out sex, age and the first ten principal components of genetic structure, and then inverse normalizing CPD. We used the whole data but also split the data into two: lower and higher ends of the GS. We used 0 as the cut-off point based on the bend apparent in the loess curve of the scatter plot (Fig. 1) and to obtain roughly the same sized groups as the mean is close to the median (Table 2). In order to have a reference point, using our Finnish data, we plotted the NMR against CPD, as well as against a biomarker for nicotine intake, the sum of the nicotine metabolites cotinine (Cot) and 3'hydroxycotinine (3HC) (Cot+3HC).

PheWAS

We ran the PheWAS of the GS for the NMR using PHESANT (version 1.1). We began by running PHESANT on the full sample using the 'save' option in order to save the variables processed and derived by PHESANT. We adjusted for age, sex and the first ten principal components of genetic structure. Out of the 4,546 UKB phenotype fields we had to begin with, PHESANT created a total of 21,094 outcome variables. We then performed the actual PheWAS analyses in two stages.

Exploratory PheWAS. In the first stage, we used the 21,094 PHESANT derived

outcome variables to run the PheWAS for the entire sample (All), and separately for those who had ever smoked daily (Ever) and those who had never smoked (Never) to obtain our initial PheWAS results. For each of these three exploratory PheWAS runs (All/Ever/Never), we used the 5 % false discovery rate (FDR) level to define statistical significance. We used the Benjamini-Hochberg method to obtain the cut-off points for the 5 % FDR. Within each PheWAS, the results were ranked based on ascending p-values. Next, for each outcome variable, we calculated the critical value $c_i = 0.05 \times i/n$, where i is the rank and n the total number of tests performed in that PheWAS. The cut-off point was then defined as the c_i corresponding to the outcome with the highest rank to satisfy the condition $p_i \leq c_i$ [32]. For each PheWAS (All/Ever/Never), we then used this cut-off value as the threshold for statistical significance also in the second stage.

Final PheWAS. In the second stage of our analyses we took all of the outcomes reaching statistical significance in any of the three PheWAS described above (All/Ever/Never). We then manually checked the coding and the distributions of these highlighted variables, as well as the appropriateness of the models used by PHESANT. Where appropriate, we split the variable into multiple variables, adjusted the coding or used a different regression model. We did these changes to obtain results that are more readily interpretable. Additionally, we reran the highlighted linear regression analyses by regressing the covariates out before inverse rank transforming the outcome variables to normality. We did this because many of these continuous outcomes showed normal distributions centred on different values depending on sex (one of our covariates). Otherwise, we used the same protocol as PHESANT for our final analyses. A detailed description of the second stage outcomes that we derived, recoded or analysed using a different model from the initial PheWAS can be found in Table S3.

Differences between the ever and never smokers

We wanted to assess whether there were any outcome variables for which there was a statistically significant difference between the effect sizes for the Ever and Never subsets. We did this in R by deriving a p-value using the function pchisq(x,df=1,lower=FALSE), where

$$x = \frac{(\hat{\beta}_{\text{ever}} - \hat{\beta}_{\text{never}})^2}{\text{se}^2_{\text{ever}} + \text{se}^2_{\text{never}}}, \text{ and } x \sim \chi^2_1 \text{ under the null hypothesis of } \beta_{\text{ever}} = \beta_{\text{never}}.$$

First, we did this for our initial PheWAS results. Of all the 21,094 PHESANT derived outcomes altogether 15,212 had been analysed for both the Ever and Never subsets using the same regression model. We used the 5 % FDR threshold to determine statistical significance. Once again, we identified the cut-off value using the Benjamini-Hochberg method. We then repeated this assessment for our outcomes from the second stage of our PheWAS analyses, using the same cut-off value as in our initial phenome-wide analysis to define statistical significance.

Follow-up analyses

Current versus Former analysis. We followed-up our Ever versus Never analyses by taking all the outcomes reaching a statistically significant difference in their effect sizes, and repeating the analysis for current versus former smokers. We did this to obtain a better understanding of the possible causal role of smoking in the association pathway. We constructed the current and former smoker subgroups from our Ever subset, and repeated the second stage PheWAS regression analyses for the chosen outcomes in these two groups.

Follow-up of the GS-Cessation association. Our PheWAS results suggested a positive association between the GS for faster nicotine metabolism and smoking cessation. To investigate further what could be driving the positive association, we examined the variable "Why did you stop smoking? (You can select more than one answer): Illness or ill health / Doctor's advice / Health precaution / Financial reasons" (UKB Field 6157), and the variable "Number of unsuccessful stop-smoking attempts" (UKB Field 2926) which captures the difficulty of quitting.

As faster nicotine metabolism is thought to associate with a greater amount of cigarettes smoked and more difficulties in quitting smoking successfully (reviewed in [10]), we began by rerunning the logistic regression model for cessation, including cigarettes smoked per day (CPD) as an additional covariate to adjust for possible confounding. Smoking more will increase the risk of severe illnesses and the need for treatments such as surgery that require abstinence. Thus, we also ran this model within a subset of the ever smokers from whom we had excluded those who listed illness or doctor's advice as the reason for stopping smoking.

We divided the Ever subset into three approximately equal sized groups based on the GS tertiles, and named these groups as slow, medium and fast metabolizers. We then compared, one reason at a time, whether the proportions differed between the slow and fast groups. We used the 2-sample test for equality of proportions using the prop.test function in R.

To assess the importance of the different reasons, we reran the cessation model among those ever smokers for whom we had the reason for stopping available. This subset consisted of all ever smokers who had at least once managed to stop for over six months. Among this subset, we added all four reasons to the model. We wanted to see how the reasons predicted former status at the time of the questionnaire.

Then, among our former smokers, we ran a negative binomial regression model with "Number of unsuccessful stop-smoking attempts" as the dependent variable and each of the four reasons, the standardized GS, age, sex and the first ten principal components of genetic structure as the predictor variables. We also ran the model including CPD as a covariate.

MRBase and FinnGen PheWASs using the top SNP for the NMR. For comparison, we ran PheWASs of the top SNP for the NMR (rs56113850) using GWAS summary data from FinnGen (Data freeze 9, from April 2022) and MRBase (Database version 0.3.0, from 25 Oct 2020). We used the open access web interfaces. The FinnGen Data Freeze 9 has a total sample size of 377,277 and consists of 2,272 disease endpoints from Finnish biobank participants [33, 34]. The MRBase comprises GWAS summary data from a large number of consortia, also including data from both UKB and the FinnGen study [35, 36]. From MRBase, we got results for 39,105 outcomes. This group of outcomes included expression quantitative trait loci (eQTLs) as well.

In both cases, we created tables of the results reaching statistical significance at the 5 % FDR level using the Benjamini-Hochberg method. We used the library biomaRt in R to annotate the eQTLs that were highlighted in the MRBase PheWAS.

Results

Descriptive statistics

For female participants, the smoking status (Ever/Experimenter/Never) category Never was the largest (44 %), whilst for males the Ever category was the largest (38 %) (Table 2). Among both sexes, over 70 % of the ever smokers were former smokers, and roughly 70 % of the current smokers wanted to stop smoking. The distribution of the GS for the

NMR was nearly identical for both sexes (Table 2). When running a linear model of the standardized GS by age, the age variable was not statistically significant ($p = 0.909$).

The correlation between the GS and the top SNP (rs56113850) was 0.73 (Fig. S2). Each standard deviation increase in the GS is associated with an increase of 0.6 cigarettes smoked per day (CPD), after adjusting for age, sex and the first ten genetic principal components (Table S4). The association seems to be stronger at the lower end of the GS than at the higher end (Table S4, Fig. 1a). This trend can also be seen in our Finnish sample when plotting NMR against CPD or Cot+3HC (Fig. 1b).

Exploratory PheWAS

A total of 61 different outcomes reached statistical significance across the initial All, Ever, and/or Never PheWASs (Fig. S4, Table S5). As PHESANT created 21,094 outcome variables out of the 4,546 UKB phenotype fields, our initial PheWAS for the full sample consisted of 21,094 tests. The initial full sample PheWAS resulted in 47 outcomes reaching statistical significance, and the threshold for significance was $0.05 \cdot 47/21094 = 1.1e-04$ (Table S5a). For the Ever subset there were 29 outcomes that reached statistical significance at the threshold of $0.05 \cdot 29/16648 = 8.7e-05$ (Table S5b). For the Never subset, only two outcomes were highlighted, and the threshold was at $0.05 \cdot 2/16103 = 6.2e-06$ (Table S5c).

Phenome-wide comparison of ever and never smokers

A total of 15,212 outcome variables had been analysed for both the Ever and Never subsets using the same regression model. Of these, only two variables showed a statistically significant difference at the 5 % FDR level in their effect sizes between the ever and never smokers (Table S5d). The threshold for significance was thus at $0.05 \cdot 2/15212 = 6.6e-06$. Both of these were lung capacity variables and were included among the variables highlighted in the All and Ever PheWAS results (Fig. S4). There was no association among the Never subset while among the Ever and All groups higher values of the GS predicted worse lung capacity.

Final PheWAS

The set of 61 outcome variables highlighted in the initial PheWAS was then taken to the second stage of our PheWAS analysis. Some of these variables were split into multiple variables in order to tease out which aspect of the phenotype was driving the association (Table S3). For example, Smoking Status (Never/Previous/Current) was split into Cessation (former vs current smoker) and Initiation (ever vs never smoker), out of which only Cessation showed a statistically significant association when we reran the analyses (Fig. 2).

Our second stage of the PheWAS analysis included 71 outcome variables (Table S6). Of these, 57 showed a statistically significant association in at least one of the three data sets based on the data specific 5 % FDR thresholds (Fig. 2, Fig. 3). We had 45 outcomes reaching statistical significance among the full sample, 26 among the Ever subset and three among the Never subset (Fig. 2, Table S6b–d).

For 14 outcomes, we did not observe a statistically significant association with the GS for faster nicotine metabolism in any of the three groups (All, Ever or Never). This set of non-significant associations included five continuous variables that had reached the 5 % FDR level of significance in at least one of the data sets in the previous stage. Of these, at least Waist circumference and the volume of Ventral Anterior (VA) nucleus of the thalamus (left hemisphere) demonstrated notable differences in their distribution by sex based on visual inspection of the histograms (not shown). The rest of the

non-significant results included the derived Smoking Initiation and Occasional vs Daily Smoker variables, as well as some of the derived tea and coffee variables.

Faster nicotine metabolism predicts worse lung capacity: causal pathway through smoking

Of the 71 outcome variables included in the second stage of our PheWAS, 58 were available for our Ever vs Never analysis. Of these, four showed a statistically significant difference between ever and never smokers in their effect sizes (Table S6a, Fig. 3). All four were lung capacity measures. For all four, there was no evidence of association among never smokers but a statistically significant association among ever smokers, suggesting a causal pathway through smoking. The results suggested that only among ever smokers higher values of the GS predicted worse lung capacity.

Never smokers: associations with liver enzymes and tea quantity

Among never smokers, a higher GS for faster nicotine metabolism was associated with decreased alkaline phosphatase and increased alanine aminotransferase, both of which are liver enzymes (Table S6d). These associations were also seen among the Ever and All groups (Fig. 2, Fig. 3), and were statistically significant for all three groups already in our initial PheWAS (Fig. S4).

Among never smokers, there was also an association with increased tea consumption. This was statistically significant in the All group too, but not among the ever smokers subset. Apart from these three outcomes, none of the other outcomes showed a statistically significant association in the never smokers subset. The direction and magnitude of association was nevertheless similar to those of the other groups for many variables such as for the other blood and urine assay variables and variables related to coffee (Fig. 3).

Ever smokers: associations with liver enzymes and smoking related variables

Among ever smokers, 26 of the final second stage outcome variables showed statistically significant associations (Fig. 2, Table S6c). In addition to the liver enzymes, a higher GS for faster nicotine metabolism was associated with a greater number of cigarettes smoked per day, an increased likelihood of having quit smoking, decreased lung capacity, an increased likelihood of taking simvastatin (cholesterol medicine), being more of a morning person than an evening person, quitting smoking at a younger age, an increased likelihood of having dentures, increased smoking compared to 10 years ago, decreased consumption of cheese, a decreased likelihood of being a worrier, decreased liking for cigarette smoking, an increased likelihood of leg artery operations, and a decreased likelihood of an anastomosis of ileum to anus.

The GS for faster nicotine metabolism was also associated with an increased likelihood of smoking less at the time of the interview ("Do you smoke tobacco now?: No / Only occasionally / Yes, on most or all days", UKB field 1239). This negative association with current smoking was driven by the higher odds of quitting. There was no association for the derived Occasional vs Daily outcome, whereas our derived Cessation outcome demonstrated one of the strongest associations (Table S6c). Our ever smokers only contained individuals who smoked daily or had smoked daily in the past. Therefore, we had better statistical power for our Occasional vs Daily phenotype in our full sample. Nevertheless, we saw no association there either ($p = 0.90$, Table S6b).

Full sample: strongest associations with liver enzymes, smoking related variables, coffee and tea

Among the full sample there were 45 outcome variables that showed an association with the GS for the NMR at the 5 % FDR level (Fig. 2,Table S6b). The strongest associations were observed for the liver enzymes, cigarettes smoked per day, cessation, and coffee and tea consumption. The GS was associated with 16 biomarkers, often used to assess overall health. For example, the GS was associated with outcomes related to fatty acids, blood sugar, kidney and liver health, impedance of arm, calcium, and white blood cells. The Never and Ever associations for these biomarkers were in line with each other but neither reached statistical significance at the 5 % FDR level, except for the liver enzymes mentioned earlier (Fig. 3).

Some additional outcome variables, that had not reached statistical significance in the smaller subsets, were highlighted in the full sample. For example, among the full set, the GS for faster nicotine metabolism was associated with an increased liking for dark chocolate and coffee without sugar, decreased father's age at death, decreased maternal smoking around birth, increased volumes of grey matter in the supracalcarine cortex, and a decreased likelihood of using Serenace (haloperidol), an anti-psychotic medication. Some outcomes were only analysed in some groups due to a lack of cases, and the use of haloperidol was one of these (only analysed in the full sample). The outcome Past tobacco smoking (Never/Tried once or twice/Occasionally/Most days or daily) was only analysed in the All group as it required answers from both ever and never smokers. The GS for faster nicotine metabolism was associated with an increased likelihood of being in a higher past tobacco smoking category.

Follow-up analyses

Comparison of current and former smokers supports the causal role of smoking on worse lung functioning. Our Ever vs Never analyses highlighted four lung capacity measures, suggesting a causal pathway through smoking, as the association was only seen in the ever smokers. We followed-up these results by rerunning the analyses for the current and former smoker subsets. For all four measures of lung capacity, the associations were attenuated among the former smokers, giving further support for the possible causal role of smoking on worse lung functioning (Fig. 4, Table S6e).

GS-Cessation association possibly explained by CPD and ill health. In our sensitivity analyses of cessation, the odds ratio of the GS remained positive after adjusting for CPD (Table S7a–b), although it did decrease from 1.07 to 1.05. The odds ratio decreased further to 1.01 and was no longer statistically significant ($p = 0.54$) once we excluded individuals who had stopped due to 'Illness or ill health' and/or 'Doctor's advice' (Table S7c).

When comparing the reasons for stopping smoking, fast metabolizers of nicotine (high GS) seemed to have more incentive to quit than slow metabolizers (low GS), though differences were small (Table S8). Regardless of reason (llness or ill health / Doctor's advice / Health precaution / Financial reasons), fast metabolizers had a slightly higher percentage of people selecting it than slow metabolizers. For the more prevalent reasons: 'Health precaution' and 'Financial reasons', the differences were statistically significant at the 0.05 level (Table S8).

We took a closer look at ever smokers who had managed to quit at least once for over six months, to see whether any of the reasons for stopping smoking were associated with remaining a former smoker still at the time of the questionnaire (Table S7d–e). Listing 'Financial reasons' was positively associated with remaining a former smoker (OR = 1.21, $p = 2.18e - 05$), while 'Doctor's advice' had a negative association (OR =

0.61, $p = 5.09e - 15$) (Table S7d). The GS for faster nicotine metabolism showed a positive association (OR = 1.04, $p = 0.0264$). After adding CPD to the model, the direction of the association remained the same for all reasons, as well as for the GS (Table S7e). Now the significant reasons were 'Doctor's advice' (OR = 0.51, $p = 6.94e - 24$), 'Illness or ill health' (OR = 0.83, $p = 0.0022$), and 'Health precaution' (OR = 1.08, $p = 0.0462$). For every additional cigarette the odds of remaining a former smoker was 6 % greater (OR = 1.06, $p = 3.37e - 122$).

Among former smokers of the Ever subset, the median value of unsuccessful smoking attempts was 2.0, the mean was 2.9 and the standard deviation was 7.1. Once we excluded those listing 'Illness or ill health' and/or 'Doctor's advice', the numbers hardly changed: 2.0, 2.8, and 6.9.

All four reasons (UKB participants could choose as many as they wanted) were statistically significant at the 0.05 level in explaining the number of unsuccessful stop-smoking attempts in the former smokers subset (Table S9). They were all associated with an increased number of attempts, with 'Health precaution' demonstrating the strongest association. Individuals listing 'Health precaution' had a 39 % higher number of attempts compared to those who did not list it (IRR = 1.39, $p = 9.42e - 221$) (Table S9a). Additionally, for every standard deviation increase in the GS for the NMR, there was an expected 1 % increase in the number of attempts (IRR = 1.012, $p = 0.02$). However, after including CPD in the model, the GS was not significant anymore ($p = 0.24$) (Table S9b). With every additional cigarette, the number of unsuccessful stop-smoking attempts increased by 1.6 % (IRR = 1.016, $p = 5.07e - 233$).

PheWAS in MRBase and FinnGen confirm our findings and highlight lung cancer and lipid outcomes. The NMR top SNP in our previous Meta GWAS of ever smokers with European ancestry was rs56113850 on chromosome 19 [15]. Based on our META GWAS, the beta for the major allele C (vs. T) was 0.682 (se = 0.02) (C allelle frequency was 0.553). Both the FinnGen and MRBase PheWAS results have been presented for the allele C so that the results can be interpreted for genetically determined faster nicotine metabolism.

Our FinnGen PheWAS had two outcomes reaching statistical significance at the 5 % FDR level (Table S10a) –these were both lung cancer diagnoses. The associations were positive, so genetically determined faster nicotine metabolism was associated with a greater odds of having lung cancer. The MRBase database pinpointed 199 outcomes as having a significant association at the 5 % FDR level (Table S10b). However, many of these outcomes were overlapping or were highly similar (Table S10c). Most of the top associations were related to smoking, lung cancer, lung functioning, liver enzyme levels, cholesterol and other lipids, as well as gene expression (See annotations for highlighted genes: Table S10d). Our own PheWAS of the UKB had not included some of these outcomes such as those related to gene expression, but otherwise the MRBase results closely mirrored our own UKB PheWAS.

Discussion

In our study we used a hypothesis-free method, GxE MR-PheWAS [22], to identify novel associations across the phenome with the NMR. Importantly, we conducted separate PheWAS analyses for ever and never smokers. This enabled us to identify which associations might be attributable to a causal effect through smoking. We explored over 21,000 outcome variables, making our PheWAS the most comprehensive to date on the rate of nicotine metabolism. Notably, our study was the first PheWAS to explicitly focus on the NMR itself, rather than solely on CYP2A6 activity.

We found associations with several smoking related traits and diseases. These were in line with existing literature, with the exception of a positive association between our GS

for faster nicotine metabolism and smoking cessation. Additionally, our study unveiled novel associations with measures not previously reported to associate with the rate of nicotine metabolism, including liver enzymes, lipids, and consumption of coffee and tea.

We did not replicate the hearing loss finding identified by the initial PheWAS of CYP2A6 activity [19]. The other PheWAS published to date on CYP2A6 activity, used the UKB similar to ours, but focused only on about 1,000 disease endpoints [20]. They did not replicate the hearing loss finding either. Their findings, limited to tobacco related diseases, were in line with ours, although none of their findings reached statistical significance in our study which included the whole phenome. Their top finding, lung cancer, ranked 56th in significance in our subset of ever smokers, for which 29 outcomes reached statistical significance after correcting for a 5 % false discovery rate. In our secondary analyses, however, we conducted a PheWAS replication study of the top GWAS SNP for the NMR using the FinnGen data. This FinnGen PheWAS consisted of roughly 2,000 disease endpoints, and here, the only two outcomes reaching FDR significance were indeed lung cancer diagnoses.

Faster nicotine metabolism is known to associate with increased smoking [8, 9]. Our PheWAS results align with this established relationship. The GS for faster nicotine metabolism was associated with a greater number of cigarettes smoked per day, increased smoking compared to ten years ago, and greater smoking in the past (inquired from all except daily current smokers).

We found both the GS and the NMR to have a non-linear association with CPD in the UKB dataset as well as our Finnish dataset (Fig. 1). This non-linear relationship, indicative of a plateau-effect on CPD, aligns with previous findings for cotinine [37], where a plateau-effect was observed on cotinine when modelling it against CPD. However, to our knowledge, this phenomenon has not been previously reported for the NMR. It is possible that, as our population-based data relied on voluntary participation from invitees, some subgroups of individuals with ill health, such as those with high GS and high CPD values, were underrepresented. More research conducted in a controlled setting would be warranted to explore this association more comprehensively.

While clinical trials and longitudinal data have shown that faster nicotine metabolism predicts lower quit rates [11, 12], our PheWAS yielded contrasting results. The GS was associated with greater odds of quitting, and quitting at a younger age (Fig. 3). To further investigate the relationship between the GS and smoking cessation, we reran the cessation analyses adding CPD as a covariate to the model. We speculated whether CPD might be acting as a possible mediator for the association. Faster nicotine metabolism is known to be positively associated with CPD. Amount smoked, in turn, impacts one's health and possibly one's inclination to quit due to the realization of adverse health effects or the necessity of abstinence, for example, for consequent surgical procedures. After adding CPD into the model, the association weakened but remained positive (Table S7a–b). Once we additionally excluded individuals who listed ill health and/or doctor's advice as the reason for quitting, the association was no longer statistically significant (Table S7c). We observed similar findings in our initial GWAS of the NMR where we had modelled cessation with a GS for faster nicotine metabolism first in the entire dataset and then in a subset where individuals were excluded based on their disease status [14].

When we looked at the reasons for quitting more closely, it did appear that fast metabolizers might have had slightly more incentive to quit (Table S8). Additionally, it looked like internal reasons (health precaution and financial reasons) were more likely to result in continued abstinence (remaining a former smoker at the time of the UKB questionnaire) (Table S7d–e). Conversely, having quit because of doctor's advice seemed to decrease the likelihood of remaining a former smoker at the time of the questionnaire.

As mentioned, those with faster nicotine metabolism are thought to have a harder time succeeding at quitting [10]. Among former smokers, the GS had a positive association with the number of unsuccessful quit attempts but the association was no longer statistically significant once we adjusted for CPD (Table S9). All four reasons increased the number of attempts, as did CPD. Health precaution and ill health however, had considerably stronger associations than doctor's advice and financial reasons.

Our results suggest that several factors are at play when it comes to cessation. Individuals with faster nicotine metabolism may smoke more, spend more money on tobacco, and experience more health problems. They may thus be more motivated to quit but succeeding can possibly be more challenging due to greater nicotine dependence which has been shown to correlate with faster nicotine metabolism by some studies [38,39]. Regardless of the difficulties, faster metabolizers of nicotine appear more likely to quit long-term (remain former smokers) than slower metabolizers of nicotine. This is likely due to their heightened motivation and more frequent attempts to quit.

In summary, our findings suggest that health precautionary reasons and higher daily cigarette consumption make quitting more challenging but also make it more likely for the individual to remain a former smoker if they do manage to quit. We believe that CPD and ill health might partially confound or mediate the positive association between the GS and cessation. Additionally, due to the nature of the UKB dataset, selection bias may be also contributing to this positive association. The UKB dataset is enriched with healthier individuals [40], and thus those who smoke more and have not quit may be underrepresented, potentially affecting the results.

Our Ever vs Never analyses highlighted four lung capacity measures, providing evidence for a causal pathway through smoking, as the associations were only seen in the ever smokers (Table S6a, Fig. 3). When we divided the ever smokers into subsets of former and current smokers, the effect sizes for all four lung capacity measures were smaller among former smokers compared to current smokers, though they remained statistically significant (Fig. 4, Table S6e). Our results align with existing research. Smoking is known to accelerate the decline in lung function, which may develop into chronic obstructive pulmonary disease [41, 42]. It is worth noting that, although smoking cessation has been shown to slow down this rate of decline, it does not appear to fully revert the rate of decline to that of never smokers [43].

Beyond variables related to smoking, our study also identified several other variables to be associated with our GS for the NMR. It is noteworthy, that for many variables our findings did not appear to depend on an association pathway through smoking. Interesting associations, where we saw no apparent differences between ever and never smokers, included liver enzymes, lipid measures and coffee consumption.

In our PheWAS, only two phenotypes, the liver enzymes alkaline phosphatase (ALP) and alanine aminotransferase (ALT), showed associations with the GS that were statistically significant in both ever and never smokers (Fig. 3). These associations ranked among the top findings in all three groups (All, Ever, Never), as well as in the PheWAS results from MRBase (Table S6b–d, Table S10b).

Both ALP and ALT are blood serum liver enzymes that are key biomarkers for assessing the extent and cause of liver damage [44–46]. One cause of liver damage is alcohol use, and alcohol use and smoking are well known to be associated behaviours. However, the strengths of the associations did not differ between the ever and never smokers, implying that the associations are not mediated by smoking or alcohol.

Building upon previous PheWAS analyses of these liver enzymes, conducted by Liu et al. [44], our study suggests that a higher GS for faster nicotine metabolism is correlated with less favourable liver enzyme levels with respect to associated diseases. We observed a negative association between the GS and ALP levels. The ALP PheWAS

by Liu et al., revealed that lower values of the genetically determined ALP were associated with an increased odds of hypercholesterolemia, pulmonary heart disease, as well as phlebitis and thrombophlebitis of lower extremities.

On the other hand, we found a positive association between the GS and ALT levels. For genetically determined ALT levels, Liu et al. found 16 associations in their ALT PheWAS. The strongest associations, based on p-values, highlighted a clear trend: higher values of genetically determined ALT levels were consistently associated with a higher risk of hepatic diseases such as hepatitis, primary liver cancer and non-alcoholic cirrhosis. Additionally, genetically determined ALT was positively associated with Type 2 Diabetes, possibly due to the role ALT plays in insulin resistance [44]. In our PheWAS of the entire sample, we identified a statistically significant positive association between the GS and glycated haemoglobin, suggesting that higher GS values correlate with poorer blood sugar control. The effect size remained consistent across the Ever and Never subsets but did not reach statistical significance in ether subset. Of the 16 associations reported by Liu et al. for ALT, only three showed an opposite direction of effect: dementias, fracture of hand and wrist, and corneal degenerations. However, all three had p-values that had only just surpassed the false discovery rate threshold, and for dementias, the direction of the effect varied depending on the Mendelian Randomization analysis method they used.

A substantial portion of the MRBase PheWAS results concerned outcome variables related to cholesterol, fatty acids, and other lipids (Table S10c). Likewise, other outcomes indicative of cardiovascular health were also highlighted. For instance, genetically determined faster nicotine metabolism (the NMR top SNP rs56113850 allele C count), was associated with elevated levels of cholesterol, triglycerides, LDL levels, and a higher apolipoprotein B/apolipoprotein A ratio (indicative of poorer cardiovascular health), as well as an increased waist circumference, higher BMI and a greater odds of using cholesterol medication. Of note, cholesterol medication was among the top five associations in the PheWAS of the FinnGen data, based on p-values, although it did not reach the FDR of 5 % significance level.

In our UKB PheWAS, the GS was associated with decreased apolipoprotein A, increased linoleic acid, a higher omega-6 to omega-3 ratio and a decreased omega-3 to total fatty acids ratio (Fig. 3). Again, the direction of association was such that higher GS values predicted worse lipid values with respect to cardiovascular health (see e.g. [47]). These results were in line with the findings from the MRBase PheWAS. However, they were only statistically significant in our All group. Nevertheless, there was no apparent difference between the effect size estimates among the ever and never smokers, alluding to a shared genetic component between nicotine metabolism and lipid levels. Interestingly, only among ever smokers, we observed a positive association between the GS and the use of cholesterol medication, leg artery bypass surgery, and endarterectomy of femoral artery (Fig. 3). This suggests that ever smokers with higher GS values may face a compounded risk for cardiovascular diseases due to their smoking history and genetics, and may thus be more likely to face adverse health effects and require medical interventions.

Moving on to another noteworthy association highlighted in our PheWAS, the GS for faster nicotine metabolism was associated with increased coffee and tea consumption (Fig. 3). We observed statistically significant associations with coffee consumption in the All group and with tea consumption in the All and Never groups. The concordant results between Ever and Never subsets suggest an association pathway distinct from smoking.

When decaffeinated coffee was included in the coffee quantity variable, there was a slight attenuation of associations, suggesting that the observed effect is possibly driven by caffeine metabolism. Notably, in a GWAS of caffeine metabolites, multiple SNPs at

19q13.2, including the *CYP2* cluster, *NUMBL*, *ADCK4*, *MIA* and *EGLN2*, showed genome wide significant associations [48]. The CYP2A6 enzyme is known to play a minor role in caffeine metabolism [48], and it is possible that our results reflect this connection.

Interestingly, our PheWAS also highlighted a positive correlation between the GS and preferences for dark chocolate and coffee without sugar. These associations were statistically significant only in the All subset. However, both the Ever and Never results were consistent with those of the All subset. Once again, this would imply that the association pathway is unlikely to involve smoking. Like coffee and tea, dark chocolate contains caffeine and tastes bitter. Therefore, these findings suggest that the association with coffee and tea consumption could be mediated through preference for bitter taste in addition to caffeine metabolism. Cornelis and van Dam (2021) found in their study that genetically inferred bitter taste perception indeed did play a role in coffee and tea drinking behaviour but to a lesser extent than genetically inferred caffeine sensitivity [49]. They also found support for conditioned taste preferences where individuals learn to associate the bitter taste with either beneficial or adverse physiological effects of caffeine.

While we did not observe direct evidence of a pathway involving smoking, it is well known that smoking and coffee consumption correlate [50]. The prevailing belief is that tobacco smoke increases an individual's caffeine metabolism and that is why smokers require more caffeine to achieve the desired effect [51]. Once an individual stops smoking, their caffeine metabolism eventually returns to normal. A previous Mendelian randomization study found support for this relationship [52]. They compared current, former and never smokers using a single SNP as a biomarker for CPD [52]. They found no association with coffee consumption among never smokers or former smokers in their meta-analysis of three datasets but did find an association among current smokers.

Although our GS for faster nicotine metabolism positively correlated with CPD, we saw only a slight increase in effect size estimates between the GS and the coffee and tea consumption measures among ever smokers compared to never smokers. It is possible that we would have seen stronger evidence for an additional pathway through smoking had we examined current and former smokers separately.

Our results align with the work of Cornelis et al., suggesting some shared genetics between nicotine metabolism and caffeine metabolism, along with potential connections to bitter taste perception. Both smoking and caffeine impact the metabolism of some medications and affect optimal dosages [51]. Additionally, it has been proposed that smokers attempting to quit, may inadvertently confuse symptoms of caffeine toxicity with nicotine withdrawal symptoms if they fail to adjust their coffee consumption [50]. Thus, understanding the interactions between smoking and caffeine consumption is of great interest.

Conclusions

In summary, our results suggest that faster nicotine metabolism is associated not only with smoking related traits but also with various other health related characteristics. Our GS for faster nicotine metabolism was associated with worse liver enzyme and lipid values with respect to associated diseases, as well as increased coffee and tea consumption. Importantly, we saw no evidence of a causal pathway through smoking for these associations. Leaning on the assumption that slow metabolizers of nicotine have an easier time quitting, our findings are promising. They imply that drugs designed to convert fast metabolizers into slower metabolizers may not involve adverse side effects but instead could offer therapeutic benefits for other health issues as well. Future research, focusing on these newly highlighted variables using different data and study

designs, involving different sources of potential bias, is warranted to confirm and extend these findings.

Availability of data and materials

Access to the UK Biobank data can be applied at http://ukbiobank.ac.uk/register-apply/. To apply for access to the YFS data see: http://youngfinnsstudy.utu.fi/, and for FINRISK see: https://thl.fi/en/web/thl-biobank. All analyses were performed in R and bash. Scripts will be deposited on GitHub upon publication of the manuscript.

Tables and figures

Table 1 The ten SNPs and their weights used for calculating the GS for the NMR.

Table 2 Descriptive statistics. Descriptive statistics for age, sex, the GS for faster nicotine metabolism, and various smoking variables in the UK Biobank data.

Fig. 1 **A** Loess curve depicting the association between the standardized genetic score (zGS) for the Nicotine Metabolite Ratio (NMR) and cigarettes smoked per day (CPD) in UK Biobank (UKB). **B** Loess curves depicting the associations between the standardized NMR (zNMR) and CPD, as well as zNMR and Cotinine + 3-Hydroxycotinine (Cot + 3HC), a biomarker for nicotine intake, using the combined data from FINRISK and Young Finns Study (YFS). The x-axis is restricted to show only data points within two standard deviations from the mean to highlight the main trends. Full data are shown in Fig S3. Note: The standardized variables (zGS and zNMR) were calculated by subtracting the mean and dividing by the standard deviation.

Fig. 2 Venn diagram of the results of the final 71 variables. Out of these variables, 57 reached the FDR of 5 % significance level in at least one of the data sets (All / Ever /Never). Those with a statistically significant (FDR of 5 %) difference in their effect sizes between the ever and never smokers have been indicated with a star and bolded text.

Fig. 3 Results of the 71 variables from the stage 3 analyses. The results of those variables (n = 57) that were statistically significant at the 5 % FDR level (solid circles) in at least one of the data sets (All / Ever / Never) are shown. Figure continues on next page. * = There was a statistically significant (FDR of 5 %) difference between the effect sizes of the ever and never smokers, n = normalised after covariates had first been regressed out, d = derived from the original UKB phenotype, c = coding corrected to be more intuitive.

Fig. 4 Current versus Former smokers comparison for the four lung capacity variables highlighted in the Ever versus Never smoker analysis. For both Current and Former subsets all four variables had a statistically significant association with the GS at the 5 % FDR level which had been counted from the PheWAS of the ever smokers. Differences between the subsets were not significant at the 5 % FDR level counted from the phenome-wide Ever versus Never analysis. Nevertheless, the association is visibly attenuated for the former smokers for all four variables. n = normalised after covariates had first been regressed out.

Supporting information

Fig. S1 Boxplots of the standardised GS for the NMR in UKB by subgroups. Experimenters includes those individuals who answered "Occasionally" or "Tried once or twice" to the questions on current and past smoking behaviour.

Fig. S2 Scatterplot of the imputed genotype dosage (allele C) at the chromosome 19 top SNP (rs56113850) for the NMR against the GS for faster nicotine metabolism in UKB.

Fig. S3 Scatterplots and loess curves presenting the full data of the association between A the standardized genetic score for the NMR (zGS) and cigarettes smoked per day (CPD) in UKB, **B** the standardized NMR (zNMR) and CPD in the Finnish data, **C** the NMR and CPD in the Finnish data, and **D** the zNMR and Cotinine + 3-Hydroxycotinine (Cot + 3HC), a biomarker for nicotine intake, in the Finnish data. Note: The standardized variables (zGS and zNMR) were calculated by subtracting the mean and dividing by the standard deviation.

Fig. S4 Venn diagram of the 61 variables highlighted in our initial PheWAS. The figure shows the 61 variables that were statistically significant at the 5 % FDR level in at least one of the data sets (All / Ever / Never). The variable Smoking status has been listed twice in the figure as it contained a different amount of categories for the All and Ever groups.

Table S1 GS distribution by subset of smoking status. sd = standard deviation. The p-values are from Mann-Whitney U tests comparing the GS distributions to the Never group.

Table S2 FINEMAP top configuration of causal SNPs for the NMR on the chromosome 19 locus in the Finnish data. FINEMAP results of the 5Mb region centred at the top SNP when only including SNPs passing quality control in UKB (n(SNP) = 10,133). The most probable configuration consisted of 9 SNPs (depicted in the table) and their heritability estimate was 32.1% (95% CI: 28.5–35.6%). FINEMAP gave a regional heritability estimate of 33.8% (95% CI: 30.0–37.8%) and suggested that there are 7–11 causal SNPs within the region. SNP, single-nucleotide polymorphism; BP, base pair position in GRCh37 coordinates; EA/NEA, the effect allele/ the non-effect allele; MINOR, the less common allele; MAF, minor allele frequency in the Finnish dataset (n = 2,119) used for the FINEMAP analysis; SNP PROB, posterior probability of being a causal SNP; BETAJ, effect estimate from the joint model including all these 9 SNPs (reported for the effect allele); SEJ, standard error for the BETAJ; P-VALUEJ, p-value for the BETAJ.

Table S3 Variable descriptions for PheWAS outcomes. A Variable information on all the third stage outcomes that we derived, recoded or analysed using a different model from our initial PheWAS (See Table S3b). **B** Variable information on all 61 variables highlighted in our initial PheWAS analyses.

Table S4 Linear regression beta coefficient for the zGS when explaining CPD by zGS in UK Biobank. Analysis was done for the current smokers subset of UK Biobank. CPD has been adjusted for sex, age and the first 10 genetic principal components, and then inverse normalized (rank-based-inverse-normal-transformation) before regressing it on zGS. The standard deviation of the adjusted CPD was 8.3.

When looking at all individuals, each standard deviation increase in the GS is associated with a 0.07516 standard deviation increase in CPD, i.e. a 0.6 increase in cigarettes smoked per day. The regression was ran for three subsets of the data including either all individuals (ALL), only those with lower zGS values (LOW GS; zGS < 0) or only those with higher zGS values (HIGH GS; zGS ≥ 0).

Table S5 Results from the initial PheWAS. A-C PheWAS results for the variables that were significant at the 5 % FDR level for the All, Ever and Never data, respectively. **D** Ever versus Never analysis results for the variables that had a statistically significant difference at the 5 % FDR level between their effect sizes for the Ever and Never subsets.

Table S6 Results from the final stage of the PheWAS. A Ever versus Never analysis results for the variables that had a statistically significant difference at the 5 % FDR level between their effect sizes for the Ever and Never subsets. **B-D** PheWAS results for the variables that were significant at the 5 % FDR level for the All, Ever and Never data, respectively. **E** Current versus Former analysis results for the four lung capacity measures that had been highlighted in the Ever versus Never analysis (See Table S6a).

Table S7 Odds ratios from the logistic regression model for cessation. A Among Ever smokers (n = 110348), model including the standardized GS, sex (1=Male, 0=Female), age and the first 10 genetic principal components as the independent variables. **B** Among Ever smokers (n = 110348), model including CPD as an additional independent variable. **C** Among subset of Ever smokers who did not stop smoking due to Illness or Doctor's advice (n = 70278), model including all the same variables as in B. **D** Among subset of Ever smokers who had at least once managed to quit for over 6 months (n = 83704), model including the 4 reasons for stopping smoking. **E** Among subset of Ever smokers who had at least once managed to quit for over 6 months (n = 79228), model including the 4 reasons for stopping smoking and CPD.

Table S8 Reason for stopping smoking by nicotine metabolism group (grouped based on the tertiles of the GS for the NMR). Ever smokers who had stopped smoking for over 6 months during the time they smoked were asked: "Why did you stop smoking? (You can select more than one answer)" (UK Field 6157). Total number of ever smokers included in this analysis was N = 110 348, of which 26 % were current smokers and 74 % former smokers. * Statistically significant difference ($p < 0.05$, 2-sample test for equality of proportions) between the Slow and Fast groups.

Table S9 Incident Rate Ratios (IRRs) from the negative binomial regression model for Number of unsuccessful stop-smoking attempts. A Among former smokers, model including standardized GS, the four reasons (1 = Yes, 0 = No), sex (1 = Male, 0 = Female), age and the first 10 genetic principal. **B** Same as A but including CPD as a covariate in the model.

Table S10 The FDR significant results from the PheWAS of the top NMR SNP, rs56113850 (allele C), using FinnGen and MRBase. A Using the FinnGen r9 there were two outcomes reaching statistical significance at the 0.05 FDR level. Results for these two outcomes are presented as they were obtained from the web interface with the exception of deleting the "m" from in front of "af_case" and "af_control" and deleting the empty "pip" column to avoid confusion. mlogp=-log10(p), Bhcritical = Benjamini-Hochberg critical value. Beta has been counted for the C allele.

Allele C frequency (vs T) was 56 % – 57 % across all phenotypes. **B** Using the MRBase (Database version: 0.3.0 from 25 October 2020) there were 199 outcomes reaching statistical significance at the 0.05 FDR level. Results for these outcomes are presented as they were obtained from the web interface. minuslogp = -log10(p), Bhcritical = Benjamini-Hochberg critical value, bf = significant at the 0.05 bonferroni level (T = TRUE, F = FALSE). **C** Same as B but ordered by trait. **D** Annotations for the 14 gene expression outcomes that were among the 199 MRBase results. Annotations were obtained from the Ensembl database using BiomaRt in R. band = Karyotype band, gene_biotype = Gene type , hgnc_symbol = HGNC (The HUGO Gene Nomenclature Committee) gene symbol. Note: MRBase id column can be searched as the GWAS ID here: https://gwas.mrcieu.ac.uk/datasets/ for more information on the study.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 22627. We thank Dr. Samuel Jones for organizing the UK Biobank data files. We want to express our gratitude to all participants in UK Biobank, YFS, FINRISK and FinnGen as well as to all people, institutes and funding bodies that have been involved in the infrastructure surrounding these cohorts, and made these valuable resources possible.

Funding

This study has been funded by the Doctoral Program in Population Health and the Doctoral School in Health Sciences from the University of Helsinki, as well as Yrjö Jahnsson Foundation to JB, by the Academy of Finland (grants 338507, 336825 and 352795) and Sigrid Juselius Foundation to MP and JK, and by the Academy of Finland Center of Excellence in Complex Disease Genetics (grants 336823, 352792), and the Finnish Foundation for Cardiovascular Research to JK. VS was supported by the Juho Vainio Foundation. The FINRISK surveys have been mainly funded from budgetary funds of THL. Important additional funding has been obtained from the Academy of Finland and from several domestic foundations. The YFS has been financially supported by the following sources: Academy of Finland (grant numbers 356405, 322098, 286284, 134309 (Eye), 126925, 121584, 124282, 255381, 256474, 283115, 319060, 320297, 314389, 338395, 330809, and 104821, 129378 (Salve), 117797 (Gendi), and 141071 (Skidi)); the Social Insurance Institution of Finland; Competitive State Research Financing of the Expert Responsibility Area of Kuopio, Tampere and Turku University Hospitals (grant X51001); the Juho Vainio Foundation; the Paavo Nurmi Foundation; the Finnish Foundation for Cardiovascular Research; the Finnish Cultural Foundation; the Sigrid Juselius Foundation; the Tampere Tuberculosis Foundation; the Emil Aaltonen Foundation; the Yrjö Jahnsson Foundation; the Signe and Ane Gyllenberg Foundation; the Diabetes Research Foundation of Finnish Diabetes Association; EU Horizon 2020 (grant 755320 for TAXINOMISIS and grant 848146 for To Aition); the European Research Council (grant 742927 for MULTIEPIGEN project); the Tampere University Hospital Supporting Foundation, Finnish Society of Clinical Chemistry; the Cancer Foundation Finland; pBETTER4U_EU (Preventing obesity through Biologically and bEhaviorally Tailored inTERventions for you; project number: 101080117); and the Jane and Aatos Erkko Foundation.

Author contributions

Jadwiga Buchwald: Conceived and designed the study; acquired funding; determined the statistical analysis methodology; curated data; conducted the analysis; interpreted and visualized the data; wrote the manuscript. Matti Pirinen: Conceived and designed the study; acquired funding; determined the statistical analysis methodology; curated data; interpreted the data; supervised the project; critically reviewed and revised the manuscript. Jaakko Kaprio: Conceived and designed the study; acquired funding; critically reviewed and revised the manuscript. Terho Lehtimäki: Acquired data. Olli Raitakari: Acquired data. Veikko Salomaa: Acquired data. All authors read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Ethics approval and consent to participate

UKB was approved by the North West Multi-center Research Ethics Committee. The Coordinating Ethics Committee of the Helsinki and Uusimaa Hospital District approved the FINRISK surveys (decision number 229/E0/06 in 2007 and 162/13/03/00/11 in 2012). The local University Hospital ethics committees in Turku and Tampere approved the YFS study. All participants provided written informed consent.

Competing interests

The authors declare that they have no competing interests.

References

1. Murray CJ, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The lancet. 2020;396(10258):1223–1249.

2. Drope J, Hamill S, Chaloupka F, Guerrero C, Lee H, Mirza M, et al. The Tobacco Atlas, Seventh edition. Vital Strategies and Tobacconomics; 2022.

3. WHO. WHO global report on trends in prevalence of tobacco use 2000–2025, Fourth edition. World Health Organization; 2021.

4. Benowitz NL. Pharmacology of nicotine: addiction, smoking-induced disease, and therapeutics. Annual review of pharmacology and toxicology. 2009;49:57–71.

5. Yuan M, Cross SJ, Loughlin SE, Leslie FM. Nicotine and the adolescent brain. The Journal of physiology. 2015;593(16):3397–3412.

6. Stolerman IP, Jarvis M. The scientific case that nicotine is addictive. Psychopharmacology. 1995;117:2–10.

7. Dempsey D, Tutka P, Jacob III P, Allen F, Schoedel K, Tyndale RF, et al. Nicotine metabolite ratio as an index of cytochrome P450 2A6 metabolic activity. Clinical Pharmacology & Therapeutics. 2004;76(1):64–72.

8. Benowitz NL, Pomerleau OF, Pomerleau CS, Jacob III P. Nicotine metabolite ratio as a predictor of cigarette consumption. Nicotine & Tobacco Research. 2003;5(5):621–624.

9. Schoedel KA, Hoffmann EB, Rao Y, Sellers EM, Tyndale RF. Ethnic variation in CYP2A6 and association of genetically slow nicotine metabolism and smoking in adult Caucasians. Pharmacogenetics and Genomics. 2004;14(9):615–626.

10. Chenoweth MJ, Tyndale RF. Pharmacogenetic optimization of smoking cessation treatment. Trends in pharmacological sciences. 2017;38(1):55–66.

11. Patterson F, Schnoll R, Wileyto E, Pinto A, Epstein L, Shields P, et al. Toward personalized therapy for smoking cessation: a randomized placebo-controlled trial of bupropion. Clinical Pharmacology & Therapeutics. 2008;84(3):320–325.

12. Chenoweth MJ, O'Loughlin J, Sylvestre MP, Tyndale RF. CYP2A6 slow nicotine metabolism is associated with increased quitting by adolescent smokers. Pharmacogenetics and genomics. 2013;23(4):232.

13. Lerman C, Schnoll RA, Hawk LW, Cinciripini P, George TP, Wileyto EP, et al. Use of the nicotine metabolite ratio as a genetically informed biomarker of response to nicotine patch or varenicline for smoking cessation: a randomised, double-blind placebo-controlled trial. The lancet Respiratory medicine. 2015;3(2):131–138.

14. Loukola A, Buchwald J, Gupta R, Palviainen T, Hällfors J, Tikkanen E, et al. A genome-wide association study of a biomarker of nicotine metabolism. PLoS genetics. 2015;11(9):e1005498.

15. Buchwald J, Chenoweth MJ, Palviainen T, Zhu G, Benner C, Gordon S, et al. Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent. Molecular psychiatry. 2021;26(6):2212–2223.

16. Hukkanen J, Jacob P, Benowitz NL. Metabolism and disposition kinetics of nicotine. Pharmacological reviews. 2005;57(1):79–115.

17. Chenoweth MJ, Novalen M, Hawk Jr LW, Schnoll RA, George TP, Cinciripini PM, et al. Known and novel sources of variability in the nicotine metabolite ratio in a large sample of treatment-seeking smokers. Cancer Epidemiology, Biomarkers & Prevention. 2014;23(9):1773–1782.

18. Denny JC, Bastarache L, Roden DM. Phenome-wide association studies as a tool to advance precision medicine. Annual review of genomics and human genetics. 2016;17:353–373.

19. Polimanti R, Jensen KP, Gelernter J. Phenome-wide association study for CYP2A6 alleles: rs113288603 is associated with hearing loss symptoms in elderly smokers. Scientific Reports. 2017;7(1):1–7.

20. Giratallah H, Chenoweth MJ, Pouget JG, El-Boraie A, Alsaafin A, Lerman C, et al. CYP2A6 associates with respiratory disease risk and younger age of diagnosis: a phenome-wide association Mendelian Randomization study. Human Molecular Genetics. 2023; p. ddad172.

21. Millard LA, Davies NM, Gaunt TR, Davey Smith G, Tilling K. PHESANT: a tool for performing automated phenome scans in UK Biobank. BioRxiv. 2017; p. 111500.

22. Millard LA, Munafò MR, Tilling K, Wootton RE, Davey Smith G. MR-pheWAS with stratification and interaction: searching for the causal effects of smoking heaviness identified an effect on facial aging. PLoS Genetics. 2019;15(10):e1008353.

23. Yamamiya I, Yoshisue K, Ishii Y, Yamada H, Chiba M. Effect of CYP2A6 genetic polymorphism on the metabolic conversion of tegafur to 5-fluorouracil and its enantioselectivity. Drug Metabolism and Disposition. 2014;42(9):1485–1492.

24. Murai K, Yamazaki H, Nakagawa K, Kawai R, Kamataki T. Deactivation of anti-cancer drug letrozole to a carbinol metabolite by polymorphic cytochrome P450 2A6 in human liver microsomes. Xenobiotica. 2009;39(11):795–802.

25. Biobank Showcase. Field 21022: [Variable description]; n.d. Website. Available from: `https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=21022`.

26. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203–209.

27. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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 medicine. 2015;12(3):e1001779.

28. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T, et al. Forty-year trends in cardiovascular risk factors in Finland. The European Journal of Public Health. 2015;25(3):539–546.

29. Raitakari OT, Juonala M, Rönnemaa T, Keltikangas-Järvinen L, Räsänen L, Pietikäinen M, et al. Cohort profile: the cardiovascular risk in Young Finns Study. International journal of epidemiology. 2008;37(6):1220–1226.

30. Gibson MJ, Lawlor DA, Millard LA. Identifying the potential causal role of insomnia symptoms on 11,409 health-related outcomes: a phenome-wide Mendelian randomisation analysis in UK Biobank. BMC medicine. 2023;21(1):1–15.

31. Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics. 2016;32(10):1493–1501.

32. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

33. FinnGen. FinnGen Documentation of R9 release; 2023. Website (accessed July 11, 2023). Available from: `https://www.finngen.fi/en/access_results`.

34. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. Nature. 2023;613(7944):508–518.

35. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. elife. 2018;7:e34408.

36. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure. BioRxiv. 2020; p. 2020–08.

37. Hovanec J, Weiß T, Koch HM, Pesch B, Behrens T, Kendzia B, et al. Smoking intensity and urinary nicotine metabolites by socioeconomic status in the Heinz Nixdorf Recall study. BMC Public Health. 2022;22(1):1–10.

38. 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. Journal of the National Cancer Institute. 2011;103(17):1342–1346.

39. Sofuoglu M, Herman AI, Nadim H, Jatlow P. Rapid nicotine clearance is associated with greater reward and heart rate increases from intravenous nicotine. Neuropsychopharmacology. 2012;37(6):1509–1516.

40. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American journal of epidemiology. 2017;186(9):1026–1034.

41. Fletcher C, Peto R. The natural history of chronic airflow obstruction. Br Med J. 1977;1(6077):1645–1648.

42. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary. American journal of respiratory and critical care medicine. 2017;195(5):557–582.

43. Oelsner EC, Balte PP, Bhatt SP, Cassano PA, Couper D, Folsom AR, et al. Lung function decline in former smokers and low-intensity current smokers: a secondary data analysis of the NHLBI Pooled Cohorts Study. The Lancet Respiratory Medicine. 2020;8(1):34–44.

44. Liu Z, Suo C, Jiang Y, Zhao R, Zhang T, Jin L, et al. Phenome-Wide Association Analysis Reveals Novel Links Between Genetically Determined Levels of Liver Enzymes and Disease Phenotypes. Phenomics. 2022;2(5):295–311.

45. Kalas MA, Chavez L, Leon M, Taweesedt PT, Surani S. Abnormal liver enzymes: A review for clinicians. World journal of hepatology. 2021;13(11):1688.

46. Lala V, Zubair M, Minter DA. Liver function tests. In: StatPearls [internet]. StatPearls Publishing; 2022.

47. Julkunen H, Cichońska A, Tiainen M, Koskela H, Nybo K, Mäkelä V, et al. Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. Nature Communications. 2023;14(1):604.

48. Cornelis MC, Kacprowski T, Menni C, Gustafsson S, Pivin E, Adamski J, et al. Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. Human molecular genetics. 2016;25(24):5472–5482.

49. Cornelis MC, van Dam RM. Genetic determinants of liking and intake of coffee and other bitter foods and beverages. Scientific reports. 2021;11(1):23845.

50. Swanson JA, Lee JW, Hopp JW. Caffeine and nicotine: a review of their joint use and possible interactive effects in tobacco withdrawal. Addictive behaviors. 1994;19(3):229–256.

51. Zevin S, Benowitz NL. Drug interactions with tobacco smoking: an update. Clinical pharmacokinetics. 1999;36:425–438.

52. Bjørngaard JH, Nordestgaard AT, Taylor AE, Treur JL, Gabrielsen ME, Munafo MR, et al. Heavier smoking increases coffee consumption: findings from a Mendelian randomization analysis. International journal of epidemiology. 2017;46(6):1958–1967.

Table 1 The ten SNPs and their weights used for calculating the GS for the NMR

SNP	CHR	BP	EA/NEA	MINOR	MAF_{UKB}	WEIGHT (MAF)
RS36103218	4	69359253	T/C	C	0.3422	-0.175 (MAF_{META} = 0.43)
RS189621498	19	41288136	A/G	A	0.0003	0.6127 (MAF_{FY} = 0.03)
RS74719953	19	41335799	T/C	T	0.0652	-0.3581 (MAF_{FY} = 0.09)
RS34945948	19	41340842	G/A	G	0.1474	1.0276 (MAF_{FY} = 0.14)
RS12985907	19	41343544	A/G	A	0.2307	-0.8003 (MAF_{FY} = 0.28)
RS1801272	19	41354533	T/A	T	0.0271	-1.0103 (MAF_{FY} = 0.02)
RS7250713	19	41355195	C/G	G	0.4386	0.3251 (MAF_{FY} = 0.37)
RS7248187	19	41437426	C/G	G	0.3023	0.3187 (MAF_{FY} = 0.24)
RS116382863	19	41534881	T/C	T	0.1242	0.2212 (MAF_{FY} = 0.13)
RS11466310	19	41861858	T/C	T	0.0244	-0.5414 (MAF_{FY} = 0.02)

For all chromosome 19 SNPs the weight has been obtained from their joint model using FINEMAP in the Finnish data. For the chromosome 4 SNP the weight is the effect size obtained from our previous GWAS [15].

SNP, single-nucleotide polymorphism; *CHR*, chromosome; *BP*, base pair position in GRCh37 coordinates; *EA/NEA*, the effect allele/ the non-effect allele; *MINOR*, the less common allele; *MAF*, minor allele frequency; *WEIGHT*, weight used to calculate the GS (reported for the effect allele); *UKB*, UK biobank; *META,* data used in our previous GWAS meta-analysis (n = 5,185) from which the weight for the chr 4 SNP is; *FY,* FINRISK and YFS data (n = 2,119) used to obtain the weights for the chr 19 SNPs.

Table 2 Descriptive statistics for age, sex, the GS for faster nicotine metabolism, and various smoking variables in the UKB data.

		All n = 343,662	Female n = 184,565 (53.7 %)	Male n = 159,097 (46.3 %)
		mean (sd), median [min, max]	**mean (sd), median [min, max]**	**mean (sd), median [min, max]**
Age		56.9 (8), 58 [39, 72]	56.7 (7.9), 58 [40, 71]	57.1 (8.1), 59 [39, 72]
GS		1.2 (0.6), 1.2 [-1.7, 2.9]	1.2 (0.6), 1.2 [-1.7, 2.9]	1.2 (0.6), 1.2 [-1.7, 2.8]
zGS		0 (1), 0.1 [-5.2, 3.1]	0 (1), 0.1 [-5.2, 3.1]	0 (1), 0.1 [-5.2, 2.9]
		n (%)	**n (%)**	**n (%)**
Smoking status				
	Never	135,890 (39.7)	81,037 (44.1)	54,853 (34.6)
	Experimenter	96,272 (28.1)	52,700 (28.6)	43,572 (27.5)
	Ever	110,348 (32.2)	50,217 (27.3)	60,131 (38.0)
Former-Current (Ever subset)				
	Former	81,179 (73.6)	36,490 (72.7)	44,689 (74.3)
	Current	29,141 (26.4)	13,711 (27.3)	15,430 (25.7)
Wants to quit (Current subset)				
	Yes	17,876 (70.3)	8,759 (72.4)	9,117 (68.4)
	No	7,543 (29.7)	3,336 (27.6)	4,207 (31.6)
		mean (sd), median [min, max]	**mean (sd), median [min, max]**	**mean (sd), median [min, max]**
Age started smoking on most days				
	Current (n = 25,522)	17.7 (5.8), 16 [5, 69]	18.1 (5.8), 17 [5, 69]	17.4 (5.7), 16 [5, 65]
	Former (n = 80,788)	17.2 (3.6), 17 [5, 63]	17.7 (3.7), 17 [5, 60]	16.8 (3.4), 16 [5, 63]
Cigarettes smoked per day				
	Current (n = 23,682)	15.7 (8.4), 15 [1, 140]	14.2 (7.3), 15 [1, 120]	17.4 (9.2), 15 [1, 140]
	Former (n = 76,865)	19.3 (10.5), 20 [1, 140]	16.8 (8.4), 15 [1, 100]	21.5 (11.6), 20 [1, 140]
Number of unsuccessful stop-smoking attempts				
	Current (n = 2,527)	4.1 (10.1), 3 [0, 200]	4 (10.7), 3 [0, 200]	4.1 (9.7), 3 [0, 200]
	Former (n = 74,108)	2.9 (7.1), 2 [0, 200]	2.4 (5.3), 2 [0, 200]	3.2 (8.3), 2 [0, 200]
Age stopped smoking on most days				
	Current (n = 3,218)	47.3 (11.6), 48 [12, 70]	47.4 (11.5), 48 [13, 70]	47.1 (11.6), 48 [12, 70]
	Former (n = 80,859)	39.5 (11.6), 39 [9, 69]	39.2 (11.5), 38 [12, 69]	39.7 (11.6), 39 [9, 69]

GS: Genetic Score for the Nicotine Metabolite Ratio. zGS: standardized GS. Smoking status: Ever: Individuals indicating smoking on most days at present (UKB Field 1239) or in the past (UKB Field 1249), Never: Individuals responding negatively to current (UKB Field 1239) and past (UKB Field 1249) tobacco smoking, Experimenters: individuals indicating occasional or minimal past (UKB Field 1249) or present (UKB Field 1239) tobacco use. Former-Current: subsets of ever smokers differentiating those who had smoked in the past and those who smoke presently based on UKB Field 20116. Wants to quit: derived from the variable “Wants to stop smoking” (UKB Field 3496) with responses grouped into “Yes” (including “Yes, definitely” and “Yes, probably”) and “No” (including “No, probably not” and “No, definitely not”). Age started smoking on most days: UKB Field 2867 for former smokers, UKB Field 3436 for current smokers. Cigarettes smoked per day: UKB Field 3456 for current smokers, UKB Field 2887 for former smokers. Number of unsuccessful stop-smoking attempts: all individuals who had indicated that in the past they smoked tobacco on most or all days and that during the time they smoked they stopped for more than 6 months were asked “How many times did you try to give up smoking before you were successful?” (UKB Field 2926). Age stopped smoking on most days: UKB Field 2897.

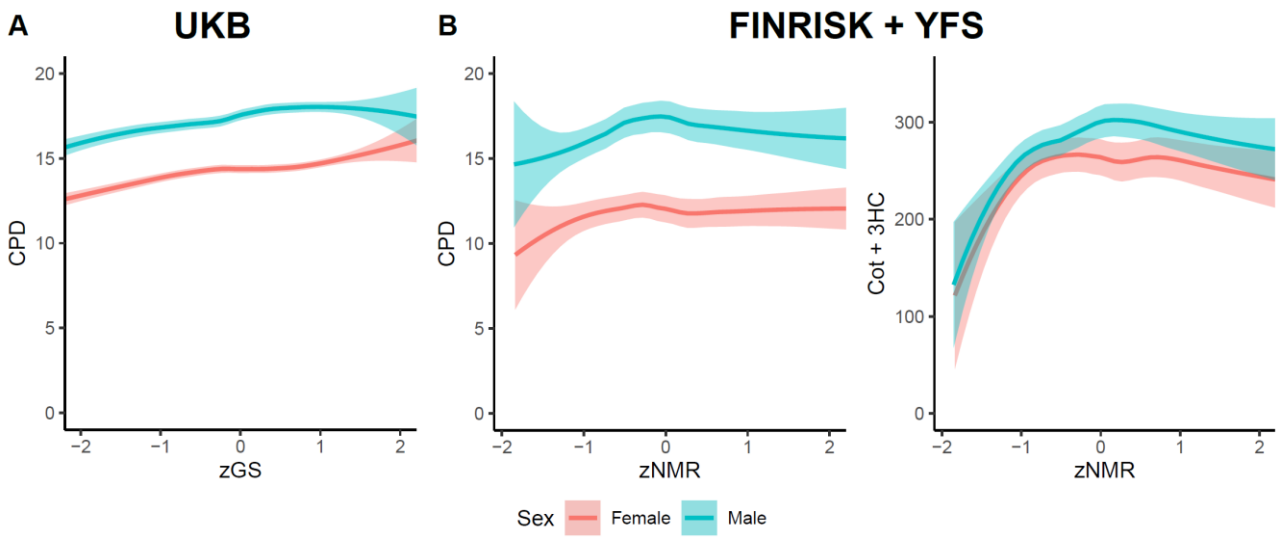

Fig. 1 A Loess curve depicting the association between the standardized genetic score (zGS) for the Nicotine Metabolite Ratio (NMR) and cigarettes smoked per day (CPD) in UK Biobank (UKB). **B** Loess curves depicting the associations between the standardized NMR (zNMR) and CPD, as well as zNMR and Cotinine + 3-Hydroxycotinine (Cot + 3HC), a biomarker for nicotine intake, using the combined data from FINRISK and Young Finns Study (YFS). The x-axis is restricted to show only data points within two standard deviations from the mean to highlight the main trends. Full data are shown in Fig S3. Note: The standardized variables (zGS and zNMR) were calculated by subtracting the mean and dividing by the standard deviation.

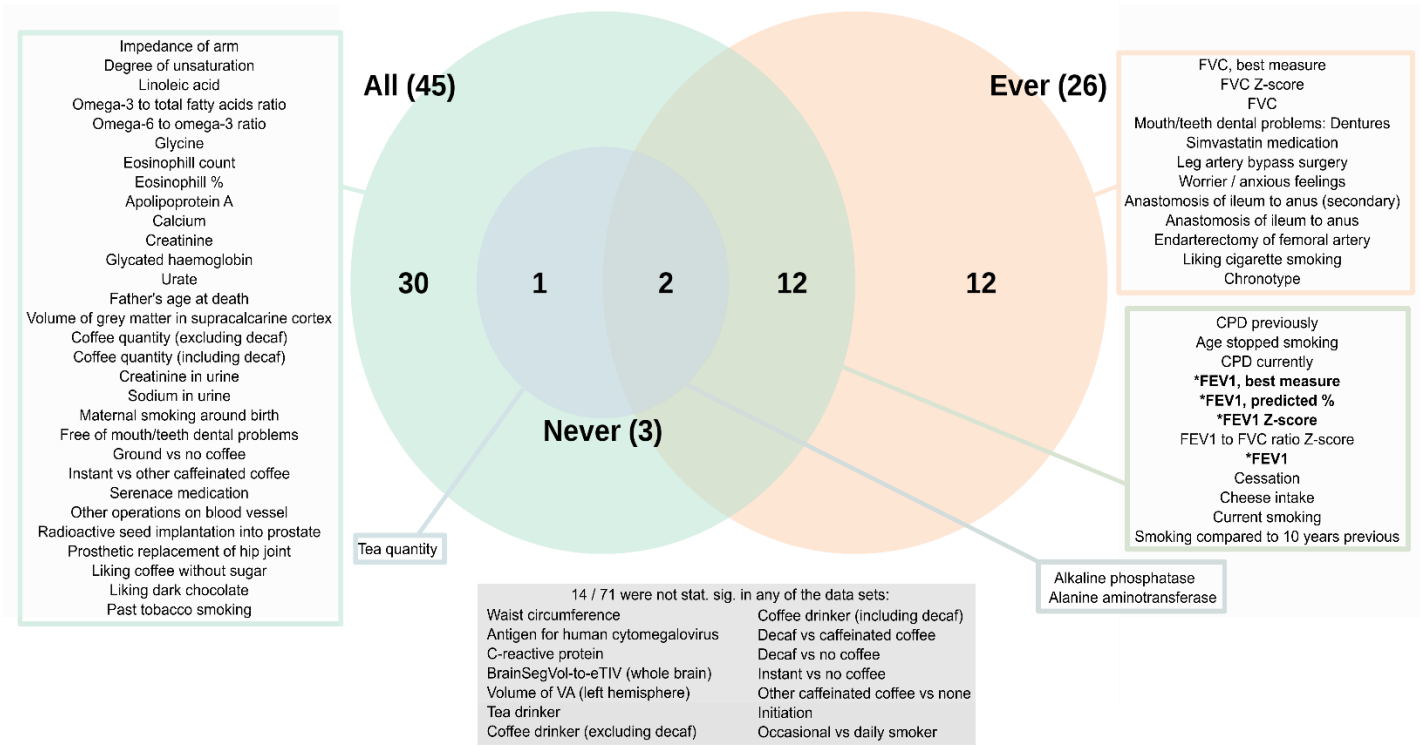

Fig. 2 Venn diagram of the results of the final 71 variables. Out of these variables, 57 reached the FDR of 5 % significance level in at least one of the data sets (All / Ever /Never). Those with a statistically significant (FDR of 5 %) difference in their effect sizes between the ever and never smokers have been indicated with a star and bolded text.

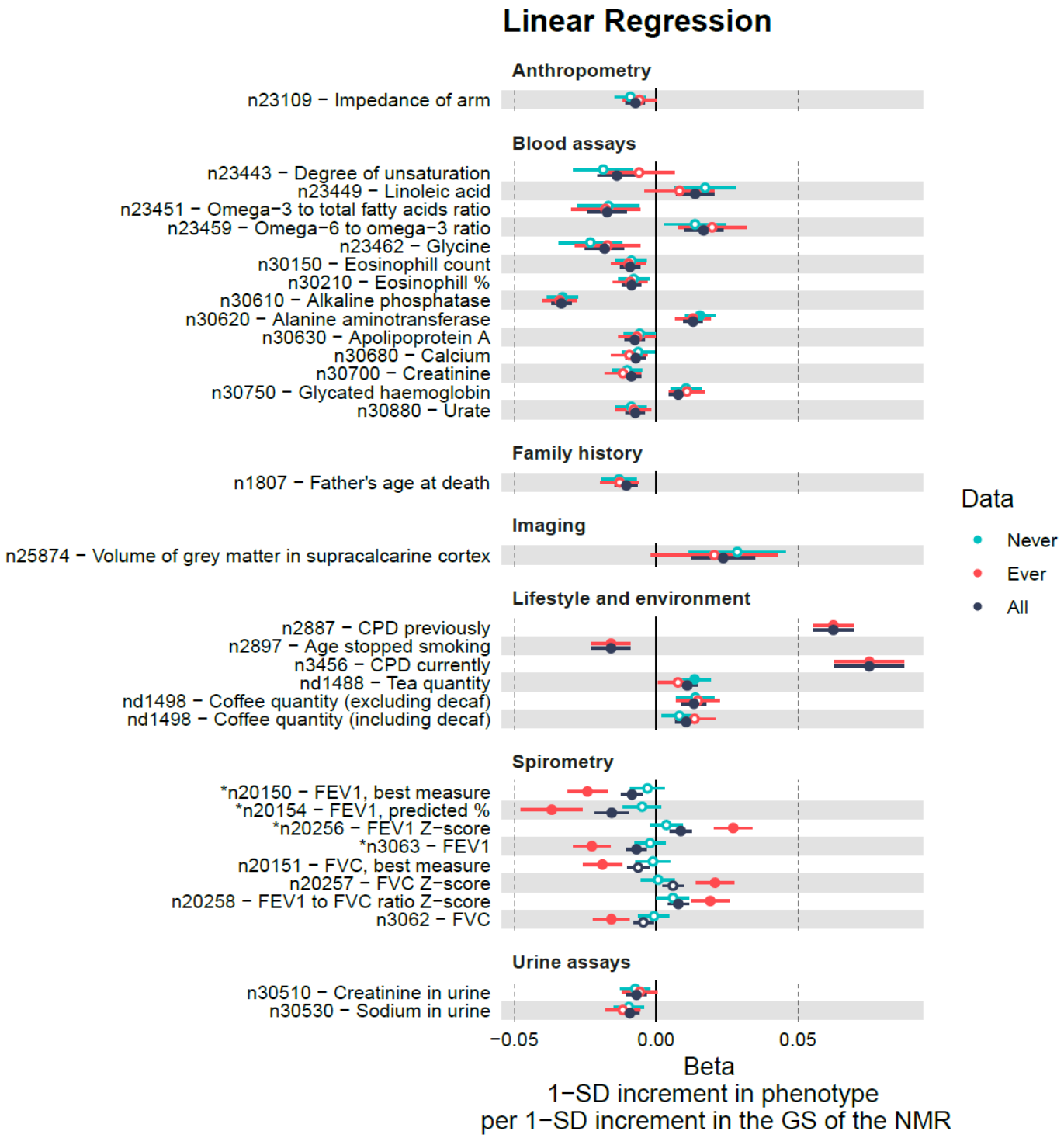

Fig. 3 Results of the 71 variables from the stage 3 analyses. The results of those variables (n = 57) that were statistically significant at the 5 % FDR level (solid circles) in at least one of the data sets (All / Ever / Never) are shown. Figure continues on next page. * = There was a statistically significant (FDR of 5 %) difference between the effect sizes of the ever and never smokers, n = normalised after covariates had first been regressed out, d = derived from the original UKB phenotype, c = coding corrected to be more intuitive.

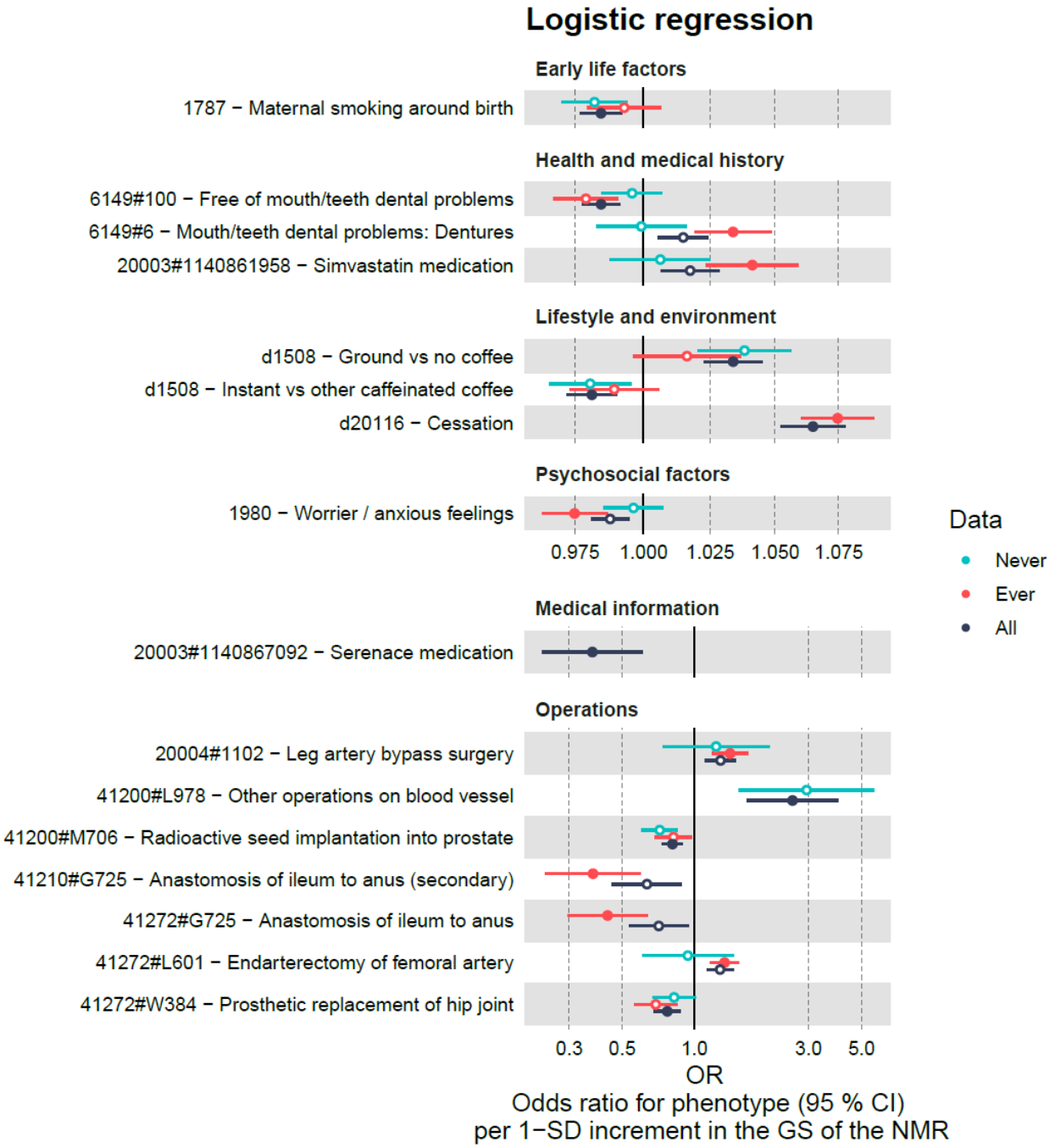

Fig. 3 continued

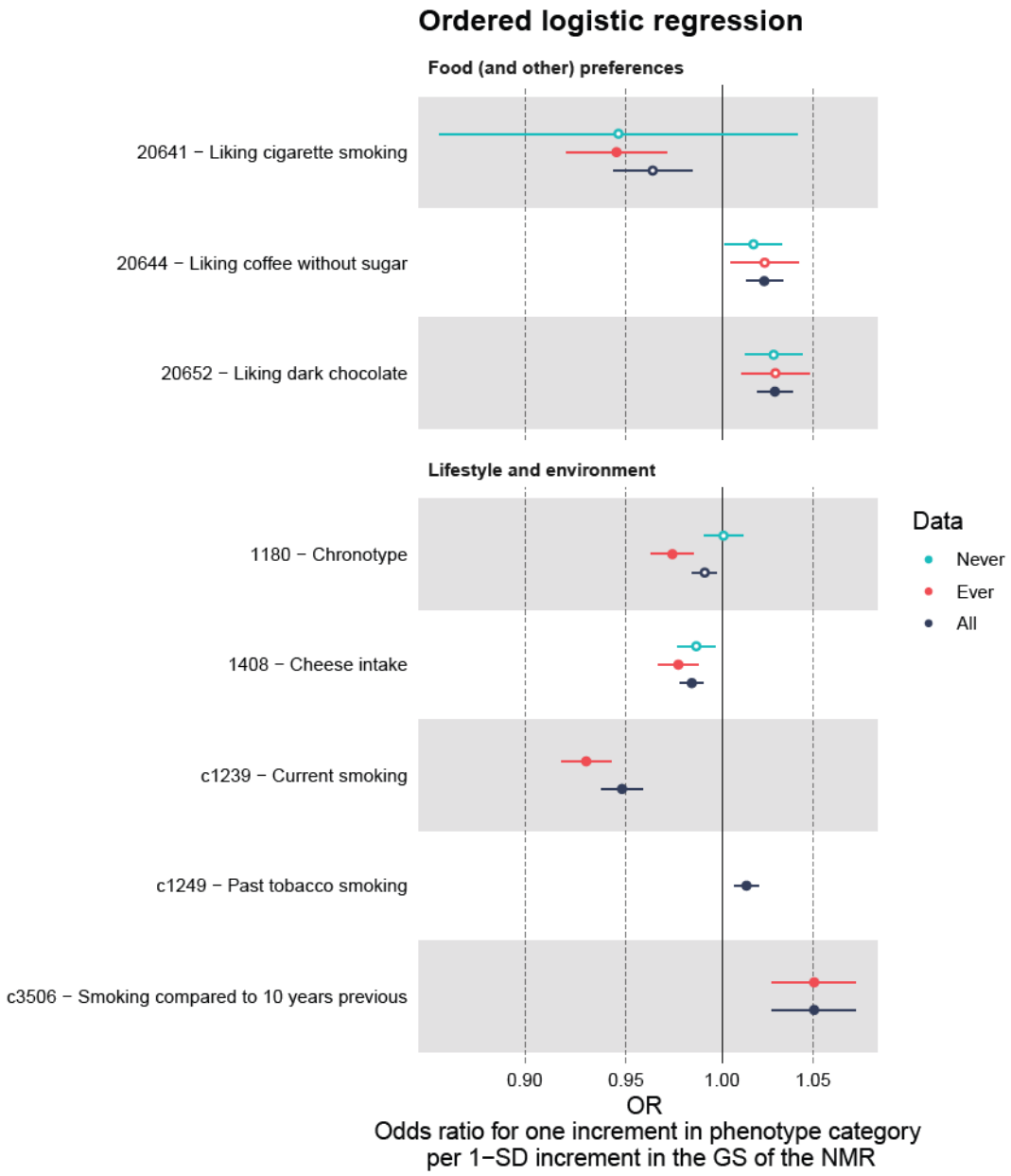

Fig. 3 continued

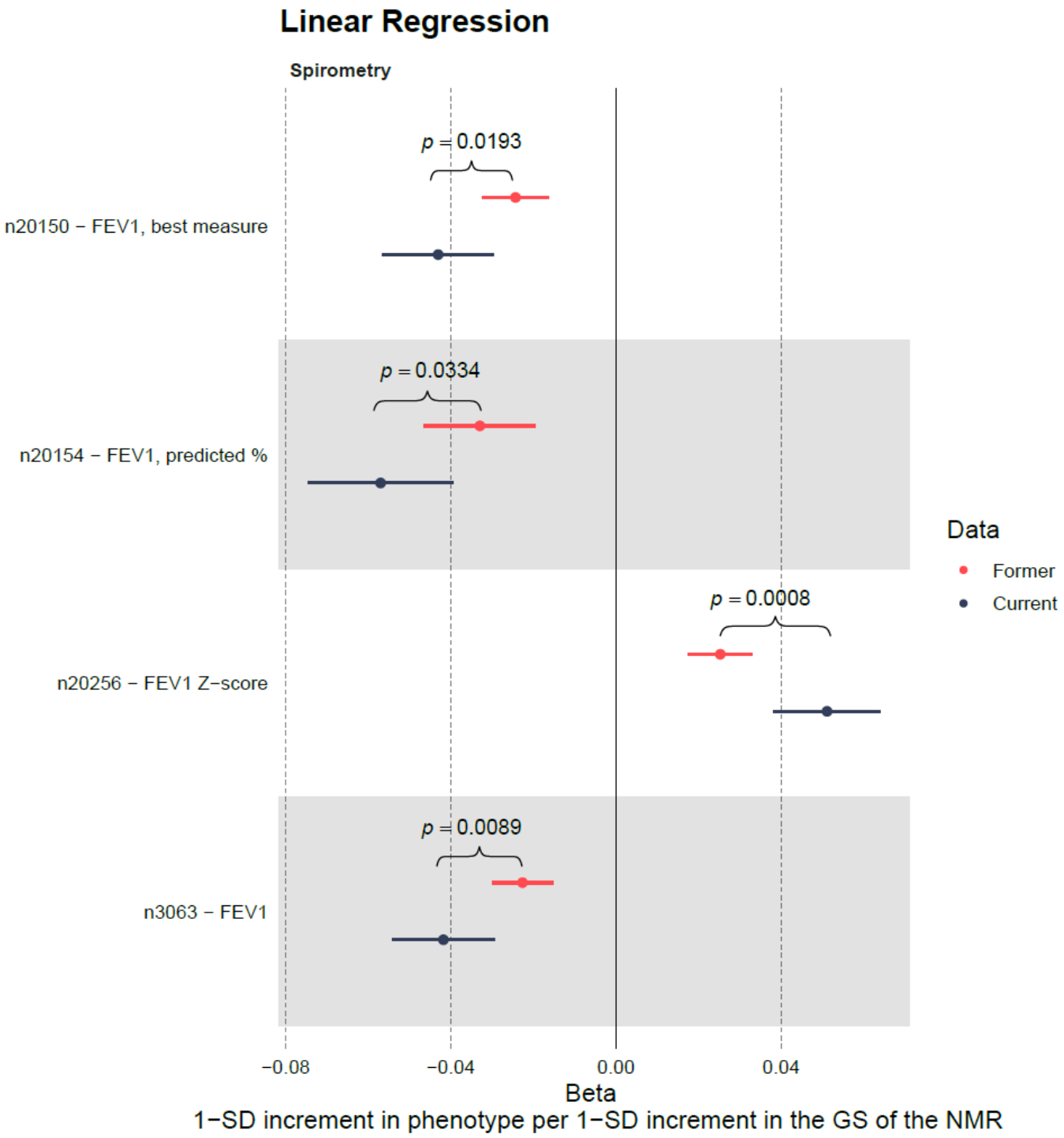

Fig. 4 Current versus Former smokers comparison for the four lung capacity variables highlighted in the Ever versus Never smoker analysis. For both Current and Former subsets all four variables had a statistically significant association with the GS at the 5 % FDR level which had been counted from the PheWAS of the ever smokers. Differences between the subsets were not significant at the 5 % FDR level counted from the phenome-wide Ever versus Never analysis. Nevertheless, the association is visibly attenuated for the former smokers for all four variables. n = normalised after covariates had first been regressed out.