Contents lists available at ScienceDirect

EClinicalMedicine

journal homepage: https://www.journals.elsevier.com/eclinicalmedicine

Research Paper

Mendelian randomization case-control PheWAS in UK Biobank shows evidence of causality for smoking intensity in 28 distinct clinical conditions

Catherine King[a,b,c], Anwar Mulugeta[a,b,d], Farhana Nabi[a], Robert Walton[e], Ang Zhou[a,b], Elina Hyppönen[a,b,c,*]

[a] Australian Centre for Precision Health, University of South Australia Cancer Research Institute, Adelaide, SA 5001, Australia
[b] South Australian Health and Medical Research Institute, Adelaide, Australia
[c] UniSA Clinical and Health Sciences, University of South Australia, Adelaide, SA, Australia
[d] Department of Pharmacology and Clinical Sciences, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia
[e] Asthma UK Centre for Applied Research, Barts Institute of Population Health Sciences, Queen Mary University of London, London, United Kingdom

ARTICLE INFO

Article History:
Received 30 April 2020
Revised 14 July 2020
Accepted 15 July 2020
Available online 31 July 2020

Keywords:
PheWAS
Phenome-wide
MR
mendelian randomisation
Smoking
Intensity
Causal effects
Genetics

ABSTRACT

Background: Smoking is one of the greatest threats to public health worldwide. We integrated phenome-wide association study (PheWAS) and Mendelian randomization (MR) approaches to explore causal effects of genetically predicted smoking intensity across the human disease spectrum.

Methods: We conducted PheWAS case-control analyses in 152,483 ever smokers of White-British ancestry, aged 39–73 years. Disease diagnoses were based on hospital inpatient and mortality registrations. Smoking intensity was instrumented by four genetic variants, and disease risks estimated for one cigarette per day heavier intakes. Associations passing the FDR threshold ($p<0{\bullet}0025$) were assessed for causality using several complementary MR approaches.

Findings: Genetically instrumented smoking intensity was associated with 48 conditions, with MR supporting a possible causal effect for 28 distinct outcomes. Each cigarette smoked per day elevated the odds of respiratory diseases by 5% to 33% (nine distinct diseases, including pneumonia, emphysema, obstructive chronic bronchitis, pleurisy, pulmonary collapse, respiratory failure) and the odds of circulatory disease by 5% to 23% (seven diseases, including atherosclerosis, myocardial infarction, congestive heart failure, arterial embolisms). Further effects were seen for cancer within the respiratory system and other neoplasms, renal failure, septicaemia, and retinal disorders. No associations were observed in sensitivity analyses on 185,002 never smokers.

Interpretation: These genetic data demonstrate the substantial adverse health impacts by smoking intensity and suggest notable increases in the risks of several diseases. Public health initiatives should highlight the damage caused by smoking intensity and the potential benefits of reducing or ideally quitting smoking.

1. Introduction

Over the last fifty years, observational studies have demonstrated associations between smoking and a wide-range of diseases, particularly lung cancer, respiratory and cardiovascular diseases [1]. However, most of the evidence has come from association studies which are susceptible to reverse causality and confounding, and therefore it is difficult to identify and measure the true causal effects of smoking or the degree of damage caused by heavier vs. lighter smoking. Mendelian randomisation (MR) provides an alternative method to determine evidence of causality. MR involves natural randomization of study participants based on a genetic instrument serving as a proxy for the exposure of interest, and akin to a randomised controlled trial, it limits the effect of potential confounders and reduces the likelihood of reverse causation [2]. Several MR studies have been carried out to determine the influence of smoking behaviours, typically on a limited range of outcomes. Phenome-wide associations studies (PheWAS) are increasingly used to evaluate associations between genetic variants across a wide range of phenotypes [3]. The approach is similar to a genome-wide association study (GWAS) which tests associations between a large number of genetic variants and a single outcome, although in PheWAS we test for associations

FundingNational Health and Medical Research Council, Australia.

* Corresponding author.
E-mail address: elina.hypponen@unisa.edu.au (E. Hyppönen).

https://doi.org/10.1016/j.eclinm.2020.100488

Research in context

Evidence before this study

A MEDLINE search using a combination of “smoking”, “phenome-wide”, and “Mendelian randomization” as key words (including synonyms and alternative spellings) identified 58 MR studies on smoking, including three PheWAS. Studies provided convincing evidence of an increased morbidity and mortality in smokers, particularly from respiratory and cardiovascular diseases but also from other conditions. Individual studies suggested protective effects on outcomes such as Parkinson’s disease, Alzheimer’s Disease, osteoarthritis, and possibly hayfever. All PheWAS used single SNPs as instruments.

Added value of this study

We provide evidence for a causal effect of smoking intensity on a wide range of respiratory, and circulatory diseases, in addition to neoplasms, mental illness, injury and poisoning, endocrine/metabolism, sense organ, genitourinary, and infectious diseases. In all cases greater smoking intensity was associated with harm, with an estimated 5% to 33% increase in the risk of disease per cigarette smoked.

Implications of all the available evidence

The public health burden by greater smoking intensity is notable. In addition to smoking intensity, the overall contribution of smoking on disease outcomes will depend on a range of smoking behaviours, including age of initiation, and duration of smoking.

between a single genetic variant (or genetic risk score) and a large number of disease outcomes. Combining MR with the hypothesis free PheWAS approach allows us to identify proof of principle for a causal effect of an exposure across a wide range of outcomes. One smoking MR-PheWAS has been published in which the authors examined the effect of smoking instrumented by rs16969968 at the *CHRNA3* locus, in ever and never smokers on 18,000 phenotypes in the UK Biobank. Using the PHESANT package Millard et al. confirmed the causal effects of smoking on decreased lung function, and also identified a novel detrimental effect on facial ageing [4].

In this two-stage hypothesis-free case-control MR-PheWAS we use information from ever smokers in the UK Biobank to investigate the effects of genetically predicted smoking intensity on a spectrum of carefully curated disease outcomes defined based on hospital inpatient registrations and mortality records [5]. Using four variants determined by a large scale GWAS meta-analysis [6], we instrument smoking intensity to approximate the effects by each cigarette smoked per day (CPD), with our multi-SNP approach increasing power and enabling us to use a number of complementary MR approaches to validate our findings [7].

2. Methods

2.1. Study design and participants

The UK Biobank contains clinical and genetic data for more than 500,000 participants [8]. Participants were aged 39–73 years, at the time of recruitment between March 13, 2006, and Oct 1, 2010. We restricted analyses to unrelated (no first-degree, second-degree, or third-degree relatives) individuals of White-British ancestry, determined on the basis of self-report and genetic data. Ever smokers were defined as participants who had smoked at least one cigarette in their life. Never smokers, had never smoked at the time of data collection (see Supplementary Methods for detailed definitions).

Disease outcomes were identified through linkage to hospital episode statistics and/or the cause of death from the Office of National Statistics mortality data. We identified International Classification of Diseases (ICD; ninth and tenth editions) codes in the dataset and converted them into phenotype codes (phecodes), considered to be more representative of the terminology used in clinical practice [9]. Participants with a given phecode were defined as cases. A total of 1859 phecodes were listed. We excluded all phecodes with less than 200 cases [10], with 904 phecodes in ever smokers retained for further analysis (Supplementary Table 1). Controls were selected as participants without the phecode of interest and without any other phecode within the same disease category (‘refined controls’). Based on the literature we identified three phecodes (cancer within the respiratory system, ischaemic heart disease, and tobacco use disorder) as positive control outcomes expected to show an association with the genetic instrument if analyses were unbiased [1].

This research was carried out using UK Biobank data (application 20,175). Explicit informed consent was obtained from all participants when they enroled in UK Biobank, which itself has approval from the North West Multicentre Research Ethics Committee and the National Information Governance Board for Health and Social Care and (11/NW/0382).

2.2. Genetic information

A GWAS meta-analysis on CPD by the ENGAGE consortium on 31,266 individuals of White European ancestry identified five independent SNPs; rs1051730 (*CHRNA3*), rs6474412 (*CHRNB3*), rs215605 (*PDE1C*), rs4105144 *(CYP2A6)*, and rs7260329 (*CYP2B6*) [6]. Genotyping in the UK Biobank was carried out using the Axiom platform (see Supplementary Methods for further details). All five SNPs (or their proxies) were identified in the UK Biobank data. However, for rs4105144 data was missing for 37% of participants and no appropriate proxy (r^2=0•8) could be identified. As such this SNP was excluded and the remaining four SNPs were retrieved from the UK Biobank GWAS data. Smoking intensity GRS was calculated based on the weighted sum of the individual risk alleles, with weights based on the measured effect sizes in the ENGAGE meta-analyses[6] (Supplementary Fig. 1).

2.3. Statistical analysis

In the first stage of the analysis, we ran the PheWAS using both the refined controls, and the default controls (see Supplementary Methods for definitions). Based on positive control analyses on respiratory cancers, ischaemic heart disease, and tobacco use disorder (Supplementary Tables 2 and 3) primary analyses were conducted using refined controls which excluded all individuals with any disease condition within a category from the control group. The ‘PheWAS’ package in R was used to carry out a logistic regression of each phecode against the GRS, adjusting for age (years), sex (male versus female), assessment centre (22 centres), type of genotyping array (Affymetrix UK BiLEVE Axiom array versus Affymetrix UK Biobank Axiom array), birth location (deciles of east-coordinate and north coordinates for place of birth), and 40 genetic principal components provided by the UK Biobank. For each phecode, its regression analysis was restricted to individuals with complete information on all variables in the model. Correction for multiple testing was based on the false discovery rate (FDR) (p< 0·0025).

In the second stage we carried out two-sample MR analyses on those outcomes which passed the FDR threshold, using the TwoSampleMR and MR PRESSO R packages. SNP - exposure effect size estimates were obtained from the primary ENGAGE GWAS data [6], and SNP - outcome effect size estimates were determined using the UK

Biobank. Primary analyses were conducted using inverse variance weighted (IVW) MR complemented by pleiotropy robust approaches based on different assumptions including MR Egger, weighted median, weighted mode, and MR PRESSO [7]. MR-Egger was used to test for directional pleiotropy, but estimates tend to be conservative and may be biased if the assumption that "the instrument strength is independent of direct effect" is violated. MR-PRESSO was used to carry out a global test for pleiotropy by excluding one SNP at a time, an outlier test to detect potentially pleiotropic outlier variants for each genetically determined exposure-outcome association, and a distortion test which determines the extent to which the causal estimates change when the pleiotropic outlier variant is excluded. Using the TwoSampleMR package, we also generated a scatterplot of SNP effect on smoking intensity against SNP effect on each outcome, and carried out leave-one-out sensitivity analysis, to further identify any potentially pleotropic SNPs. Effect estimates are per a unit increase in the genetically instrumented number of cigarettes smoked per day. As part of instrument validation, we checked for evidence of association between the smoking intensity GRS and potential confounders (Supplementary methods, supplementary Table 4). To further explore whether the smoking intensity GRS was associated with outcomes other than through the exposure of interest, smoking intensity, we repeated the PheWAS in never smokers.

Power for the MR was estimated using the method by Burgess et al [11]. The current study was estimated to have 80% power to detect a 20% increase in risk per each genetically determined CPD increase for 92, and a 50% increase for 488 outcomes, using the refined controls (α=0·05, r^2=0·03, Supplementary Table 5). Statistical analyses were carried out in R (version 3.6.1), and Stata (version 16).

2.4. Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. CK, AM, FN, RW, AZ, and EH had full access to all the data in the study. EH had final responsibility for the decision to submit for publication.

3. Results

We identified 337,484 participants for inclusion, of which 53·7% were women (Supplementary Fig. 2. There were 152,482 ever smokers, with a higher prevalence in men compared to women (50·8% vs. 40·4%, Table 1). Higher prevalence of smoking was associated with older age, greater body mass index (BMI), and reduced general health (Table 1). The GRS explained 3·94% of the variation in smoking intensity in ever smokers (as defined by CPD). We found no evidence of association between the smoking intensity GRS and any of the potential confounders investigated (Supplementary Table 4).

PheWAS analyses using the refined controls identified forty-eight phecodes which were associated with the GRS in ever smokers after FDR correction ($p<$ 0·0025). Signals were observed for disorders of the respiratory, circulatory, endocrine/metabolic, and genitourinary systems, and neoplasms, mental disorders, injuries and poisoning, infectious diseases, and sense organs (Fig. 1). The associations with the lowest P values were for emphysema (p_{PheWAS}=9·24 × 10^{-16}), chronic airway obstruction (p_{PheWAS}=4·12 × 10^{-14}), pneumonia (p_{PheWAS}=2·16 × 10^{-10}), and cancer within the respiratory system (p_{PheWAS}=6 × 10^{-10}). Greater genetically predicted smoking intensity was associated with an increased risk of all the outcomes. We found no associations in never smokers (Supplementary Fig. 3).

We proceeded to carry out two-sample MR analysis of the forty-eight phecodes, representing thirty-two distinct conditions (full data in Supplementary Fig. 4 and Supplementary Table 6). IVW MR provided evidence of a causative effect of genetically instrumented smoking intensity and increased risk for twenty-eight conditions (Fig. 2). The lowest P values were observed for pneumonia (OR per cigarette smoked 1·12, 95% CI 1·07–1·17), abnormal findings examination of lungs (OR 1·19, 95% CI 1·11–1.30), cancer within the respiratory system (OR 1·21, 95% CI 1·12–1·30), emphysema (OR 1·33, 95% CI 1·18–1·49), and atherosclerosis (OR 1·23, 95% CI 1·13 – 1·34). IVW MR, weighted median, weighted mode and MR-PRESSO provided reasonably consistent associations (Supplementary Fig. 4, Supplementary Table 6). We additionally extrapolated our findings to reflect differences in smoking 5 or 10 cigarettes per a day for heavier vs. lighter smokers (Supplementary Table 7).

Emphysema was the only disease outcome associated with genetically instrumented smoking intensity using all five MR methods (Supplementary Fig. 4, Supplementary Table 6). Twelve disease outcomes (cancer within the respiratory system, radiotherapy, secondary malignant neoplasm, hypoosmolality and/or hyponatremia, other mental disorder, alcoholism, atherosclerosis, other disorders of arteries and arterioles, pneumonia, obstructive chronic bronchitis, respiratory insufficiency, and abnormal findings examination of lungs) were supported by four of the MR methods. A further thirteen outcomes (septicaemia; cancer, suspected or other; other retinal disorders; myocardial infarction; congestive heart failure, non-hypertensive; peripheral vascular disease, unspecified; empyema and pneumothorax; pleurisy; pleural effusion; pulmonary collapse; interstitial and

Table 1
Characteristics of individuals and prevalence of smoking.

Characteristic		*n* (%)	Ever smokers		
			Prevalence (n)	Median (IQR)	*P*
Total		337,484	152,482		
Sex	Women	181,236 (53.7)	40.4 (73,199)	15 (10, 20)	<1.0E-300
	Men	156,248 (46.3)	50.8 (79,283)	20 (15, 25)	
Age (in years)	39–49	73,849 (21.9)	38.5 (28,394)	15 (10, 20)	9.54E-148
	50–59	111,272 (33.0)	43.3 (48,166)	20 (10, 20)	
	60–73	152,363 (45.2)	49.8 (75,922)	20 (10, 20)	
BMI (kg/m^2)	< 18.5	1673 (0.5)	42.5 (711)	15 (10, 20)	<1.0E-300
	18.5 - 24.5	109,787 (32.5)	40.3 (44,205)	15 (10, 20)	
	25 - 29.5	143,785 (42.6)	46.6 (66,915)	20 (10, 20)	
	>= 30	81,145 (24.1)	49.4 (40,087)	20 (15, 25)	
	Missing	1094 (0.3)	51.6 (564)		
General health	Excellent	56,531 (16.8)	36.8 (20,794)	15 (10, 20)	<1.0E-300
	Good	197,169 (58.4)	43.7 (86,145)	16 (10, 20)	
	Fair	68,621 (20.3)	53.0 (36,351)	20 (12, 20)	
	Poor	13,983 (4.2)	60.9 (8520)	20 (15, 30)	
	Missing	1180 (0.4)	57.0 (672)		

Median and IQR (interquartile range) are considered for cigarettes smoked per day. *P*-values are generated from likelihood ratio test for models adjusted for age, sex, birth location and assessment centres.

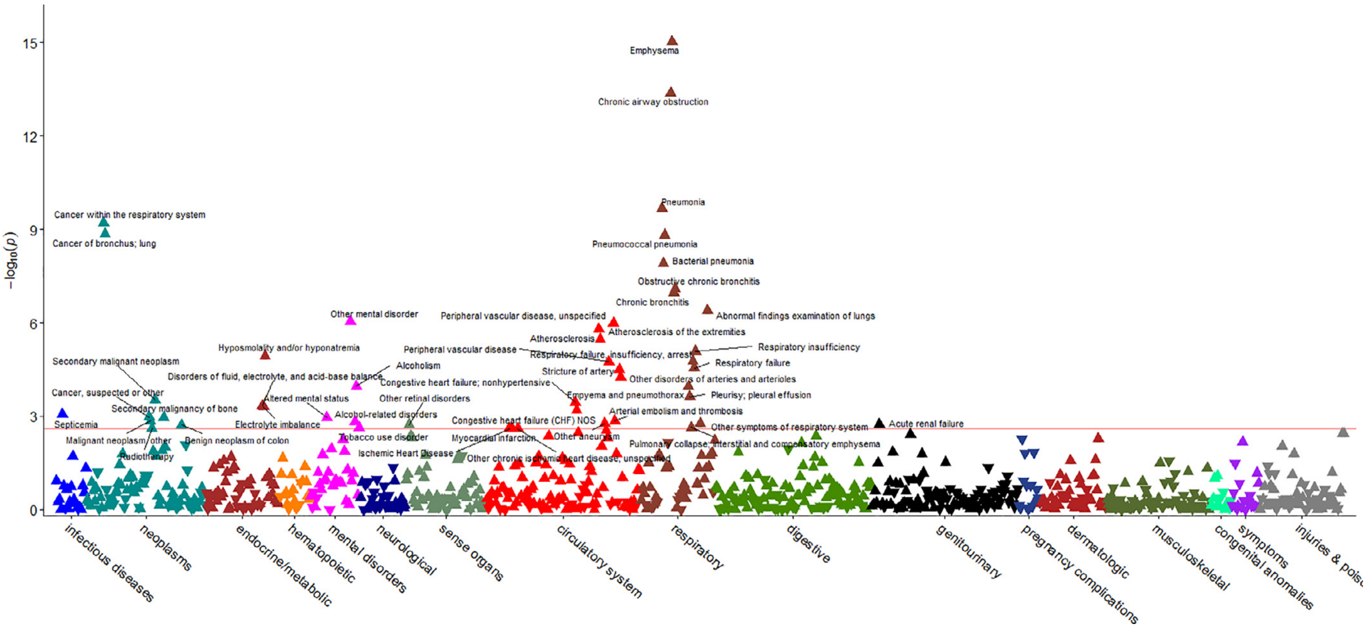


Fig. 1. Manhattan plot illustrating the outcomes of the PheWAS analysis of the smoking intensity GRS in ever smokers. The red line indicates the FDR threshold ($p<0.0025$). Y-axis is minus log transformed P-value of the association between smoking intensity genetic risk score (GRS) and disease outcomes; the X-axis provides the list of labels of 17 diseases groups. Arrows pointing up indicate that the smoking intensity GRS is associated with increased odds of disease. Arrows pointing down indicate that the smoking intensity GRS is associated with decreased odds of disease.

compensatory emphysema; respiratory failure; other symptoms or respiratory failure; acute renal failure; and complication of surgical and medical procedures) were associated with smoking intensity based on three MR methods. An increased risk of benign neoplasm of the colon, altered mental status, tobacco use disorder, other chronic ischaemic heart disease, unspecified, other aneurysm, and other arterial embolism and thrombosis were supported by two analyses methods. Of all methods, MR-Egger tended to be more conservative, and only emphysema, obstructive chronic bronchitis, and empyema and pneumothorax were associated with genetically predicted smoking intensity based on this analysis method.

No evidence of pleiotropy was observed using MR Egger or leave one out analysis (Supplementary Fig. 5). MR PRESSO identified rs7260329 as a potential pleiotropic outlier in analyses on chronic airway obstruction and other aneurysm. However, MR-PRESSO distortion test did not identify significant changes in the effect estimates when this SNP was removed from the analysis.

4. Discussion

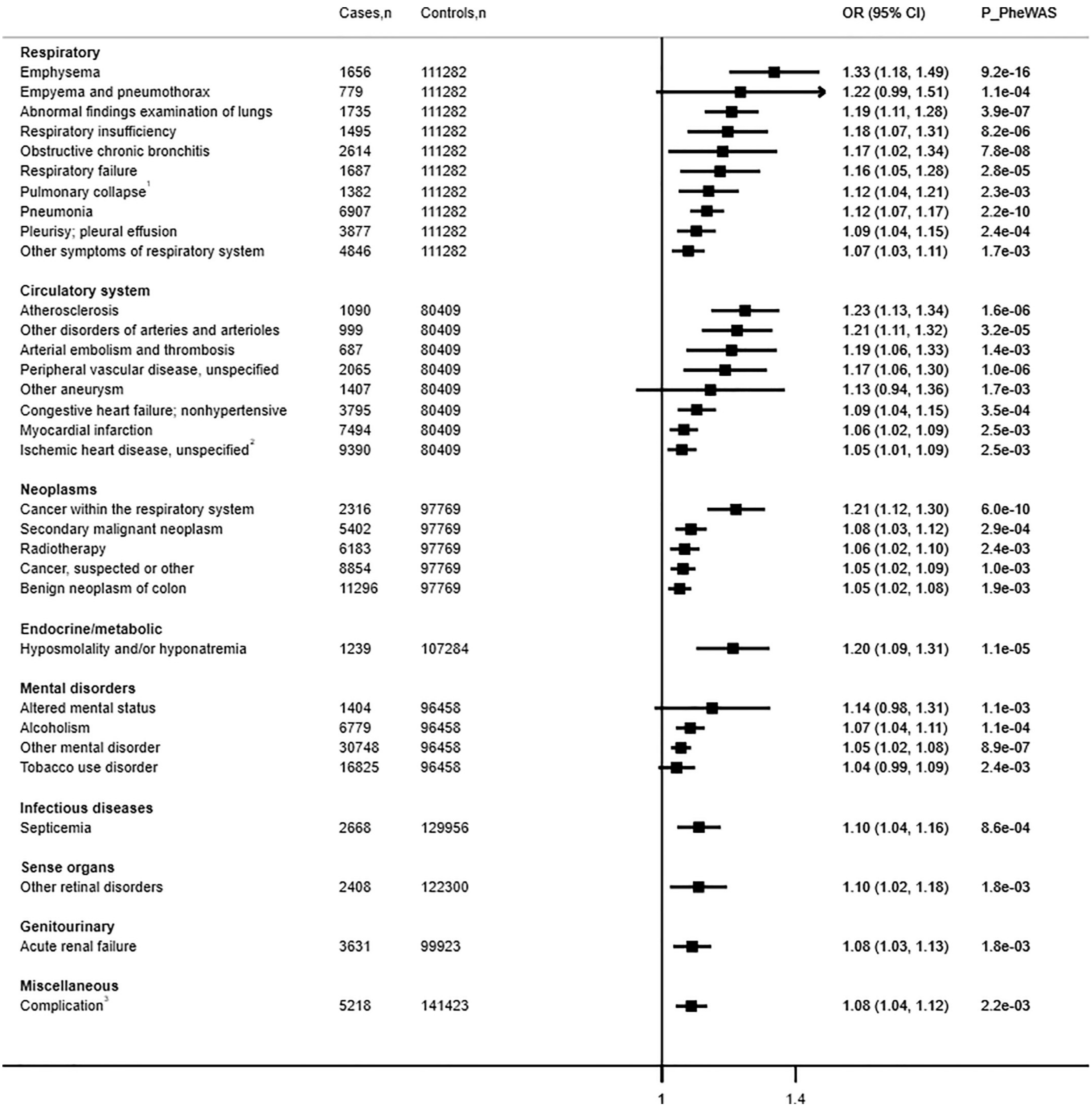
Tobacco smoking is the leading preventable cause of death worldwide, and smokers typically die 10 years earlier than non-smokers. Despite some decline over the last 15–20 years, the global prevalence of tobacco smoking is still estimated to be about 20% of the population aged ≥ 15 years [12]. In the U.S. alone 40 million people smoke, and 16 million people are living with a disease caused by smoking, costing the economy more than \$300 billion per annum through direct medical costs and loss of productivity [13].

Our hypothesis free MR-PheWAS analysis demonstrated the extensive consequences of genetically predicted smoking heaviness on the risk of respiratory diseases, cancers, and cardiovascular diseases. It has also highlighted potential effects on other conditions such as septicaemia, acute renal failure, electrolyte disturbances, retinal disorders, and complications of surgery or medical procedures. For some conditions each genetically predicted cigarette per day elevated the odds of disease by over 30%. If taken to reflect larger differences in smoking, 10 additional cigarettes per a day would predict a nearly 17-fold increase in the odds of emphysema, 8-fold for atherosclerosis (OR 8·13), and 6.5-fold for cancer within the respiratory system (Supplementary Table 7). This highlights the significant damage caused by smoking, beyond smoking initiation, and the potential health benefits which could be gained by reducing the number of cigarettes smoked, or ideally, complete smoking cessation.

Our analysis has confirmed findings from observational studies relating to an increased risk of respiratory disease by heavier smoking [14], particularly chronic obstructive pulmonary disease (COPD), cancer in the respiratory system, and pneumonia. Our results also provide evidence for an increased risk of many other respiratory conditions. The effects on respiratory health are supported by an earlier MR PheWAS of smoking intensity which found causal associations with poor lung function, COPD, and increased cancer of the bronchus and lung [4]. Furthermore, in a MR study by Vie et al., smoking intensity was associated with increased mortality from respiratory diseases [15].

Greater genetically predicted smoking intensity increased the risk of many circulatory diseases. Our results show that increased smoking intensity contributes to all forms of ischaemic heart disease including conditions characterized by both coronary atherosclerosis (myocardial infarction), and those characterized primarily by coronary vasospasm, including variant and microvascular angina (other chronic ischaemic heart disease, unspecified), although concurrent atherosclerosis is not uncommon in angina patients [16]. Other circulatory diseases associated with smoking intensity were peripheral vascular disease, atherosclerosis, arterial embolism, aneurysm, other disorders of arteries and arterioles, and congestive heart failure. Associations for smoking intensity and peripheral vascular disease and aortic aneurysm in the UK Biobank were also previously reported by Millard et al [4].

Smoking is the primary cause of lung cancer and it has also been associated with increased risk of many other types of cancer [17]. In line with earlier MR studies,[4,15] we confirmed the strong effect on cancers of the respiratory tract, but we also report genetic evidence for causal associations more broadly with radiotherapy, secondary malignant neoplasms, and other suspected cancer, as well as benign neoplasm of colon. An important limitation with our study is the ability to detect only relatively strong associations, and we are unable to



	Cases,n	Controls,n	OR (95% CI)	P_PheWAS
Respiratory				
Emphysema	1656	111282	1.33 (1.18, 1.49)	9.2e-16
Empyema and pneumothorax	779	111282	1.22 (0.99, 1.51)	1.1e-04
Abnormal findings examination of lungs	1735	111282	1.19 (1.11, 1.28)	3.9e-07
Respiratory insufficiency	1495	111282	1.18 (1.07, 1.31)	8.2e-06
Obstructive chronic bronchitis	2614	111282	1.17 (1.02, 1.34)	7.8e-08
Respiratory failure	1687	111282	1.16 (1.05, 1.28)	2.8e-05
Pulmonary collapse[1]	1382	111282	1.12 (1.04, 1.21)	2.3e-03
Pneumonia	6907	111282	1.12 (1.07, 1.17)	2.2e-10
Pleurisy; pleural effusion	3877	111282	1.09 (1.04, 1.15)	2.4e-04
Other symptoms of respiratory system	4846	111282	1.07 (1.03, 1.11)	1.7e-03
Circulatory system				
Atherosclerosis	1090	80409	1.23 (1.13, 1.34)	1.6e-06
Other disorders of arteries and arterioles	999	80409	1.21 (1.11, 1.32)	3.2e-05
Arterial embolism and thrombosis	687	80409	1.19 (1.06, 1.33)	1.4e-03
Peripheral vascular disease, unspecified	2065	80409	1.17 (1.06, 1.30)	1.0e-06
Other aneurysm	1407	80409	1.13 (0.94, 1.36)	1.7e-03
Congestive heart failure; nonhypertensive	3795	80409	1.09 (1.04, 1.15)	3.5e-04
Myocardial infarction	7494	80409	1.06 (1.02, 1.09)	2.5e-03
Ischemic heart disease, unspecified[2]	9390	80409	1.05 (1.01, 1.09)	2.5e-03
Neoplasms				
Cancer within the respiratory system	2316	97769	1.21 (1.12, 1.30)	6.0e-10
Secondary malignant neoplasm	5402	97769	1.08 (1.03, 1.12)	2.9e-04
Radiotherapy	6183	97769	1.06 (1.02, 1.10)	2.4e-03
Cancer, suspected or other	8854	97769	1.05 (1.02, 1.09)	1.0e-03
Benign neoplasm of colon	11296	97769	1.05 (1.02, 1.08)	1.9e-03
Endocrine/metabolic				
Hyposmolality and/or hyponatremia	1239	107284	1.20 (1.09, 1.31)	1.1e-05
Mental disorders				
Altered mental status	1404	96458	1.14 (0.98, 1.31)	1.1e-03
Alcoholism	6779	96458	1.07 (1.04, 1.11)	1.1e-04
Other mental disorder	30748	96458	1.05 (1.02, 1.08)	8.9e-07
Tobacco use disorder	16825	96458	1.04 (0.99, 1.09)	2.4e-03
Infectious diseases				
Septicemia	2668	129956	1.10 (1.04, 1.16)	8.6e-04
Sense organs				
Other retinal disorders	2408	122300	1.10 (1.02, 1.18)	1.8e-03
Genitourinary				
Acute renal failure	3631	99923	1.08 (1.03, 1.13)	1.8e-03
Miscellaneous				
Complication[3]	5218	141423	1.08 (1.04, 1.12)	2.2e-03

Fig. 2. Inverse-variance weighted mendelian randomisation analyses on the top 32 distinct smoking intensity–disease associations. Number of cases and number of controls are shown below each disease outcome as (cases, controls). Risk estimates are reported as odds ratios (OR; 95% CI) per one genetically predicted extra cigarette smoked per day. P_PheWAS is P-value from phenome-wide association analysis.

discount a causal effect on other types of cancer. While we only found evidence for an effect of smoking intensity on a relatively limited number of specific cancers, there were a further seven phecodes in the neoplasms category which showed some evidence of association ($p>0{\cdot}0025$ to $p<0{\cdot}05$).

In line with our findings, evidence to date has not supported a strong causal effect of smoking on mental health outcomes. We found some evidence for an association between greater genetically instrumented smoking intensity and selected disorders under the mental health category including altered mental status, other mental disorder, tobacco use disorder, and alcoholism. Altered mental status seemed to relate to disorientation, while the majority of cases in the 'other mental disorder' category were individuals with a history of psychoactive substance abuse, mainly tobacco or alcohol. However, there is evidence that alcohol and tobacco misuse share some genetic aetiology, notably in relation to reward pathways [18]. Although we conducted analyses using several complementary MR approaches, it is possible that these statistical sensitivity analyses may not have been able to fully account for related pleiotropic effects, and that this may have particularly biased associations observed with mental disorders.

In observational studies smoking is associated with an increased risk of renal injury, changes in renal haemodynamics and vascular injury [19], and is associated with an increased risk of death from renal failure [17]. A previous MR study suggested a causal relationship between smoking intensity and glomerular hyperfiltration [20], while our study is the first to report a causative association between genetically instrumented smoking intensity and acute renal failure. An increased risk of acute renal failure may be a direct result of heavy smoking or it could be mediated through other smoking related outcomes such as cardiovascular disease or infection. In line with an earlier study [4], we found evidence for a causative effect of genetically predicted smoking intensity on septicaemia. Further associations, not previously reported in smoking MR studies, include hypoosmolality and/or hyponatremia retinal disorders, and complications of surgery and medical procedures.

Key strengths of our study include the large sample, which enabled us to test for causal effects across the spectrum of disease outcomes, and the analyses with pleiotropy robust methods facilitated by use of a multi-SNP instrument for smoking intensity. Furthermore, we implemented a case-control design, only focusing on carefully curated disease outcomes based on hospital admissions and mortality registrations and selecting controls as participants free of related conditions. One earlier MR-PheWAS on smoking intensity was conducted in the UK Biobank [4], however, this study used a single SNP (rs16969968) approach and conducted analyses with the PHESANT package to identify associations across all available data fields, including self-reported and other information. Nevertheless, many of our findings are consistent with this earlier study, and for example associations with lung cancer, COPD, respiratory failure, peripheral vascular disease, and septicaemia, were picked up using both approaches. While our analysis was not designed to detect associations with facial ageing which came up in the PHESANT analyses, we identified several novel associations, including acute renal failure, retinal disorders, complications of surgery and medical procedures, in addition to confirming the extensive range of respiratory and cardiovascular conditions affected by smoking intensity. A particular strength of our methodology is the use of a conservative control group, and validation of our outcomes using positive control phecodes. Our refined control groups did not include any phecodes relating to the broad disease category. With reference to hypertension as an example, our controls excluded anyone with any type of circulatory disease. We have demonstrated, using positive controls, that this conservative approach is more robust than using the default control parameters. No associations were identified in a further sample consisting of 185,002 never smokers, which further substantiates the plausibility of our findings. As the effects of smoking on disease risk may take many years to develop, we conducted our analyses in ever smokers. As a further sensitivity analysis we repeated the PheWAS in current smokers, and despite the notably smaller sample, results were similar (data not shown).

Several known smoking related outcomes, including stroke, were not picked up by our PheWAS, potentially suggesting important limitations with our study. One disadvantage of this study is that our analyses were only powered to detect relatively strong effects. The threshold of 200 cases was based on an earlier PheWAS examining simple SNP-disease risk associations [10]. In the formal Mendelian randomization setting where statistical power heavily depends on the strength of the genetic instrument (ie SNP-exposure association), we were underpowered for many disease outcomes, and hence could not provide conclusive evidence for many null associations observed in our PheWAS analysis. We also focused on one smoking related behaviour, smoking intensity. It is plausible that smoking intensity may have a more prominent effect on certain disease outcomes, while other smoking behaviours (for example smoking initiation, age of smoking initiation, or duration of smoking) may be more important for others. For example, genetically instrumented smoking initiation, but not smoking intensity, has previously been associated with increased risk of ischaemic stroke [21].

An additional methodological limitation to this study was that we were not able to include *CYP2A6* SNP rs4105144 as neither it nor a proxy were available in the dataset. However, we used a multi-SNP approach, which compared to the earlier single SNP PheWAS instrument [4], provides increased power and facilitates testing for horizontal pleiotropy. Even though MR-PRESSO detected some heterogeneity between variants, which may signal presence of pleiotropy, exclusion of the potentially pleiotropic variants did not affect observed associations. Further, we implemented several modelling strategies (each with different assumptions on pleiotropic effects) and observed broadly consistent effect estimates across these approaches. It was also reassuring that we observed no GRS-disease associations amongst never smokers (i.e. negative controls), supporting the role of smoking as a mediator of the observed effects. Population stratification may introduce spurious associations in MR analyses. To minimize this possibility, we restricted our analysis to white-British ancestry and adjusted all analyses for assessment centre, birth location and top 40 genetic principal components to account for subtle population structure. We also found no evidence for association between the GRS and any of the confounders tested. Moreover, given 5% participation rate in the UK Biobank and evidence of a healthy volunteer effect it is possible that the same factors that contribute to differences in smoking intensity and the odds of disease affect the likelihood of taking part in the study, with the possibility of introducing collider bias. However, collider bias is believed to be less of an issue in MR analyses than other sources of bias such as pleiotropy or population stratification [22]. An earlier smoking PheWAS estimated the effects of related biases under different strengths of simulated confounding and found that even in the scenario where confounder-smoking status association was very strong (with an odds ratio of 10), there was still no evidence of inflation in false positive rate in the ever smokers [4]. In addition, we acknowledge that misclassification may have occurred at both the level of applying ICD codes and also in the automated process of converting them to phecodes [23,24] Furthermore, MR assumes a linear effect, which would not be able to precisely capture the detrimental effects of smoking intensity if the effect is non-linear. Finally, we acknowledge that this study was carried out in participants of White-British ancestry and other studies are required to confirm these associations and their magnitude in other populations.

In the last 15–20 years, the proportion of heavy smokers smoking a pack or more per day has decreased in countries such as the US and Australia, while there has been an increase in those smoking less than 10 CPD [25,26] While this reflects progress, our genetic study suggests that each additional cigarette smoked matters, notably increasing the risks of cancer, respiratory, circulatory and many other diseases. Indeed, the risks of many of the leading causes of morbidity and mortality worldwide are increased by smoking. Public health initiatives targeting smoking cessation can reduce the burden of these smoking related diseases, with enormous potential for health and economic benefits.

Funding

This project was supported by grants from the National Health and Medical Research Council, Australia (GNT1157281, GNT1123603).

Contributors

EH and CK conceptualized the study. CK carried out the literature review. EH, CK, AM, AZ, FN, and RW planned, analysed and interpreted the data. CK drafted the paper. EH, CK, AM, AZ, FN, and RW were responsible for critical revision of the paper and approval of the manuscript.

Data sharing statement

Data will be made available through application to the UK Biobank.

Declaration Competing Interest

Dr. King, Dr Mulugeta, Miss Nabi, and Dr Zhou have nothing to disclose. Dr. Walton reports grants from NIHR Health Technology Asessment, grants from NIHR Programme Grants for Applied Research, personal fees and other from TTS Pharma, outside the submitted work; In addition, Dr. Walton has a patent WALTON R, MCKINNEY E, MARSHALL S, MURPHY M, WELSH K, others. GENETIC INDICATORS OF TOBACCO CONSUMPTION. Patent number: 2001,038,567. Filed date: 24 Nov 2000. Publication date: 01 Jun 2001 licensed to gNostics, and a patent Tucker MR, Walton R, Matthews H, Miskin A. Method and Kit for Assessing a Patient's Genetic Information, Lifestyle and Environment Conditions, and Providing a Tailored Therapeutic Regime. Patent number: US20110251243 A1. Application number: US 12/944,372. Filed date: 11 Nov 2010. Publication date: 13 Oct 2011 issued to None. Dr. Hypponen reports grants from National Health and Medical Research Council, during the conduct of the study; grants from Tour de Cure, grants from Medical Research Future Fund, grants from Australian Research Council, outside the submitted work; .

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.eclinm.2020.100488.

References

[1] National Center for Chronic Disease P, Health Promotion Office on S, Health. Reports of the surgeon general. The health consequences of smoking-50 years of progress: a report of the surgeon general. Atlanta (GA): Centers for Disease Control and Prevention (US); 2014.
[2] Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 2018;362:k601.
[3] Roden DM. Phenome-wide association studies: a new method for functional genomics in humans. J Physiol (Lond) 2017;595(12):4109–15.
[4] Millard LAC, 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 Genet 2019;15(10):e1008353.
[5] Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013;31(12):1102–10.
[6] Thorgeirsson TE, Gudbjartsson DF, Surakka I, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat Genet 2010;42(5):448–53.
[7] Slob EAW, Burgess S. A comparison of robust Mendelian randomization methods using summary data. Genet Epidemiol 2020.
[8] Biobank U. Genetic data. 2019. https://www.ukbiobank.ac.uk/scientists-3/genetic-data/2019).
[9] Wei W-Q, Bastarache LA, Carroll RJ, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS ONE 2017;12(7):e0175508.
[10] Verma A, Bradford Y, Dudek S, et al. A simulation study investigating power estimates in phenome-wide association studies. BMC Bioinform 2018;19(1):120.
[11] Burgess S. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. Int J Epidemiol 2014;43(3):922–9.
[12] WHO. WHO global report on trends in prevalence of tobacco smoking 2000-2025 second edition, 2018.
[13] CDC. Smoking and Tobacco Use - Fast facts. 2019. https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm.
[14] Jayes L, Haslam PL, Gratziou CG, et al. SmokeHaz: systematic reviews and meta-analyses of the effects of smoking on respiratory health. Chest 2016;150(1):164–79.
[15] Vie GÅ, Wootton RE, Bjørngaard JH, et al. The effect of smoking intensity on all-cause and cause-specific mortality—A Mendelian randomization analysis. Int J Epidemiol 2019.
[16] Slavich M, Patel RS. Coronary artery spasm: current knowledge and residual uncertainties. Int J Cardiol Heart Vasc 2016;10:47–53.
[17] Carter BD, Abnet CC, Feskanich D, et al. Smoking and mortality — beyond established causes. N Engl J Med 2015;372(7):631–40.
[18] Schlaepfer IR, Hoft NR, Ehringer MA. The genetic components of alcohol and nicotine co-addiction: from genes to behavior. Curr Drug Abuse Rev 2008;1(2):124–34.
[19] Orth SR. Cigarette smoking: an important renal risk factor - far beyond carcinogenesis. Tob Induc Dis 2002;1(2):137–55.
[20] Åsvold BO, Bjørngaard JH, Carslake D, et al. Causal associations of tobacco smoking with cardiovascular risk factors: a Mendelian randomization analysis of the HUNT study in Norway. Int J Epidemiol 2014;43(5):1458–70.
[21] Larsson SC, Burgess S, Michaëlsson K. Smoking and stroke: a Mendelian randomization study. Ann. Neurol. 2019;86(3):468–71.
[22] Gkatzionis A, Burgess S. Contextualizing selection bias in Mendelian randomization: how bad is it likely to be. Int J Epidemiol 2018;48(3):691–701.
[23] Beesley L.J., Salvatore M., Fritsche L.G., et al. The emerging landscape of health research based on biobanks linked to electronic health records: existing resources, statistical challenges, and potential opportunities. 2020;39(6):773—800.
[24] Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. JMIR Med Inform 2019;7(4):e14325.
[25] CDC. Smoking is down, but almost 38 million American adults still smoke. 2018.
[26] ABS. National health survey: first results, 2017-18 - smoking. 2019.