The causal effect of cigarette smoking on healthcare costs

Padraig Dixon[1,2]*, Hannah Sallis[2,3], Marcus Munafo[2,3,4], George Davey Smith[2,4,5], Laura Howe[2,5]

1: Nuffield Department of Primary Care Health Sciences, University of Oxford
2: MRC Integrative Epidemiology Unit, University of Bristol
3: School of Psychological Science, University of Bristol
4: NIHR Biomedical Research Centre, University of Bristol of Bristol
5: Population Health Sciences, University of Bristol

July 2022

ABSTRACT

Knowledge of the impact of smoking on healthcare costs is important for establishing the external effects of smoking and for evaluating policies intended to modify this behavior. Conventional analysis of this association is difficult because of omitted variable bias, reverse causality, and measurement error. We approached these challenges using a Mendelian Randomization study design, in which genetic variants associated with smoking behaviors were used as instrumental variables. We undertook genome wide association studies to identify genetic variants associated with smoking initiation and a composite index of lifetime smoking on up to 300,045 individuals in the UK Biobank cohort. These variants were used in two-stage least square models and a variety of sensitivity analyses. All results were concordant in indicating a substantial impact of each smoking exposure on annual inpatient hospital costs Our results indicate a substantial impact of smoking on hospital costs. Genetic liability to initiate smoking – ever versus never having smoked – was estimated to increase mean per-patient annual hospital costs by £477 (95% confidence interval (CI): £187 to £766). A one unit change in genetic liability a composite index reflecting the cumulative health impacts of smoking was estimated to increase these costs by £204 (95% CI: £105 to £303). Models conditioning on the causal effect of risk tolerance were not robust to weak instruments for this exposure. Our findings have implications for the scale of external effects that smokers impose on others, and on the probable cost-effectiveness of smoking interventions.

Keywords: Genetics, smoking instrumental variables; healthcare costs, Mendelian Randomization

*Corresponding author. Padraig Dixon, Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford OX2 6GG, padraig.dixon@phc.ox.ac.uk

1 Introduction

Over one billion people smoked tobacco products – mostly cigarettes – in 2015 [1]. Despite declines in prevalence in many industrialized countries [1-3], smoking continues to be associated with substantial morbidity and mortality [4, 5]. Smoking is arguably the most damaging of all voluntary health behaviors [6, 7] and is associated with a variety of adverse economic and socioeconomic outcomes [8]. Here we study the causal association of smoking with one such economic outcome: healthcare costs.

The association between smoking and healthcare costs is important [7]. Accurate estimates of the healthcare expenditure attributable to smoking are necessary to calculate the external effects associated with smoking, and these externalities inform and underpin many government interventions intended to prevent smoking [9-12]. Causal evidence on the effect of smoking on healthcare costs is also necessary for the robust evaluation of specific interventions that aim to prevent smoking and to treat its downstream consequences. Decision-making by individual smokers may be improved with better information about the non-health consequences of smoking [13].

However, establishing the causal effect of cigarette smoking on healthcare costs is challenging [7, 14]. Observed associations of smoking with healthcare costs may arise because smoking is indeed a cause of healthcare costs, because smoking is itself partly determined by healthcare costs, or because smoking is associated with causes or consequences of processes that influence healthcare costs. In general, it is not clear if the factors that may predispose an individual to smoke are themselves independent determinants of healthcare costs. For example, smoking tends to cluster with other behaviors known or suspected to affect healthcare costs, including high body mass index (BMI), poor diet, alcohol consumption and low physical activity [15-19]. Smoking is also heavily socially patterned, and lower socio-economic status groups are more likely to initiate and continue smoking [20, 21]. These groups may have access to fewer resources to ameliorate the consequences of incident disease and its progression [22] [23, 24].

Smoking is also more prevalent amongst groups defined by health status. For example, smoking is more common amongst individuals with depression and schizophrenia [25, 26] [27, 28]. These associations may affect all or some of smoking initiation, smoking intensity, and smoking cessation. For example, smoking influences disease incidence (such as lung cancer) which in turn may prompt cessation. Smoking may reflect elements of self-medication [29] or a desire to control one's weight [30]. Smoking may therefore be both a cause and consequence of health status and other circumstances [14, 31].

Analysis of the effects of smoking may also be complicated by measurement error, including from inaccurate recollection of smoking history. More subtly, even if declarations of smoking habits are accurate, self-reported smoking patterns ("two packs a day for fifteen years") may not fully capture the consequences of smoking on health. Instead, the cumulative physiological insult of lifetime smoking (reflecting both duration and intensity) on respiratory and other functions may be a better measure for this exposure.

Measured associations of smoking with healthcare cost may also partly reflect wider attitudes to risk tolerance, including impulsivity and behavioral disinhibition. In the Grossman model of health capital accumulation [32], differences between individuals in attitudes to risky behaviors such as smoking may reflect differences in risk attitudes. While empirical support for between-individual differences of this type in line with the Grossman model is sparse [33], it does point to the empirical challenges of distinguishing between smoking and risk tolerance in causal analyses.

These difficulties with robust causal inference comprise the classic issues of omitted variable bias, simultaneity and measurement error [7]. We used Mendelian Randomization to address these issues. Using Mendelian Randomization, and subject to the assumptions of instrumental variable analysis, causal effects of smoking on healthcare cost can be robustly identified by perturbations to genetic variants associated with liability to smoke cigarettes. Since this genetic variation is fixed at conception, it cannot be affected by reverse causation. By virtue of the quasi-random allocation of genetic variation at conception, it will be independent of many confounding variables that might otherwise affect the association between smoking and healthcare cost. We attempted to account for measurement error by using a newly developed index of lifetime smoking that reflects the impact of both smoking duration (encompassing initiation and cessation) and smoking intensity on mortality and health.

2 Genetic variants as instrumental variables

We very briefly review the requirements for valid instrumental variable analysis in the context of Mendelian Randomization. Many introductions are available [34-37] and a more extended discussion accompanied by directed acyclic graphs is provided in supplementary material.

An allele refers to the specific genetic code at particular location or locus in the genome. Individuals have two alleles at each location, one inherited from each parent according to Mendel's first law (the random segregation of alleles) and second law (independent assortment of alleles for different traits) of inheritance. Single nucleotide polymorphisms (SNPs) are

examples of genetic variation subject to Mendel's first and second laws. SNPs are single changes to the nucleotides that make up the genomic code. Some SNPs are known to be associated with disease, traits, or specific behaviors. These associations are usually obtained using genome-wide association studies (GWASs). A variety of different SNPs have been found to robustly associate with various smoking phenotypes.

For example, the *CHRNA5–A3–B4* genomic locus contains the rs16969968 SNP, which is associated with the heaviness of smoking. Individuals with two copies of the less common allele (the "AA" genotype) of this SNP increase the amount of smoke inhaled when nicotine content in cigarettes is increased, whereas individuals with two copies of the more common or major allele do not exhibit this compensatory behavior [14]. Each copy of this allele amounts to approximately one additional cigarette per day in smokers [38]. The *CHRNA5–A3*–B4 locus was identified in various GWAS of different diseases relating to smoking including chronic obstructive pulmonary disease and it is associated both with an earlier diagnosis and an increased risk of lung cancer [39]. It is now apparent that smoking is responsible for the association of this locus with smoking-related diseases.

These associations of SNPs with disease, and their conditionally random allocation at concept, establish the potential for this form of genetic variation to be used as instrumental variables in causal analysis. Briefly, variants should affect the exposure of interest, and have no effects on the outcome (healthcare costs in this case) that are not mediated via its influence on the exposure. The latter assumption embodies two further assumptions – first that the variant does not influence confounders the exposure and the outcome (sometimes referred to as correlated pleiotropy), and second that the variant influences the outcome only via the exposure and not through channels that are independent of the exposure (sometimes referred to as uncorrelated pleiotropy).

Correlated pleiotropy may arise when a SNP affects more than one trait through a shared heritable factor and may lead to false positives when assessing the association between the genetic instruments we use for smoking and risk phenotypes; that is, the appearance of a causal relationship when none may exist in truth. For example, a SNP may influence healthcare costs by its influence on smoking but also via its influence on risk attitudes.

Uncorrelated pleiotropy will arise if SNPs are correlated with other SNPs that themselves influence the outcome independently of the exposure of interest. This correlation arises as a consequence of linkage disequilibrium, which typically occurs when variants located in close physical proximity on the genome are inherited together [40]. SNPs may also influence more

than one outcome phenotype through independent pathways, and this will also give rise to this type of pleiotropy.

We expand on these potential violations of the instrumental variable assumptions in supplementary material. We describe below in the Methods section how we assessed and accounted for their presence.

3 Methods

3.1 Conventional multivariable analysis

We implemented conventional multivariable linear regression as well as Mendelian Randomization. The conventional linear regression models related healthcare cost to both smoking status exposures, controlling in each case for age, sex, and Biobank recruitment centre. We implemented this minimally-adjusted model as a basis for comparison with the instrumental variable models.

3.2 Instrumental variable analysis

Instrumental variables were calculated using versions of the Wald ratio. In Mendelian Randomization, the Wald ratio is calculated as the ratio of the association between the SNP(s) and outcome to the association between the SNPs and the exposure [41]. Denoting the SNP-exposure association as β_{exp}, and the SNP-outcome association as β_{out}, the Wald ratio for the causal instrumental variable estimate is

$$\beta_{IV} = \frac{\beta_{out}}{\beta_{exp}}$$

This ratio is equivalent to the two-stage least squares (2SLS) model when a single instrument is used. The first stage of a 2SLS regresses an exposure variable on the instrument variables. The predicted values of the smoking exposures from this regression are then used in the second stage regression with costs as the outcome. The F-statistic from the first stage regression tests the joint significance of all instruments and is an indicator of instrument strength. Weak instruments will lead to estimates biased in the direction of the conventional multivariable estimate.

For the 2SLS models, we developed polygenic risk scores (PRSs), also known as genetic risk scores or allele scores. Each PRS was calculated as the weighted sum of the effect alleles for all

SNPs. Each SNP was weighted by the regression coefficient from the respective GWAS in which the SNP was identified. These estimates were adjusted with age, sex and the first 40 genetic principal components.

The interpretation of the 2SLS effect estimates is that of a unit change in genetic liability to the exposure variable. For the initiation exposure, this reflects a change in genetic liability in case status; that is, the causal effect estimate represents the costs per person, per year of becoming a smoker. For the composite index of lifetime smoking, the 2SLS reflects a unit change in the composite smoking index, equivalent to (for example) a current smoker who has smoked 22 cigarettes per day for 10 years, or an individual who smoked 30 cigarettes per day for 11 years who quit smoking 5 years ago. See Wootton et al [42] for further details on this variable.

3.3 Sensitivity analysis

We conducted several sensitivity analyses. An important potential violation of the requirements for valid instrumental variable analysis in Mendelian Randomization is horizontal pleiotropy, in which a variant is associated with the outcome other than via the exposure of interest. The presence of pleiotropy may be indicated by the presence of heterogeneity in effect estimates across SNPs. This may be formally tested by comparing Cochran's Q statistic (Formula 7) to the critical values of a chi-squared distribution:

$$Q = \sum_{j=1}^{J} \frac{1}{\sigma^2_{\hat{Y}_j}} (\hat{\beta}_j - \hat{\beta}_{IVW})^2$$

Here, *J* indexes the number of SNPs, $\hat{\beta}_j$ is the effect estimate for SNP *j*, $\hat{\beta}_{IVW}$ is the overall inverse variance (IVW) weighted effect calculated for *J* SNPs, and the variance of the SNP-outcome association is denoted by $\sigma^2_{\hat{Y}_j}$. We assessed the sensitivity of our results to potential violations of the exclusion restriction by estimating a variety of over-identified Mendelian randomization models. We implemented a random-effects inverse-variance weighted (IVW) estimator. This estimator calculates the Wald ratio for each SNP separately, and then combines these SNPs using weights determined by the precision of the association between the SNP and healthcare costs. More precisely estimated SNPs receive more weight.

This estimator assumes that no pleiotropy in violation of the exclusion restriction is present, or that any such pleiotropy has a net zero effect on point estimates. The consequence of the latter form of horizontal pleiotropy – in which any effects "balance out" – is that estimates are unbiased but will have somewhat larger variance than in the no-pleiotropy case. We calculated three other estimators that relax this assumption.

The Mendelian Randomization Egger estimator includes and conditions on an intercept, which may be interpreted as the average pleiotropic effect of all included SNPs. Horizontal pleiotropy is indicated if this intercept is statistically different from zero. This estimator is consistent even if all SNPs violate the exclusion restriction, provided further assumptions on the relationship between instrument strength and any direct pleiotropic effect of SNPs are met – for details see [43, 44].

If the Wald ratio estimates for all SNPs are ordered, then the median estimate will be consistent if at least 50% of SNPs are valid instrumental variables. We implemented a penalized and weighted version of a median estimator in which effect estimates are weighted by precision and in which outlying variants (measured by their contribution to the Cochran Q heterogeneity statistic) are penalized by down-weighting their contribution to the overall effect estimate. Finally, the mode estimator is given by the mode of the Wald ratio estimates. Unlike the median estimator, the mode estimator is consistent even if more than 50% of the SNPs are invalid instrumental variables, provided that the largest homogenous cluster of SNPs are valid instrumental variables. We implemented a version of the estimator that down-weighted the contribution of less precisely estimated SNPs.

We also assessed whether genetic variants associated with smoking and smoking heaviness affected costs in “ever smoker” compared to “never smokers” – if these variants influence costs only via smoking heaviness, their effect should only be apparent in ever smokers [45]. We assessed this association by interacting in each of the split GWAS samples the smoking initiation phenotype with variants associated with smoking and located closest to the CHRNA5 locus. Evidence of an interaction in smokers but not non-smokers would constitute some evidence – albeit not definitive evidence – in support of the exclusion restriction for the initiation phenotype.

Smoking SNPs identified in GWASs may also reflect genetic liability to risk tolerance. Risk tolerance may therefore influence smoking behaviour, as well as other risk-taking behaviours that themselves cause healthcare costs. We therefore implemented multivariable Mendelian Randomization, in which the casual effect of both smoking and risk tolerance were jointly estimated in a single instrumental variable model. Multivariable Mendelian Randomization allows for the direct effect (not via the other exposure) of each exposure to be determined. It also allows the causal effect of each exposure to be expressed as conditional on the other. We implemented separate models for smoking initiation and risk tolerance, and for lifetime smoking and risk tolerance. We used the methods of Sanderson and colleagues [46] to

implement our models, as implemented in the MVMR R package (https://github.com/WSpiller/MVMR).

We also explored potential consequences of correlated pleiotropy. We applied Steiger filtering [47] to each smoking exposure as a simple test of the assumption that instruments affect the outcome via the influence of these instruments on the exposure. This test calculates the variance explained by the SNPs, and tests if the variance explained by the instrument in the outcome is less than the variance explained in the exposure. If so, this provides some evidence that the SNPs influence the exposure via the outcome, and not because the direction of causation is running from the outcome to the exposure. The weighted median and weighted mode estimators, described above, are robust to some types of correlated pleiotropy, provided that this (and other forms of horizontal pleiotropy) affect less than half of variants (for the median estimator) or that the modal pleiotropic effect is zero (for the modal estimator). We also explored the use of the "MR CAUSE" methodology [48] as a means of accounting for correlated pleiotropy but could not obtain stable estimates.

4 Data

We used the UK Biobank study as our primary source of data (13–15). This is a population-based cohort of some 500,000 individuals recruited between 2006 and 2010. All adults aged 40-69 living in defined catchment areas were invited to participate [49, 50], with a final response rate of 5.45%. Participants provided a wide variety of personal and phenotypic data at baseline recruitment, and most consented to genotyping. Individuals in the cohort are generally healthier, wealthier and more educated than the wider UK general population [49, 51].

We studied two measures of smoking: smoking initiation and a composite index of lifetime smoking. The composite index of lifetime smoking was created by Wootton and colleagues [42] to reflect smoking initiation, duration of smoking, heaviness of smoking and smoking cessation (if any). These different measures were aggregated into a composite lifetime smoking index with a half-life constant to reflect the exponential declining impact of smoking on health at a given time. The half-life was obtained from a simulation of the effect of smoking on lung cancer and on all-cause mortality in the UK Biobank.

The smoking initiation phenotype measure captures whether cohort participants ever smoked, without further distinction according to duration or intensity. A binary variable indicating smoking initiation ("ever" smoked versus "never" smoked) was created from the lifetime smoking index, with non-zero values of the lifetime index indicating an individual had initiated smoking.

Mendelian Randomization studies may be biased if the same sample used to select SNPs is also used as the sample in which those SNPs are analyzed as instrumental variables [52]. Sample overlap will tend to bias associations towards the observational exposure-outcome association. We randomly split our available Biobank sample into two non-overlapping samples. On one of these sets, we conducted *de novo* GWASs to identify SNPs, which we then analyzed on the other set of data. We then performed fixed-effect meta-analysis to obtain a single summary measure of effect (and the associated confidence intervals) across both samples. We used a set of standard in-house [53, 54] quality control procedures for conducting the *de novo* UK Biobank GWASs. We used a clumping threshold of R^2<0·001 to account for potential linkage disequilibrium between SNPs.

Genetic data on lifetime smoking (and the related binary smoking initiation variable) was obtained from two *de novo* GWASs of using the composite lifetime smoking index undertaken on 318,067 individuals in the UK Biobank. We also conducted a *de novo* GWAS for a measure of risk tolerance to use in multivariable Mendelian Randomization (N=274,450). This was constructed using responses to UK Biobank questionnaires. A score was created from response to questions relating to days per week of moderate and vigorous physical activity, hours of TV viewing, breaking of motorway speed limits, illicit drug use, alcohol consumption, self-harm, and sexual activity. The precise details of the construction of this variable are given in the supplementary material.

Genetic and other data from UK Biobank were linked to records of inpatient hospital care. A patient undergoing an inpatient hospital visit will occupy a bed but does not necessarily stay overnight. The process by which episodes of care were coded to reflect costs are described elsewhere [55, 56]. Briefly, episodes of inpatient hospital care were coded to create hospital resource groups (HRGs), which denote episodes of care with similar diagnoses, operations and procedures. These HRGs were then cross-referenced to unit costs of care to create a per-person, per-year overall figure for inpatient hospital costs.

All analyses were performed in *R* version 4.02 and Stata version 16.1. Analysis code is available at www.github.com/pdixon-econ/MR_smoking_costs.

5 Results

Up to 300,045 individuals were analyzed, of whom 54% were female (n=161,022). Mean age at recruitment was 57 years (standard deviation = 8.0 years). Mean costs per year were £478 and median costs £87. Some 45% of the sample had zero inpatient NHS costs. There were 87,651

individuals in this sample who had ever smoked. The range of the cumulative smoking index is from 0 to 4, with a score of 0 indicating a never-smoker. The mean value of this index amongst all individuals was 0.33 (standard deviation: 0.67) and amongst smokers 1.14 (standard deviation: 0.78).

Conventional multivariable models, estimated using linear regression and adjusted for age and sex but without any genetic information, are summarized in Table 1.

Table 1 **Multivariable estimates of association between smoking initiation, lifetime smoking, risk tolerance, and annual inpatient hospital costs**

	N	Effect estimate	95% confidence interval
Phenotype			
Smoking initiation	300,045	£208	£196 to £219
Lifetime smoking amongst smokers	87,651	£142	£131 to £153

Note to Table 1: These models adjust for age, sex and recruitment centre only.

The effect estimate on the smoking initiation phenotype indicates the observational association between becoming a smoker compared to never smoking on per-person, per-year inpatient hospital costs. Likewise, a unit change in the lifetime smoking index in the intensity, duration and cessation (if any) is associated with £142 increase in annual per-person inpatient costs. Given that median per-person costs in this sample are £87 per year, both smoking exposures are associated with a substantial increase in annual costs. These conventional multivariable models will be confounded by any variables other than age, sex and UK Biobank recruitment centre that jointly influence smoking and healthcare costs.

Table 2 details information relating to the polygenic instrumental variable models.

Table 2 **Cases, number of SNPs and strength of instruments**

	N	N of SNPs	% of variance explained by polygenic risk score	F-statistic from first stage of 2SLS polygenic risk score 2SLS model
Phenotype				
Smoking initiation sample 1	149,995	10	0.19%	223
Smoking initiation sample 2	150,050	11	0.25%	293
Lifetime smoking sample 1	149,995	17	0.40%	584
Lifetime smoking sample 2	150,050	15	0.44%	620

Note to table : Percentage of variance obtained from pseudo R^2 from a logistic regression of initiation status on the PRS for initiation, and from R^2 obtained from linear regression of lifetime smoking on the PRS for this exposure.

The number of SNPs refers to genome-wide significant hits in each split sample of the UK Biobank cohort. The proportion of variance explained by the polygenic risks scores is modest, being less than 1% in all cases. However, the first stage F-statistic from the 2SLS indicates that these are strong instruments.

5.1 Instrumental variable results

Table 3 summarises the causal effect estimates and associated 95% confidence intervals representing the effect of genetic liability to each smoking phenotype, estimated using polygenic risk score 2SLS models.

Table 3 2SLS allele score estimates

	Beta	95% confidence interval
Phenotype		
Smoking initiation	£477	£187 to £766
Lifetime smoking	£204	£105 to £303

Note to Table 3: Estimates and confidence intervals from fixed-effects meta analysis over each split sample.

The betas in Table 3 reflect inpatient hospital costs per person per year of a (genetically influenced) change in liability to smoking initiation and to the lifetime smoking index. Note that these represent the effect of genetic liability to each smoking phenotype, and are not directly comparable to the corresponding conventional multivariable estimates [57, 58]. Note also that the wide confidence intervals are a consequence of the modest proportion of variance that each polygenic risk scores explains in the respective phenotypes. Nevertheless, the effect estimates are consistent with substantial effects of genetic liability to each smoking phenotype on healthcare costs.

5.2 Sensitivity analysis

Inspection of Cochran's Q revealed little evidence of heterogeneity (Table 4).

Table 4 Cochran's Q statistic and heterogeneity by phenotype

	N of SNPs	Q statistic	Q p-value
Phenotype			

Smoking initiation sample 1	10	13.2	0.16
Smoking initiation sample 2	11	8.7	0.47
Lifetime smoking sample 1	17	25.2	0.07
Lifetime smoking sample 2	15	12.2	0.51

Note to table: There were insufficient SNPs to calculate the Q statistic for the risk tolerance exposure.

The data do not suggest the presence of pleiotropy, with the possible exception of sample 1 for the lifetime smoking phenotype. We nevertheless report the findings of the various pleiotropy robust estimators in supplementary material. These data are generally concordant in indicating a similarity of causal effects across estimators for both samples and for both smoking phenotypes. They are similar to the 2SLS analysis in indicating a substantial effect of smoking on costs. The results of Steiger filtering tests indicated that the direction of causality was more likely to be from each smoking exposure to outcome rather than vice versa. These results are reported in supplementary material. Interactions between initiation and variants located near the CHRNA5 locus were consistent with the null (see supplementary material).

We used multivariable Mendelian Randomization to examine whether risk tolerance may influence the association between smoking and healthcare costs. Multivariable Mendelian Randomization requires that the instrumental variable assumptions hold jointly for all included exposures. In particular, SNPs must be strongly associated with each exposure, conditional on the other included exposures. Sanderson and colleagues [46] recommend that each exposure included have an F-statistic greater than 10. In the event, the conditional strengths of the SNPs for risk tolerance (but not the two smoking exposures) were too weak to make reliable inferences. Further details of these results, including the F-statistics, are reported in supplementary material.

6 Discussion

Conventional multivariable estimates in simple linear models indicated a substantial association of both smoking initiation and the composite smoking index on annual per-patient inpatient costs in the UK Biobank cohort. Mendelian Randomization analyses of both of genetic liability exposures were estimated with more uncertainty, but were consistent with potentially large impacts on hospital costs.

Under the assumption of monotonicity – that the effect of the instrumental variables on each respective exposure is in the same direction for all subjects – these estimates are local average treatment effect estimates that capture lifelong genetic liability to smoke. In the present study, this assumption means for those individuals whose smoking differs by levels of the respective instruments, then the association change in smoking behaviour is in the same direction for all individuals. This assumption may be reasonable although we do not formally test it.

The Mendelian Randomization estimates refer to the impact on costs of a genetic liability to smoking and do not necessarily correspond with the effect estimates that would be obtained from a hypothetical experiment to compel people to smoke (or to compel smoking cessation) and in which healthcare costs are observed over a period of time. Mendelian Randomization estimates do not necessarily comply with the stable unit treatment value assumption (SUTVA) [59]. Our effect estimates relate to the cumulative lifetime impact of a genetic liability to smoke, and cannot necessarily be used to make inferences about the effects of smoking at particular phases of life. This also complicates comparisons between conventional multivariable models and Mendelian Randomization estimates on the liability scale.

Studies using similar smoking exposures (e.g. [60]) have found evidence of pleiotropic effects, particularly in relation to smoking initiation. There was little evidence of heterogeneity associated with pleiotropy in the present study. We further examined estimates from pleiotropy robust estimators, which were similar and suggested that the same causal effect was being identified by each such estimator. However, if risk attitudes (or other variables) confound the association between smoking and healthcare costs, then interventions that address smoking may not necessarily influence healthcare costs [61]. Moreover, confounding the effects of smoking on healthcare cost with those of attitudes toward risk may lead to biased estimates of the external effects of smoking. There was no attenuation of the effects of smoking on healthcare cost when including instruments for risk tolerance, although this finding was not robust to weak instrument bias. We cannot rule out bias arising from associations between risk attitudes and smoking, including via correlated pleiotropy.

We estimated all associations on the UK Biobank cohort, which is a self-selected sample of individuals who are healthier, wealthier and more educated than the wider UK population of adults. This may lead to sample selection bias. Simulation studies [62, 63] suggest that biases other than those attributable to selection, such as violations of the exclusion restriction caused by pleiotropy, may have a greater impact on effect estimates. We were also limited to studying individuals of European (White British) ancestry, and to the analysis of inpatient hospital costs. The relationships we estimate on this ethnicity may not reflect associations between smoking and costs in other ancestral groups.

Mendelian Randomization requires that variants are conditionally independent of local environments. This assumption will be violated if there is clustering of alleles in certain environments, or amongst certain types of individual. In relation to the former, geographical stratification that is not wholly accounted for by genetic principal components is a feature of UK Biobank [64] and may impart bias to our results. In relation to the latter, alleles that indicate genetic liability to smoking will tend to cluster amongst individuals if assortative mating means that smokers are more likely to marry other smokers than they are to marry non-smokers [65, 66].

There may also be biases from dynastic effects [67]; for example, children raised by smokers are more likely to smoke independently of the genetic liability that a child has to smoke. Howe et al [68] use a within-family GWAS to control for demographic and indirect genetic effects, such as the impact of non-transmitted parental alleles. These GWAS found smaller effect estimates for smoking than in GWASs of unrelated individuals. This would suggest that the effect estimates in the present study, drawn from GWASs of unrelated individuals, may be upward biased.

7 Conclusion

Mendelian Randomization analysis of the causal effect of genetic liability smoking initiation and lifetime smoking indicated a substantial impact of each exposure on annual inpatient hospital costs, although these associations may be inflated by indirect genetic effects and potentially also by risk attitudes. Nevertheless, the costs we attribute to smoking suggest that externalities associated with smoking, and the cost-effectiveness of interventions to prevent smoking initiation and encourage cessation, may be considerable.

Declarations

Funding statement: PD, HMS, GDS, MM and LDH are members of the MRC Integrative Epidemiology Unit at the University of Bristol which is supported by the Medical Research Council and the University of Bristol (MC_UU_00011/1, MC_UU_00011/7). PD acknowledges support from a Medical Research Council Skills Development Fellowship (MR/P014259/1). HMS is supported by the European Research Council (Reference: 758813 MHINT). LDH was supported by Health Foundation grant "Social and economic consequences of health status - Causal inference methods and longitudinal, intergenerational data", awarded under the Social and Economic Value of Health programme (Award reference 807293) and a Career Development Award from the UK Medical Research Council (MR/M020894/1).

Conflict of interest statement: The authors declare no conflicts of interest.

Acknowledgments: This research was conducted using the UK Biobank Resource under Application Number 29294 and Application Number 9142.

References

1. World Health O. WHO global report on trends in prevalence of tobacco smoking 2015: World Health Organization; 2015.
2. Reitsma MB, Fullman N, Ng M, Salama JS, Abajobir A, Abate KH, et al. Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: a systematic analysis from the Global Burden of Disease Study 2015. The Lancet. 2017;389(10082):1885-906.
3. Roser M, Ritchie H. Smoking. . 2013; Available from: Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/smoking' [Online Resource]
4. Doll R, Peto R, Boreham J, Sutherland I. Mortality in relation to smoking: 50 years' observations on male British doctors. BMJ. 2004;328(7455):1519.
5. Jha P, Peto R. Global Effects of Smoking, of Quitting, and of Taxing Tobacco. N Engl J Med. 2014 2014/01/02;370(1):60-8.
6. Danaei G, Ding EL, Mozaffarian D, Taylor B, Rehm J, Murray CJL, et al. The Preventable Causes of Death in the United States: Comparative Risk Assessment of Dietary, Lifestyle, and Metabolic Risk Factors. PLoS Med. 2009;6(4):e1000058.
7. Cawley J, Ruhm CJ. Chapter Three - The Economics of Risky Health Behaviors. In: Pauly MV, McGuire TG, Barros PP, editors. Handbook of Health Economics: Elsevier; 2011. p. 95-199.
8. Harrison S, Davies AR, Dickson M, Tyrrell J, Green MJ, Katikireddi SV, et al. The causal effects of health conditions and risk factors on social and socioeconomic outcomes: Mendelian randomization in UK Biobank. Int J Epidemiol. 2020.
9. Zohrabian A, Philipson TJ. External costs of risky health behaviors associated with leading actual causes of death in the U.S.: a review of the evidence and implications for future research. Int J Environ Res Public Health. 2010;7(6):2460-72.
10. Ekpu VU, Brown AK. The Economic Impact of Smoking and of Reducing Smoking Prevalence: Review of Evidence. Tob Use Insights. 2015;8:1-35.
11. Sherry G, Peter CS, Donald SK, Jody S. Economics of Health Behaviors and Addictions: Contemporary Issues and Policy Implications. Oxford University Press; 2011.
12. Jin L, Kenkel D, Liu F, Wang H. Retrospective and Prospective Benefit-Cost Analyses of U.S. Anti-Smoking Policies. Journal of Benefit-Cost Analysis. 2015;6(1):154-86.
13. Chaloupka FJ, Warner KE. Chapter 29 The economics of smoking. Handbook of Health Economics: Elsevier; 2000. p. 1539-627.
14. Lassi G, Taylor AE, Timpson NJ, Kenny PJ, Mather RJ, Eisen T, et al. The CHRNA5–A3–B4 Gene Cluster and Smoking: From Discovery to Therapeutics. Trends Neurosci. 2016 2016/12/01/;39(12):851-61.
15. Audrain-McGovern J, Benowitz NL. Cigarette smoking, nicotine, and body weight. Clin Pharmacol Ther. 2011;90(1):164-8.
16. Dare S, Mackay DF, Pell JP. Relationship between smoking and obesity: a cross-sectional study of 499,504 middle-aged adults in the UK general population. PLoS One. 2015;10(4):e0123579.

17. Healton CG, Vallone D, McCausland KL, Xiao H, Green MP. Smoking, obesity, and their co-occurrence in the United States: cross sectional analysis. BMJ. 2006;333(7557):25-6.
18. Barendregt JJ, Bonneux L, van der Maas PJ. The Health Care Costs of Smoking. N Engl J Med. 1997 1997/10/09;337(15):1052-7.
19. Carreras-Torres R, Johansson M, Haycock PC, Relton CL, Davey Smith G, Brennan P, et al. Role of obesity in smoking behaviour: Mendelian randomisation study in UK Biobank. BMJ. 2018 May 16;361:k1767.
20. Balia S, Jones AM. Catching the habit: a study of inequality of opportunity in smoking-related mortality. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011 2011/01/01;174(1):175-94.
21. Hiscock R, Bauld L, Amos A, Platt S. Smoking and socioeconomic status in England: the rise of the never smoker and the disadvantaged smoker. Journal of Public Health. 2012;34(3):390-6.
22. Laudicella M, Siciliani L, Cookson R. Waiting times and socioeconomic status: Evidence from England. Soc Sci Med. 2012 2012/05/01/;74(9):1331-41.
23. Moscelli G, Siciliani L, Gutacker N, Cookson R. Socioeconomic inequality of access to healthcare: Does choice explain the gradient? J Health Econ. 2018 2018/01/01/;57:290-314.
24. Sheringham J, Asaria M, Barratt H, Raine R, Cookson R. Are some areas more equal than others? Socioeconomic inequality in potentially avoidable emergency hospital admissions within English local authority areas. J Health Serv Res Policy. 2016 2017/04/01;22(2):83-90.
25. Lasser K, Boyd JW, Woolhandler S, Himmelstein DU, McCormick D, Bor DH. Smoking and mental illness: A population-based prevalence study. JAMA. 2000 Nov 22-29;284(20):2606-10.
26. Smith PH, Mazure CM, McKee SA. Smoking and mental illness in the U.S. population. Tob Control. 2014 Nov;23(e2):e147-53.
27. Royal College of Physicians RCoP. Smoking and mental health. London2013.
28. McClave AK, McKnight-Eily LR, Davis SP, Dube SR. Smoking characteristics of adults with selected lifetime mental illnesses: results from the 2007 National Health Interview Survey. Am J Public Health. 2010 Dec;100(12):2464-72.
29. Audrain-McGovern J, Rodriguez D, Kassel JD. Adolescent smoking and depression: evidence for self-medication and peer smoking mediation. Addiction. 2009;104(10):1743-56.
30. Cawley J, Dragone D, Von Hinke Kessler Scholder S. The Demand for Cigarettes as Derived from the Demand for Weight Loss: A Theoretical and Empirical Investigation. Health Econ. 2016 2016/01/01;25(1):8-23.
31. Auld MC. Using observational data to identify the causal effects of health-related behaviour. The Elgar Companion to Health Economics, Second Edition: Edward Elgar Publishing; 2012.
32. Grossman M. On the concept of health capital and the demand for health. Journal of Political economy. 1972;80(2):223-55.
33. Cutler DM, Glaeser E. What Explains Differences in Smoking, Drinking, and Other Health-Related Behaviors? Am Econ Rev. 2005;95(2):238-42.
34. Pingault J-B, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nature Reviews Genetics. 2018 2018/09/01;19(9):566-80.

35. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601.
36. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014 Sep 15;23(R1):R89-98.
37. Richmond RC, Smith GD. Mendelian randomization: Concepts and scope. Cold Spring Harb Perspect Med. 2021:a040501.
38. Vie GÅ, Wootton RE, Bjørngaard JH, Åsvold BO, Taylor AE, Gabrielsen ME, et al. The effect of smoking intensity on all-cause and cause-specific mortality—a Mendelian randomization analysis. Int J Epidemiol. 2019;48(5):1438-46.
39. Lebrett MB, Crosbie EJ, Smith MJ, Woodward ER, Evans DG, Crosbie PAJ. Targeting lung cancer screening to individuals at greatest risk: the role of genetic factors. J Med Genet. 2021:jmedgenet-2020-107399.
40. Visscher Peter M, Brown Matthew A, McCarthy Mark I, Yang J. Five Years of GWAS Discovery. The American Journal of Human Genetics. 2012 2012/01/13/;90(1):7-24.
41. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press, ; 2016.
42. Wootton RE, Richmond RC, Stuijfzand BG, Lawn RB, Sallis HM, Taylor GMJ, et al. Evidence for causal effects of lifetime smoking on risk for depression and schizophrenia: a Mendelian randomisation study. Psychol Med. 2020;50(14):2435-43.
43. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015 Apr;44(2):512-25.
44. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. Eur J Epidemiol. 2017;32(5):377-89.
45. 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 Genetics. 2019;15(10):e1008353.
46. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int J Epidemiol. 2019;48(3):713-27.
47. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS genetics. 2017;13(11):e1007081.
48. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nat Genet. 2020 2020/07/01;52(7):740-7.
49. 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 Med. 2015;12(3):e1001779.
50. Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank: Current status and what it means for epidemiology. Health Policy and Technology. 2012 2012/09/01/;1(3):123-6.
51. 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 the General Population. American journal of epidemiology. 2017 Jun 21.

52. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. Genet Epidemiol. 2016;40(7):597-608.
53. Mitchell R, Hemani G, Dudding T, Corbin L, S. H, Paternoster L. UK Biobank Genetic Data: MRC-IEU Quality Control, version 22019.
54. Mitchell R EB, Raistrick CA, Paternoster L, Hemani G, Gaunt TR MRC IEU UK Biobank GWAS pipeline version 2.2019.
55. Dixon P, Hollingworth W, Harrison S, Davies NM, Davey Smith G. Mendelian Randomization analysis of the causal effect of adiposity on hospital costs. J Health Econ. 2020 2020/03/01/;70:102300.
56. Dixon P, Davey Smith G, Hollingworth W. The Association Between Adiposity and Inpatient Hospital Costs in the UK Biobank Cohort. Appl Health Econ Health Policy. 2018 2018/12/31.
57. Dixon P, Harrison S, Hollingworth W, Davies NM, Smith GD. Estimating the causal effect of liability to disease on healthcare costs using Mendelian Randomization. Econ Hum Biol. 2022 2022/06/30/:101154.
58. Davey Smith G, Munafò MR. In: TARG Blog - The Tobacco and Alcohol Research Group blog, editor. Does schizophrenia influence cannabis use? How to report the influence of disease liability on outcomes in Mendelian randomization studies
University of Bristol2019.
59. Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences: Cambridge University Press; 2015.
60. Gage SH, Sallis HM, Lassi G, Wootton RE, Mokrysz C, Davey Smith G, et al. Does smoking cause lower educational attainment and general cognitive ability? Triangulation of causal evidence using multiple study designs. Psychol Med. 2022;52(8):1578-86.
61. Khouja JN, Wootton RE, Taylor AE, Davey Smith G, Munafò MR. Association of genetic liability to smoking initiation with e-cigarette use in young adults: A cohort study. PLoS Med. 2021;18(3):e1003555-e.
62. 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.
63. Hughes RA, Davies NM, Davey Smith G, Tilling K. Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis. Epidemiology. 2019;30(3).
64. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nature Communications. 2019 2019/01/18;10(1):333.
65. Ask H, Rognmo K, Torvik FA, Røysamb E, Tambs K. Non-Random Mating and Convergence Over Time for Alcohol Consumption, Smoking, and Exercise: The Nord-Trøndelag Health Study. Behav Genet. 2012 2012/05/01;42(3):354-65.
66. Clarke T-K, Adams MJ, Howard DM, Xia C, Davies G, Hayward C, et al. Genetic and shared couple environmental contributions to smoking and alcohol use in the UK population. Mol Psychiatry. 2019 2019/11/25.
67. Fletcher JM. The promise and pitfalls of combining genetic and economic research. Health Econ. 2011;20(8):889-92.

68. Howe LJ, Nivard MG, Morris TT, Hansen AF, Rasheed H, Cho Y, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. Nat Genet. 2022 2022/05/09.