Causal Inference: An Investigation into the Relationship between Smoking and Heart Disease

Junyi Liao

20307110289@fudan.edu.cn

Abstract

★ **Background:** Heart disease is an international public health concern, and smoking has long been recognized as a leading modifiable risk factor associated with it. In this study, I investigated the effect of smoking on the presence of heart disease using methods of causal inference.

★ **Methods:** I conducted my analysis on an open-source dataset from Kaggle. The analytical sample included 2,398 patients across the US. Patients were divided to two groups of "smokes" and "never smoked". Logistic regression models adjusted for various set of covariates were fitted to estimate the odds ratio (OR) of heart disease in the two groups. Effect modification on gender was also studied. Additionally, a matched sample created by one-to-one propensity score matching was used to fit regression models. Mediation analysis was conducted to estimate the effect mediated through cholesterol, another risk factor associated with both smoking and heart disease. A sensitivity analysis based on E-value was applied to evaluate the robustness of results.

★ **Results:** In the primary result produced by a logistic regression with adjustment for demographic factors, smoking appeared to increase the risk of heart disease with statistical significance (odds ratio 2.18, 95% confidence interval 1.47-3.23). No evidence of multiplicative interaction of smoking and gender was observed, and additional adjustment for examination indicators attenuated estimates. Results in matched sample showed little difference than in total sample. Mediation analysis suggested that around 12.9-16.3% of the effect was mediated through cholesterol.

★ **Conclusions:** Smoking exposure showed a strong and significant effect on increasing the risk of heart disease. A considerable proportion of this effect was mediated through elevating cholesterol level.

★ **Keywords:** Smoking, Heart disease, Cholesterol, Propensity score, E-value.

Background

Heart disease remains a significant public health concern worldwide, contributing to substantial morbidity and mortality. According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people)[1].

Numerous factors have been identified as potential contributors to the development of heart disease, ranging from lifestyle choices to genetic predispositions. About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking.[2] Other key indicators include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Among these factors, smoking has long been recognized as a leading modifiable risk factor associated with cardiovascular ailments. [3] Understanding the causal relationship between smoking and heart disease is of utmost importance, as it can inform preventive strategies and public health interventions.

The detrimental effects of smoking on various aspects of health have been extensively studied [3, 4], with rich evidence supporting its association with lung cancer [5], chronic obstructive pulmonary disease (COPD) [6], and other respiratory disorders [7]. How-

ever, the intricate mechanisms linking smoking to heart disease have been a subject of ongoing investigation. While observational studies have consistently shown a positive correlation between smoking and the incidence of heart disease [3, 4, 8], establishing a causal relationship necessitates rigorous causal inference methodologies that account for confounding variables and potential biases.

This study aimed to investigate the relationship between smoking and heart disease by employing causal inference techniques and rigorous study design. Since the smoking rate and the risk of heart disease are not consistent in women and men, this study examined whether the relationship varied by gender. Furthermore, observing in smokers a higher cholesterol level, which is a key risk factor for heart disease, the degree to which the effect of smoking is mediated through it was also studied.

By carefully designing analysis scheme and addressing confounding factors, this study attempts to find robust conclusions that can inform medical professionals, policymakers, and individuals seeking to reduce their risk of developing heart disease.

Methods

Dataset

For this study, I chose the Stroke Prediction Dataset from Kaggle, which is publicly accessible via `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset` [9]. The dataset consists of 5,110 patient records (276 presenting with heart disease and 3,834 not presenting with heart disease) and contains 12 features. In my study, the outcome of interest is the presence of heart disease, and the exposure is the smoking status.

Covariates

The baseline covariates were divided to 2 parts. There were **6 demographic factors** including: the gender (male, female or other), age (years), ever married (yes or no), work type (private, self-employment, government job, never worked or children), residence type (urban or rural), and the body mass index, and **4 physical examination indicators** including: the presence of hypertension (1:yes, 0: no), stroke (1:yes, 0: no), the average blood glucose level (mg/dL), and the cholesterol level (mmol/L). The remained 2 features are the smoking status (never smoked, formerly smoked or smokes), which was the exposure I studied, and the presence of heart disease (1: yes, 0: no), the outcome of interest.

Analytical Sample

Before analysis, I conducted a sanity check on the original data to select the analytical sample. I found that the dataset had some outliers and missing data. For example, there was one patient of gender 'other', and some patients had BMI greater than 50, which were likely to be mistakenly recorded. Moreover, some children was recorded as 'smokes' or 'formerly smoked'.

For the validity of my analytical data, I excludes the patient of gender 'other', and the patients with BMI larger than 50 (this is an empirically chosen number). To dichotomize the exposure, I excluded the patients with smoking status 'formerly smoked' or 'unknown'. Then, 'smokes' and 'never smoked' were encoded as 1 and 0, respectively. Because the exposure was longitudinal, and heart diseases observed in children are usually congenital (not caused by smoking), I also excluded the patients of age smaller than 18, so that my sample consisted of only adults.

After sample selection, the final analytical data included 2,398 patients, with 135 presenting of heart disease.

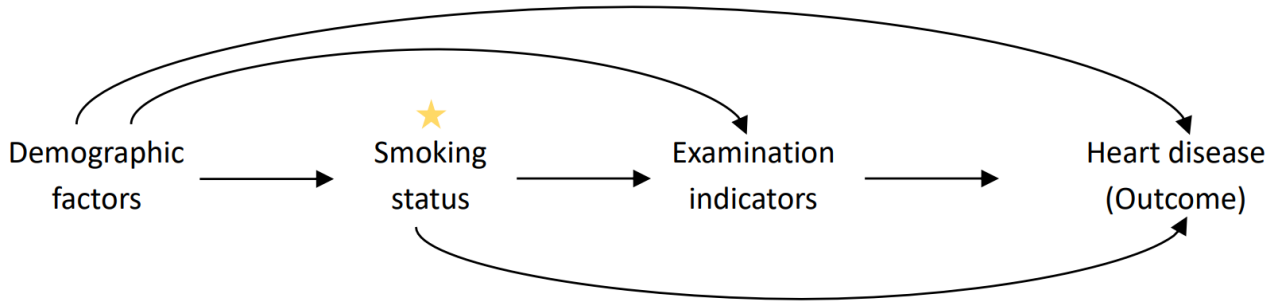
Causal DAG

In the analytical sample, the age of patients ranged from 18 to 65 years, all in their adult stage. Hence, their demographic factors were relatively stable, compared with other covariates such as physical examination indicators. These demographic factors influence other covariates directly or indirectly.

According to some prior knowledge, the smoking status may affect both the examination indicators and the outcome of interest, here, the presence of heart disease. In this path, the presented examination indicators can be the mediators between the smoking status (exposure) and the presence of heart disease (outcome).

With the discussion above, I drew a simplified causal directed acyclic graph (DAG) in this study to conceptualize the relationships among the recorded covariates, as is shown in Figure 1.

In this DAG, controlling for the demographic factors suffices to eliminate the confounding between smoking status and the presence of heart disease. Conversely, conditioning on the potential mediators, here, the physical examination indicators, may introduce bias.



- Demographic factors: age, gender, marital status, work type, residence type, BMI;
- Smoking status: the exposure of interest, 1 (yes) or 0 (no);
- Examination indicators: hypertension, stroke, average blood glucose, cholesterol;
- Heart disease: the outcome of interest, 1 (yes) or 0 (no).

Figure 1. A simplified DAG underlying this study

Statistical Analysis

The main statistical methods I applied in this study include regression analysis, effect modification with stratification, propensity score matching and mediation analysis. Sensitivity analyses were also conducted to evaluate the validity of these methods.

The baseline covariates in the analytical sample and by categories of exposure (the smoking status) were compared. The result is shown in table 1.

In the tabular analysis, the **incidence odds ratio (IOR)** of heart disease in the smoking group was calculated, compared with the non-smoking group. Note this result is unadjusted. To examine the effect modification by gender, the unadjusted IRRs stratified by gender is also reported.

To find the covariate-adjusted effect of smoking status in sense of odds ratio, we fitted logistic regression models, which was assessed using the `glm()` function in R. To show how various sets of covariates impacted exposure estimates when added to the model, I created two regression models. Model 1 adjusted for demographic factors, which was able to account for confounding shown in the previous DAG. Model 2 adjusted for all the covariates included in model 1 and additionally adjusted for physical examination indicators including hypertension, stroke, average blood glucose level and cholesterol level. For effect modification, regression models were fitted within subsamples stratified by gender.

Between the two groups (smoking or not), I observed imbalance across covariates, which might undermine my analysis. To deal with the imbalance, I fitted a propensity score model and applied propensity score matching on the analytical sample. The definition of propensity score is the probability that a participant was in the exposure group, conditional on her/his covariates. Formally,

$$e_i(C_i) = \mathbb{P}(A_i = 1|C_i),$$

where A is the exposure (smoking status). The propensity score of the patients were fitted by a logistic regression model, as an estimation of exposure probability. I only included demographic factors in the propensity score models, since physical examination indicators were not identified as confounders in the underlying DAG.

Based on the estimated scores, I applied one-to-one matching without replacement on the analytical sample. I checked the covariate balance before and after matching to evaluate the effectiveness of the matching approach. To estimate the causal effect of smoking status within the matched exposure group (also referred to as average treatment effect in the treatment group, ATT), a regression model adjusted for demographic factors was fitted over the matched sample as a doubly robust approach.

Mediation analysis was also applied in this study. The original dataset included some examination indicators including hypertension, stroke, cholesterol and blood glucose, which might lie on the causal paths from smoking to the presence of heart disease. I fitted regression models to find the relevance between smoking status and these potential mediators. Logistic regression was applied on dichotomous indicators, while linear was applied on continuous indicators. The examination indicator with significant correlation to both smoking status and heart disease was selected to conduct mediation analysis.

Along the exposure-mediator-outcome path, the natural direct effect (NDE) and natural indirect effect (NIE) were calculated by the formula given in Vander-Weele and Vansteelandt, 2010 [10]:

$$\begin{aligned}&\log NDE(a,a^*;a^*|c)\\&=\left(\theta_1+\theta_3\left(\beta_0+\beta_1a^*+\beta_2^\top c+\theta_2\sigma^2\right)\right)(a-a^*)\\&\quad+0.5\,\theta_3\sigma^2(a^2-a^{*2}),\\&\log NIE(a,a^*;a|c)=(\theta_2\beta_1+\theta_3\beta_1a)(a-a^*);\end{aligned}$$

with the exposure A of interest and covariates C, the mediator M is fitted by

$$\mathbb{E}[M|a,c]=\beta_0+\beta_1a+\beta_2^\top c,$$

and the outcome Y is fitted by

$$\text{logit}\left\{\mathbb{P}(Y=1|a,m,c)\right\}=\theta_0+\theta_1a+\theta_2m+\theta_3am+\theta_4^\top c.$$

This helped to evaluate the degree to which the effect of smoking was mediated by a specific physical indicator, with the outcome event being rare, and the mediator M being normally distributed. I tested the normality of selected mediator to verify the validity of my estimate. Both cases with and without interaction between exposure and mediator were included in my analysis, and the P-value for interaction was also reported. Bootstrapping was used to calculate an approximate 95% confidence interval of those effects.

In sensitivity analysis, E-values (proposed by Vanderweele and Ding, 2017, [11]) were calculated to evaluate the robustness of my results if there had been confounding from unmeasured covariates. I also compared the E-value with a measured strong confounder in this study to show whether my conclusion would be undermined assuming the strong confounder have been unmeasured.

All my analysis were conducted in R, version 4.1.2. Through this study, statistical significance level was defined as $P < 0.05$.

Table 1. Distribution of Baseline Covariates Across Categories of Smokers and Non-smokers

	Smoking status		
	Smokes (n=716)	**Never smoked (n=1682)**	**Total analytical sample (n=2398)**
Gender, No. (%)			
Female	409 (57.1)	1109 (65.9)	1518 (63.3)
Male	307 (43.9)	573 (35.1)	880 (36.7)
Age, mean (SD), years	47.4 (16.0)	49.1 (18.2)	48.6 (17.6)
Ever married, No. (%)	562 (78.5)	1301 (77.3)	1863 (77.7)
Work type, No. (%)			
Private	491 (68.6)	1104 (65.6)	1595 (66.5)
Self-employed	113 (15.8)	310 (18.4)	423 (17.6)
Government job	112 (15.6)	265 (15.8)	377 (15.7)
Never worked	0 (0.0)	3 (0.2)	3 (0.1)
Residence type, No. (%)			
Rural	323 (45.1)	844 (50.2)	1167 (48.7)
Urban	393 (54.9)	838 (49.8)	1231 (51.3)
BMI[1], mean (SD), $\text{kg}\cdot\text{m}^{-2}$	30.2 (6.3)	29.8 (6.4)	29.9 (6.3)
Hypertension, No. (%)	79 (11.0)	209 (12.4)	288 (12.0)
Stroke, No. (%)	39 (5.4)	83 (4.9)	122 (5.1)
Average blood glucose, mean (SD), $\text{mg}\cdot\text{dL}^{-1}$	106.4 (45.9)	107.3 (47.0)	29.9 (6.3)
Cholesterol, mean (SD), $\text{mmol}\cdot\text{L}^{-1}$	4.75 (0.56)	4.65 (0.54)	4.68 (0.55)

[1] Abbreviations: BMI, body mass index, calculated as weight in kilograms divided by height in meters squared.

Results

This analysis involved 2,398 patients (age range, 18-65 years; 880 males [36.7%]). At the baseline, compared with patients who had no smoking history, those smoked at the survey period had a smaller age, had higher BMI, lower average blood glucose and higher cholesterol level. Compared with non-smokers, they reported a higher proportion of stroke history but a lower proportion of hypertension. Furthermore, these smokers were more likely to be male, have ever been married, have private work type, and live in urban area. Descriptive statistics within strata of smoking status are available in Table 1.

Table 2. Heart Disease Incidence Odds Ratios Based on Smoking Status

Smoking status exposure group and subsample	No. of Heart disease	No. of patients	Incidence (%)	Unadjusted IOR (95% CI)[1]
Total sample	135	2398	5.63	
Smokes	55	716	7.68	1.67 (1.17-2.38)
Never smoked	80	1682	4.76	1 [Reference]
Female patients	60	1518	3.95	
Smokes	23	409	5.62	1.73 (1.01-2.94)
Never smoked	37	1109	3.34	1 [Reference]
Male patients	75	880	8.52	
Smokes	32	307	10.42	1.43 (0.89-2.32)
Never smoked	43	573	7.50	1 [Reference]

[1] Abbreviations: IOR, incidence odds ratio, computed by dividing incidence odds in smokers by that in non-smokers; CI, confidence interval, computed by likelihood ratio test.

Table 3. Association between Smoking Status and Heart Disease in the Total Sample and Within Strata of Gender

Smoking status exposure group and subsample	Unadjusted		Model 1 OR[1](95% CI)	P for interaction[2]	Model 2 OR (95% CI)	P for interaction
	No. of patients	Incidence (%)				
Total sample	2398	5.63				
Smokes	716	7.68	2.18 (1.47-3.23)		1.98 (1.32-2.97)	
Never smoked	1682	4.76	1 [Reference]		1 [Reference]	
Female patients	1518	3.95				
Smokes	409	5.62	2.85 (1.57-5.16)		2.67 (1.45-4.92)	
Never smoked	1109	3.34	1 [Reference]		1 [Reference]	
Male patients	880	8.52				
Smokes	307	10.42	1.80 (1.06-3.04)	0.250	1.59 (0.92-2.74)	0.195
Never smoked	573	7.50	1 [Reference]		1 [Reference]	

[1] Abbreviations: OR, odds ratio, estimated by a logistic regression model;

[2] P for interaction: the P value for the cross-product term for the exposure and strata variable (smoking status $\times$ gender, where female and never smoked were the reference groups).

[3] Reference groups in model 1: female, not ever married, private work and live in rural area. Reference groups in model 2: groups in model 1 plus no hypertension and no stroke.

In tabular analysis, in the total sample and within strata of gender, increased incidence odds ratio (IOR) was observed in the smokers compared with patients having no smoking history (Table 2). This result is statistically significant. Between gender strata, incidence odds ratio was highest in female smokers (1.73, 95% 1.01-2.94). Looking at the lower bound of IOR, I found that the significance of effect was weaker within both strata (In female, 1.01 and in male, 0.89). One possible reason was the shrinkage of sample size within each stratum.

In the total sample, adjusting for baseline demographic covariates (Table 3, Model 1), the estimated odds ratio (OR) of heart disease for smokers was 2.18 (95% CI, 1.47-3.23). Additional adjustment for physical examination indicators attenuated estimates (OR, 1.98, 95 % CI, 1.32-2.97), since these indicators could act as mediators along the causal path from smoking status to heart disease.

In analysis stratified by gender, smoking exposure for both female and male patients was associated with an elevated risk of heart disease. The effect of smoking was stronger and statistically significant among female patients (OR, 2.85, 95% CI, 1.57-5.16); however, there was no evidence of multiplicative interaction by gender ($P = 0.250$). Similar result was observed with additional adjustment for physical examination indicators, where the OR reduced in both strata.

Table 4. Covariate Balance before and after Propensity Score Matching

	Before Matching				**After Matching**			
	Smokes	**Never smoked**	**P-value[1]**	**SMD[2]**	**Smokes**	**Never smoked**	**P-value**	**SMD**
Size	716	1682			716	716		
	Mean (SD)				Mean (SD)			
gender	0.43 (0.50)	0.34 (0.47)	< 0.001	0.182	0.43 (0.50)	0.41 (0.49)	0.593	0.028
age	47.43 (16.02)	49.14 (18.25)	0.030	0.099	47.43 (16.02)	46.48 (17.86)	0.287	0.056
BMI	30.20 (6.31)	29.77 (6.36)	0.133	0.067	30.20 (6.31)	29.99 (6.52)	0.532	0.033
ever married	0.77 (0.42)	0.78 (0.41)	0.539	0.028	0.76 (0.43)	0.78 (0.41)	0.313	0.053
residence	0.50 (0.55)	0.55 (0.50)	0.023	0.102	0.56 (0.50)	0.55 (0.50)	0.791	0.014
work type	No. (%)		0.269	0.095	No. (%)		0.479	0.064
private	491 (68.6)	1104 (65.6)			491 (68.6)	512 (71.5)		
goverment	112 (15.6)	265 (15.8)			112 (15.6)	101 (14.1)		
self-employed	113 (15.8)	310 (18.4)			113 (15.8)	103 (14.4)		
never worked	3 (0.2)	0 (0.0)			0 (0.0)	0 (0.0)		

[1] P-values were accessed in contingency tests.

[2] SMD is standardized mean difference, which is computed by $\frac{|\mu_t - \mu_c|}{\sqrt{(s_t^2 + s_c^2)/2}}$, of which μ_t,μ_c are intra-group means, and s_t,s_c the standard deviations.

Table 5. Association between Smoking Status and Heart Disease in the Matched Sample

Smoking status exposure group	No. of Heart disease	Incidence (%)	OR (95% CI)	
			No Adjustment	With Adjustment[1]
Matched Sample	85	5.94		
Smokes ($n = 716$)	55	7.68	1.90 (1.20-3.01)	2.26 (1.37-3.70)
Never smoked ($n = 716$)	30	4.19	1 [reference]	1 [reference]

[1] The model adjusted for demographic factors: gender, age, BMI, ever married, work type and residence type.

One-to-one propensity score matching without replacement was applied to create groups of smokers and non-smokers who were similar with respect to observed covariates. After matching, all smokers ($n = 716$) and 716 paired non-smokers were included in the matched dataset.

The distribution of covariates before and after propensity score matching is shown in Table 4. As is shown, after matching, the standardized mean difference was no greater than 0.064, and all covariates showed no significant difference between the two groups. Hence, a satisfactory balance was achieved.

In the propensity score matched sample, a remarkable increase in the risk of heart disease was observed in smokers (OR, 1.90; 95% CI, 1.20-3.01; Table 5), which was statistically significant compared with non-smokers. Adjustment for demographic factors yielded a stronger effect (OR, 2.26; 95% CI, 1.37-3.70). Due to the reduction in sample size, this doubly robust approach gave a slacker 95% confidence interval, compared with the previous model 1 (OR, 2.18; 95% CI, 1.47-3.23) over the total sample.

The association of potential mediators (physical examination indicators) with smoking status and heart disease is shown in Table 6. Of the 4 potential mediators, only the cholesterol level showed strong association with both smoking status ($P = 3.29 \times 10^{-5}$) and heart disease ($P = 1.49 \times 10^{-5}$). In average, the blood cholesterol level of smokers was 0.1 $\text{mmol} \cdot \text{L}^{-1}$ greater than that of nonsmokers. Hence, the cholesterol level was most likely to be the mediator which was measured in this study.

I checked the normality of cholesterol level by smoking status. The result is shown in Figure 2. In the QQ-plot, the quantiles cholesterol level in both smokers and nonsmokers were well aligned to the theoretic quantiles of standard normal distribution. The density plot revealed that the distribution of cholesterol level was approximately normal in both smokers and nonsmokers.

Table 6. Association of Physical Examination Indicators with Smoking Status and Heart Disease

Physical Examination Indicators	Association with Smoking Status[1]		Association with Heart Disease[2]	
Dichotomous	**OR [95% CI]**	P-**value**	**OR [95% CI]**	P-**value**
Hypertension	0.97 [0.77, 1.38]	0.841	0.95 [0.60, 1.51]	0.820
Stroke	1.53 [1.01, 2.35]	0.047	1.49 [0.86, 2.59]	0.153
Continuous	**Coef. [95% CI]**	P-**value**	**OR [95% CI]**	P-**value**
Average blood glucose	-0.53 [-4.49, 3.43]	0.792	1.002 [0.998, 1.005]	0.272
Cholesterol	0.100 [0.053, 0.147]	3.29×10^{-5}	2.34 [1.59, 3.43]	1.49×10^{-5}

[1] The association between physical examination indicators (as mediator) and smoking status (as exposure) was assessed by fit a regression from exposure to mediator. For dichotomous mediators, logistic regression was applied, and for continuous mediators, linear regression was applied.

[2] The association between physical examination indicators and heart disease was assessed by the previous model 2, which adjusted for both demographic factors and physical examination indicators.

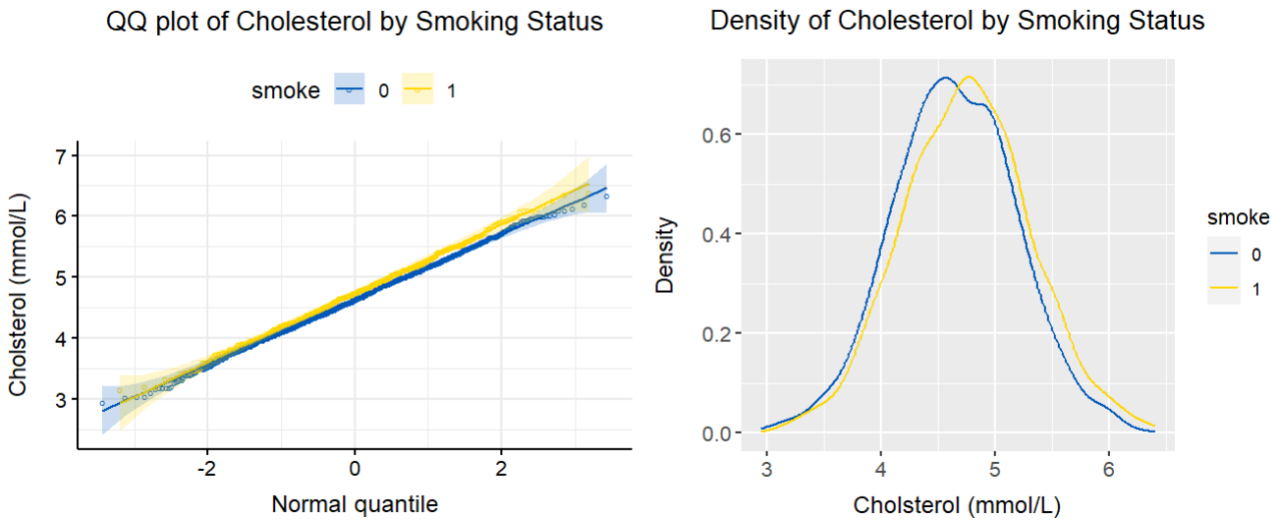


Figure 2. QQ-plot and Density of Cholesterol Covariate by Smoking Status

Fixing the cholesterol level as the mediator of interest, the estimated NDE and NIE of smoking in the total sample were shown in Table 7. With a multiplicative interaction between smoking status and cholesterol level included in the outcome model, an increased risk of heart disease was observed in smokers, in sense of both direct and indirect effects. The total effect made little difference from result produced by model 1. Furthermore, in the total effect of smoking on heart disease, 12.9% was mediated by elevating the cholesterol level. However, the multiplicative interaction of smoking and cholesterol was not significant ($P = 0.354$).

The model without interaction gave a similar result, with a slightly higher NIE and lower NDE. In this model, a proportion of 16.3% of the effect was mediated through cholesterol.

The estimates of effect of smoking on heart disease produced by the previous methods were summarized in Table 8. For sensitivity analysis, I also reported the corresponding E-values to evaluate the robustness. As is shown, all these methods produced statistically significant results in the level of 0.05, indicating a strong association between smoking and heart disease. (I do not use the term "causal" here because this is an observation study, where I have only very limited number of covariates, and no randomization.) To evaluate the minimum strength of an unmeasured confounder which could explain away association and significance, both the E-values of OR and lower bound of 95% confidence interval were reported.

Table 7. Natural Direct and Indirect Effects in the Total Sample, with Cholesterol Level Fixed as the Mediator

Outcome Model	Odds Ratio	95% CI[2]	Proportion Mediated (%)[3]	P for interaction
With Interaction[1]			12.9	0.354
Natural direct effect	2.01	[1.32, 3.02]		
Natural indirect effect	1.07	[1.01, 1.17]		
Total Effect	2.16	[1.43, 3.23]		
No Interaction			16.3	
Natural direct effect	1.95	[1.32, 3.13]		
Natural indirect effect	1.10	[1.01, 1.16]		
Total Effect	2.14	[1.41, 3.36]		

[1] In the case of interaction, a cross-product term for the exposure and mediator (smoking status × cholesterol) was included in the regression model.

[2] The confidence interval was estimated by bootstrapping, hence was approximate. I ran 1,000 simulations in bootstrap, and construct a percentile confidence interval for each estimand.

[3] The proportion mediated on the risk difference scale is calculated by $\frac{NDE \times (NIE-1)}{NDE \times NIE-1}$.

Table 8. Sensitivity Analysis: Total Effect of Smoking on Heart Disease and E-value

Model	Total OR	95% CI	E-value (Lower bound)[2]
Unmatched, Adjusted for De.	2.18	[1.47, 3.23]	3.78 (2.30)
Unmatched, Adjusted for De. + Ex.	1.98	[1.32, 2.97]	3.37 (1.97)
Matched, Unadjusted	1.90	[1.20, 3.01]	3.21 (1.69)
Matched, Adjusted for De.	2.26	[1.37, 3.70]	3.95 (2.08)
Mediated by Chol., Adjusted for De. + Chol. × Smoke	2.16	[1.43, 3.23]	3.74 (2.21)
Mediated by Chol., Adjusted for De.	2.14	[1.41, 3.36]	3.70 (2.17)

[1] Abbreviations: De. for demographic factors (gender, age, ever married, work type, residence type, BMI), Ex. for examination indicators (hypertension, stroke, average blood glucose, cholesterol), Chol. for cholesterol;

[2] The calculation of E-value was discussed in VanderWeele and Ding, 2017: When the outcome (here, the presence heart disease) is relatively rare (e.g., <15%) by the end of follow-up, the E-value formula of OR (odds ratio) is the same as RR (risk ratio), that is, E-value = $RR + \sqrt{RR \times (RR-1)}$.

Here I take the adjusted model 1 on the total sample as the primary result, the interpretation for other models is analogous. In the primary result, I observed an odds ratio of 2.18, with the 95% confidence interval being $[1.47, 3.23]$. To explain away this estimate, an unmeasured confounder had to be associated with both smoking status (the exposure) and heart disease (the outcome) by an odds ratio of at least 3.78 folds. Also, an unmeasured confounder associated with both the exposure and the outcome by an odds ratio of 2.30 folds could move the 95% confidence interval to include the null, hence impact on the significance of my result.

An immediate question is that how strong such a confounder is, and how possibly it exists in this study. Take a measured confounder as an example. While fitting model 1 (adjusted for demographic factors) and propensity score model, I found that gender was a very strong confounder in this study. In model 1, it produced an odds ratio of 2.22 ($P = 3.06 \times 10^{-5}$) on the presence of heart disease, and in propensity score model, it produced an odds ratio of 1.46 ($P = 3.80 \times 10^{-5}$) on smoking status. Both the two odds ratios did not exceed the E-value (3.78 for the estimate, and 2.30 for the lower bound of 95% CI). Therefore, even an unmeasured confounder as strong as gender in this study would not change the conclusion and impact the significance of my estimate. This reflects the robustness of my result.

Discussion

In this study, smoking exposure was associated with higher risk of heart disease in logistic regression model adjusted for baseline demographic confounders. Estimates attenuated with additional adjustment for physical examination indicators, although, as noted in the Methods section, these indicators might lie on the pathway through which smoking influences the presence of heart disease. A higher odds ratio of heart disease was observed in female smokers after stratification by gender, though the interaction between smoking and gender was not significant in my analysis.

Matching as a non-parametric method was applied in this study, and the result revealed a higher risk of heart disease in smokers, compared with the case where they had not been exposed to smoking. This effect would be causal if the covariates included in this study had sufficed to account for confoundings.

Mediation analysis gave similar estimates of total effect of smoking on heart disease. A proportion of 12.9% - 16.3% was mediated through cholesterol, suggesting that cholesterol might be a considerable risk factor lying on the pathway from smoking to heart disease. This finding helps researchers to investigate the mechanism that how tobacco impacts cardiovascular health.

This study has several strengths. (a) Diverse approaches and methods were applied to estimate the effect of smoking on heart disease, and I also sequentially adjusted for different covariates to show how various sets of covariates impacted exposure estimates when added to the regression model. My findings were robust across different methods and models with similar point estimates and confidence intervals. (b) Based on some prior but common knowledge, an underlying causal DAG was proposed in my analysis. This DAG helped researchers to see into the structure of confounding and mediation, compared with traditional statistical approaches. (c) Stratification on gender examined whether the effect varied by sex and assisted

in eliminating confounding to some degree, while mediation analysis aided in an attempt to distinguish the effect mediated through some known risk factor. (d) Sensitivity analysis was conducted to evaluate the robustness of my findings. It was shown that even an unmeasured confounder as strong as gender would neither change my conclusion nor impact on the significance of my result.

Limitations

This study has several limitations. (a) Although robust results were obtained across different methods and sensitivity analysis, unmeasured confounding is always a flaw which may undermine the causality. Here only 10 covariates were collected in the dataset I used, but baseline variables related to smoking and heart disease could be far more than this quantity. (b) The causal DAG was empirical and oversimplified, hence it might not reflect the true structure of variables, which could be very complicated. For example, I regarded BMI as a baseline demographic factor, but it is possibly influenced by smoking. (c) The sample selection criteria restricted the analytical sample size and may have introduced slight selection bias. Even if my result is causal, its generalizability is limited. (d) This is an observational study, in which the used data were collected at a time point. However, smoking, as the exposure of interest in this study, is longitudinal. In my analysis, the smoking status was dichotomized, with little longitudinal information included. Moreover, exposures of different doses and frequencies were indistinguishably classified as "smokes". Such a dichotomization would have underestimated the effect in my analysis. (e) Heart disease, as the outcome of interest, could occur at unobserved time points, hence the collected data may not reflect the real risk of heart disease of the patients.

On ethic grounds, randomization in an analogous research is infeasible. Though a number of limitations exist in this study, I have done my best to exploit the available data and conduct my analysis. In further research, some improvements can be made by accounting for more potential confounders and collecting data in a longitudinal manner. Approaches in survival analysis can be utilized to estimate the hazard ratio (HR) of heart disease for patients with different exposing status. Also, instrumental variables (IVs) can be applied in these studies. Considering the key assumptions of IV (relevance, exclusion restriction and exogeneity), some genetic characteristics such as SNP (single-nucleotide polymorphism) on a certain allele could be selected as IV. Such a selection needs prior knowledge from genetics and molecular biology. Variables related tobacco market conditions are also top candidates since they are likely to affect the smoking behavior, but unlikely to influence the presence of heart disease in a direct way.

Conclusions

Smoking exposure showed a strong and significant effect on increasing the risk of heart disease in this study. A higher cholesterol level was associated with both smoking and the presence of heart disease. From the effect of smoking on heart disease, a considerable proportion was mediated through elevating cholesterol level. If causal, these findings may provide reference for public health interventions aimed at reducing the burden of heart disease and promoting healthier lifestyles.

References

[1] Centers for Disease Control and Prevention. Heart Disease Facts. https://www.cdc.gov/heartdisease/facts.htm. Accessed June 10, 2023.

[2] Fryar CD, Chen TC, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010. NCHS Data Brief. 2012 Aug;(103):1-8. PMID: 23101933.

[3] Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, Baker-Smith CM, Beaton AZ, Boehme AK, Buxton AE, Commodore Mensah Y, Elkind MSV, Evenson KR, Eze-Nliam C, Fugar S, Generoso G, Heard DG, Hiremath S, Ho JE, Kalani R, Kazi DS, Ko D, Levine DA, Liu J, Ma J, Magnani JW, Michos ED, Mussolino ME, Navaneethan SD, Parikh NI, Poudel R, Rezk-Hanna M, Roth GA, Shah NS, St-Onge M-P, Thacker EL, Virani SS, Voeks JH, Wang N-Y, Wong ND, Wong SS, Yaffe K, Martin SS; on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2023 update: a report from the American Heart Association [published ahead of print January 25, 2023]. Circulation. doi: 10.1161/CIR.0000000000001123

[4] U.S. Department of Health and Human Services. The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2014.

[5] Walser T, Cui X, Yanagawa J, Lee JM, Heinrich E, Lee G, Sharma S, Dubinett SM. Smoking and lung cancer: the role of inflammation. Proc Am Thorac Soc. 2008 Dec 1;5(8):811-5. doi: 10.1513/pats.200809-100TH. PMID: 19017734; PMCID: PMC4080902.

[6] Laniado-Laborín R. Smoking and chronic obstructive pulmonary disease (COPD). Parallel epidemics of the 21 century. Int J Environ Res Public Health. 2009 Jan;6(1):209-24. doi: 10.3390/ijerph6010209. Epub 2009 Jan 9. PMID: 19440278; PMCID: PMC2672326.

[7] Zeng X, Ren Y, Wu K, Yang Q, Zhang S, Wang D, Luo Y, Zhang N. Association Between Smoking Behavior and Obstructive Sleep Apnea: A Systematic Review and Meta-Analysis. Nicotine Tob Res. 2023 Feb 9;25(3):364-371. doi: 10.1093/ntr/ntac126. PMID: 35922388; PMCID: PMC9910143.

[8] Mons U, Müezzinler A, Gellert C, Schöttker B, Abnet CC, Bobak M, de Groot L, Freedman ND, Jansen E, Kee F, Kromhout D, Kuulasmaa K, Laatikainen T, O'Doherty MG, Bueno-de-Mesquita B, Orfanos P, Peters A, van der Schouw YT, Wilsgaard T, Wolk A, Trichopoulou A, Boffetta P, Brenner H; CHANCES Consortium. Impact of smoking and smoking cessation on cardiovascular events and mortality among older adults: meta-analysis of individual participant data from prospective cohort studies of the CHANCES consortium. BMJ. 2015 Apr 20;350:h1551. doi: 10.1136/bmj.h1551. PMID: 25896935; PMCID: PMC4413837.

[9] Federico S. Stroke Prediction Dataset. Retrieved June 5, 2023 from `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`.

[10] Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol. 2010 Dec 15;172(12):1339-48. doi: 10.1093/aje/kwq332. Epub 2010 Oct 29. PMID: 21036955; PMCID: PMC2998205.

[11] VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med. 2017 Aug 15;167(4):268-274. doi: 10.7326/M16-2607. Epub 2017 Jul 11. PMID: 28693043.

Supplementary Contents

Some additional figures about propensity score matching are shown in this section. Figure 3 shows the distribution of propensity scores in matched and unmatched samples in a jitter plot. Figure 4 shows the joint distribution of different covariates and propensity scores in smokers and non-smokers. Lowess smoothing were applied to all these plots.

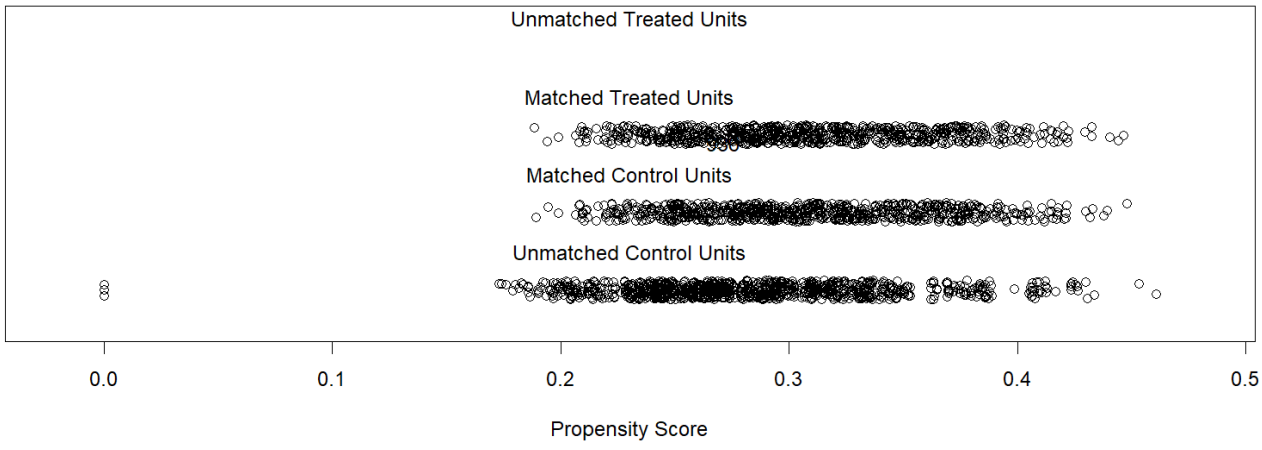


Figure 3. The Distribution of Propensity Score in Smokers and Nonsmokers

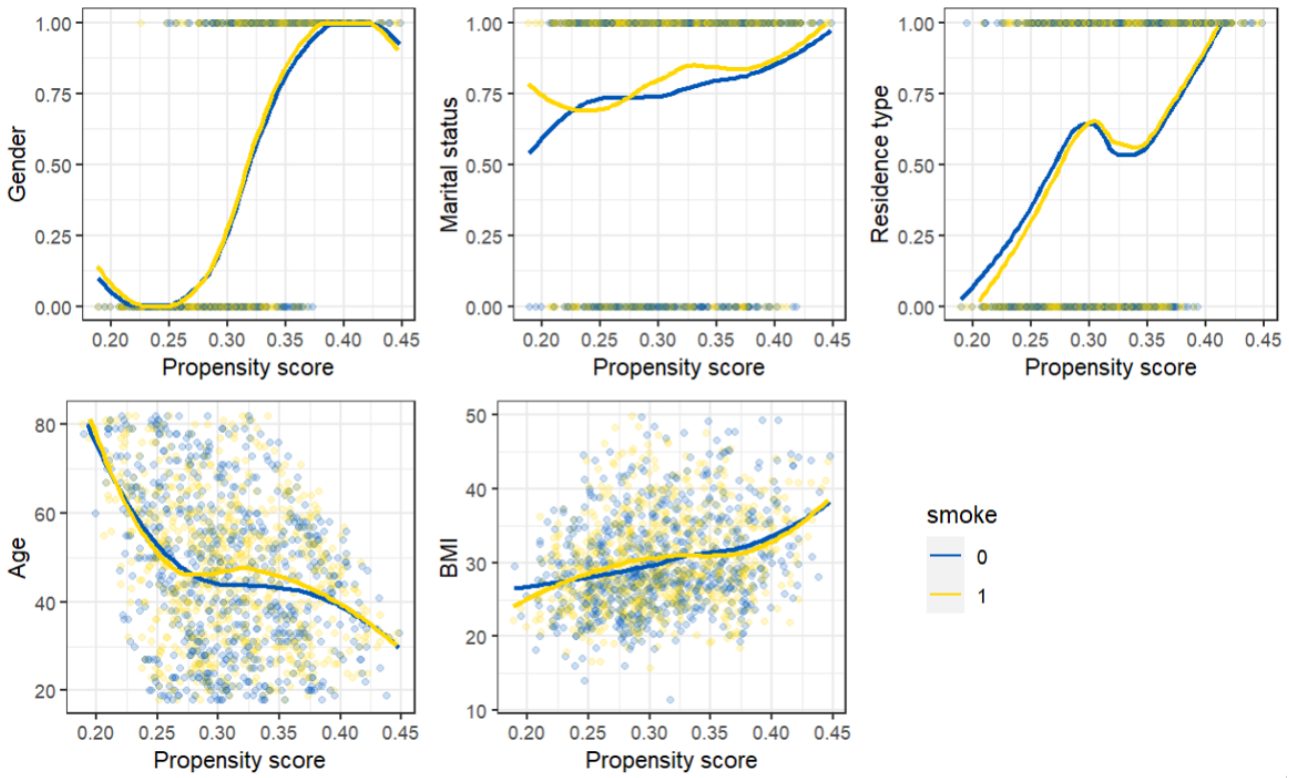


Figure 4. The Distribution of Covariates Stratified by Propensity Score