Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/44/20240424

Effect of smoking on lung cancer: A causal inference approach

Ziyang Qiu[1,6,7,†], Tianyi Ge[2,8,†], Zhaopeng Zhang[3,9,†], Zhenchen Lu[4,10,†], Louis Cao[5,6,11,†]

[1]School of Computer Science & Technology, Hua Zhong University of Science and Technology, Wuhan, 430074, China
[2] Hubei University of Technology, Wuhan, 430070, China,
[3]Keystone Academy, Beijing, 101300, China
[4]Northeastern University, Boston, 02115, China
[5]Wuxi BigBridge Academy, Wuxi, 214121, China
[6]Ziyang Qiu and Louis Cao are corresponding authors.

[7]1700445923@qq.com
[8]1398948593@qq.com
[9]jimzhang3n@gmail.com
[10]lzc2311330405@gmail.com
[11]louiscao2006@163.com
[†]These authors are co-first author.

Abstract. Smoking is associated with an increased chance of developing lung cancer. Three causal inference methods, backdoor adjustment, front-door adjustment, and counterfactual are used to analyze observational data on smoking, lung cancer, and related risks factors. Backdoor adjustment fails to allow for possible presence of unobserved confounders, which is merited by front-door adjustment. Counterfactual harnesses individual patient statistics to establish causal relationships between smoking and cancer on the individual level, so as to evaluate lung cancer risks after changes in individual smoking habits. Results by different methodology are in good agreement and showcase a strong causation between smoking and lung cancer at both group and individual level.

Keywords: Smoking, Lung Cancer, Backdoor Adjustment, Front-door Adjustment, Counterfactual

1. Introduction

The tobacco industry is expanding prosperously with rising smoking population and overall reaping profits from one third of adults worldwide [1]. Under the innocent wrapping lies the lethal danger that is wittingly or unwittingly overlooked: smoking is the predominant cause of lung cancer, resulting in 90 percent of male and 79 percent female lung cancer [1] and constituting more than 30 percent of lung cancer death [2]. After 1950, people started to realize the malicious impact of the smoking on human body. An increasing number of research on the effect of smoking on lung cancer has been conducted, focusing on different aspects such as chemical component of cigarettes and the statistical analysis of relative risk for different population groups.

DOI: 10.54254/2753-8818/44/20240424

Researchers found that the specific gene types interact with one another to either increase or decrease the risk. The effect is widely studied by analyzing the chemicals in cigarettes [3,4]. Though some biological mechanisms have been identified, the effect remains relatively elusive and informulable scientifically. Statistical analysis upon the effect of cigarettes consumption on lung cancer predisposition has been studied extensively, stretching to a large body of related work on smoking and cancer. In a comprehensive analysis by Grandini and colleagues using relative risk, smoking one more cigarette per day would occasion an increment of 7% in the risk of lung cancer in male and a slightly larger increment in female. [5] But in general the previous work only considers the correlation between the two variables and does not analyse the interaction between the causal components in the complex causal mechanism by which smoking ultimately affects lung cancer. Such a mechanism is essential for accessing the causal relationship because many variables, if not considered, may obscure the causality. For instance, the positive correlation between smoking and lung cancer may not imply that smoking causes cancer but rather that smokers usually have higher level of stress, which leads to cancer. The causal mechanism demanded remains untouched until recently.

Prior to our study, there have been studies upon the causation between smoking and lung cancer using causal analysis. Raghu and colleagues-built Lung Cancer Causal Model (LCCM) to access the effect of smoking by analyzing biological transformation within patients [6]. However, few studies had been comprehensive enough to include different methodologies to affect the causal relationship both on the macro and micro level.

We conducted our investigation using an international medical data set [7] on risk factors and symptoms of lung cancer. A summary of the data set is provided by Prof. Ahmad and Prof. Mayya at Al Andalus University for Medical Science [8]. We apply inferential methodologies to analyze the connection between smoking and lung cancer, using three major analytical tools in causal inference: backdoor adjustment, front door adjustment, and counterfactual.

2. Causal Graph

To assess the causal relationship between smoking (S) and lung cancer (C), we propose a causal graph (Figure 1), with adjoining nodes implying *direct causation*, that casts more thoroughly the relationship between the smoking and sufferance of cancer. Both personal and environmental reasons cause smoking and at the same time leads to lung cancer. Smoking also directs leads to cancer.

2.1. Personal Cause

Fatigue (F) is the "most prevalent cancer-related symptom" and hold direct leverage on a person's living quality and internal functions, according to Wagner and Cella [9]. Therefore, it points directly to lung cancer. Fatigue also leads to depression and anxiety, causing increased risk of alcohol consumption [10]. The relationship between alcohol and smoking is also evident. The study conducted by DiFranze and Guerrera found that 83% of alcoholics were smokers, whereas only 34% of the nonalcoholic subject's smoke. The cause can be attributed to individuals' genetic propensity for additive drug, tendency to behave irresponsibly, and social anxiety [11]. Also, alcohol users are less resistant to the first cigarette and are less able to quit smoking after initiation [12].

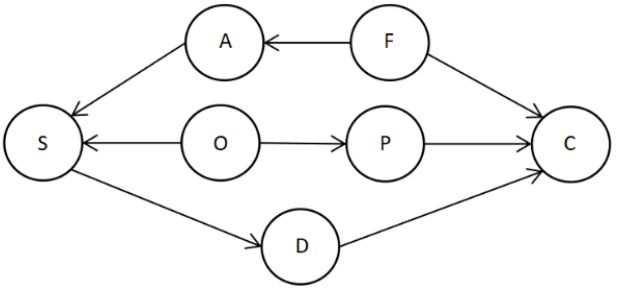

Figure 1. The causal graph of the effect of smoking on lung cancer. Variables represent: : smoking, : alcohol use, : fatigue, : occupational hazard, : passive smoking, chronic lung disease, : lung cancer.

DOI: 10.54254/2753-8818/44/20240424

2.2. Environmental Cause

Occupation hazard (O) refers to the risk of developing lung cancer based on the person's occupation. It has profound effect on smoking and passive smoking (P) intensity [13, 14]. Howard noted the discrepancy in the level of smoking between workers that blue-collar workers smoke significantly more. This is caused by insanitation, noise and the influence of smoking coworkers, according to Kim. The situation is further exasperated by the saliently higher level of "environmental tobacco smoke (ETS)" for manual workers, and a "higher risk of developing cancer *even if they are non-smokers.*" The overall situation conforms to the *fork structure* ($S \leftarrow O \rightarrow P$) and *chain structure* ($O \rightarrow P \rightarrow C$). The evident risk of passive smoking to lung cancer is also scientifically illustrated [15].

2.3. From Smoking to Cancer

Smoking directly causes an increased level of chronic lung disease (D) and then lung cancer. Durham and Adcock's work, for example, examines how smoking is the driving force for chronic obstructive pulmonary disease (COPD) and lung cancer. They assert that "Lung cancer and COPD may be different aspects of the same disease, with the same underlying predispositions" and alternatively COPD is "the driving factor in lung cancer". Therefore, smoking logically leads to chronic lung disease and to lung cancer [16]. Other studies also shows that chronic lung disease is an important risk factor for developing lung cancer, even after accounting for tobacco consumption [17,18]. Research by Daniels and colleagues' further points to genetic alternation to explain causality between lung disease and lung cancer [19].

3. Backdoor adjustment

Backdoor adjustment is a powerful analytic tool in causal inference where the relationship between the two variables, namely smoking and lung cancer, is not immediately clear. By conditioning on to be and adjusting for (fixing and summing over all combinations of) *confounding variables* that affect both the variable S and the outcome variable C, we can measure $\mathrm{P}(\mathrm{C}|\mathrm{S}=\mathrm{s})$. To perform backdoor adjustment, we identify the relevant confounding variables and adjust for them accordingly to assess the unperturbed causal effect of S on C.

Requirements for performing backdoor adjustment need to be met beforehand. The requirements are summarized in Pearl, Glymour and Jewell's *"Causal Inference in Statistics: A Primer"* and restated below [20].

The Backdoor Criterion

Define to be the set of observed variables that we condition on.

1. Z should block all backdoor paths from S to C. In our graph, F needs to be adjusted for so that A and C becomes *d-separated* (independent with each other), blocking the path $\mathrm{S} \leftarrow \mathrm{A} \leftarrow \mathrm{F} \rightarrow \mathrm{C}$. O needs to be adjusted for so that the path $\mathrm{S} \leftarrow \mathrm{O} \rightarrow \mathrm{P} \rightarrow \mathrm{C}$ is clogged.

2. All directed paths from S to C ($\mathrm{S} \rightarrow \mathrm{D} \rightarrow \mathrm{C}$) should be left undisturbed to ensure that we are not conditioning on descendants of S (D).

3. No new backdoor path should be created.

Fatigue, a non-descendant of S, needs to be adjusted to block the backdoor path $\mathrm{S} \leftarrow \mathrm{A} \leftarrow \mathrm{F} \rightarrow \mathrm{C}$. Casual effect of smoking on lung cancer by adjusting for variable F is shown below:

$$\mathrm{P}\big(\mathrm{C}=\mathrm{c}\big|\mathrm{do}(\mathrm{S}=\mathrm{s})\big)=\sum_{\mathrm{f}}\mathrm{P}(\mathrm{C}=\mathrm{c}|\mathrm{S}=\mathrm{s},\mathrm{F}=\mathrm{f})\ \mathrm{P}(\mathrm{F}=\mathrm{f}) \tag{1.1}$$

We then condition on to render variable S and P independent, blocking the backdoor path $S \leftarrow O \rightarrow P \rightarrow C$. We apply chain rule for this path:

$$\mathrm{P}\big(\mathrm{C}=\mathrm{c}\big|\mathrm{do}(\mathrm{S}=\mathrm{s})\big)=\sum_{\mathrm{o}}\mathrm{P}(\mathrm{C}=\mathrm{c}|\mathrm{S}=\mathrm{s},\mathrm{O}=\mathrm{o})\,\mathrm{P}(\mathrm{O}=\mathrm{o}) \tag{1.2}$$

Combining the two equations blocks all backdoor paths and gives the overall formula for backdoor adjustment:

DOI: 10.54254/2753-8818/44/20240424

$$P\big(C = c\big|do(S = s)\big) = \sum_{f}\sum_{o} P(C = c|S = s, F = f, O = o)\, P(F = f)P(O = o) \tag{1.3}$$

The normalized results (Table 1):

Table 1. P(C = c|do(S = s)) using backdoor adjustment (normalized)

	$(C = 1\|do(S))$	$(C = 2\|do(S))$	$(C = 3\|do(S))$
$S = 1$	0.554570	0.445430	0.000000
$S = 2$	0.312069	0.459429	0.228502
$S = 3$	0.556143	0.443857	0.000000
$S = 4$	0.761045	0.000000	0.238955
$S = 5$	0.000000	1.000000	0.000000
$S = 6$	0.500000	0.250000	0.250000
$S = 7$	0.017327	0.000000	0.982673
$S = 8$	0.000000	0.015770	0.984230

$P > 0.9$

$0.5 < P < 0.8$

$0.3 < P < 0.5$

$0.2 < P < 0.3$

$0.1 < P < 0.2$

$P = 0.00$

4. Front-door adjustment

The previous section provides a simple way to estimate causal effect from non-experimental data by identifying the ensemble of variables to condition on and using backdoor adjustment. However, in theoretical settings the existence of unobserved confounders cannot be denied entirely. Other factors not identified by the graph or unrecognized scientifically could potentially influence both the likeliness for one to smoke and suffer cancer, e.g., genotype, which we denote as. Because variable U is unobserved, the spurious path S ← U → C cannot be blocked by other variables in its path, i.e., the model does not meet the backdoor criterion. Therefore, we seek an alternative to perform conditioning operation--front-door adjustment--in this section to estimate P(C = c|do(S = s)). Note that S and C are d-separated in all the paths other than S → D → C. Therefore, we can apply Bayesian chain rule over the direct path S → D → C.

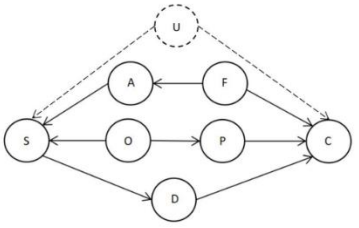

Figure 2. Causal graph - front-door criterion

Based on preliminary data processing, chronic lung disease strongly correlates with lung cancer, and the effect is not completely homologous within every specific value for D after considering the level of nicotine intake. By conditioning on S = s and thus separating all other paths from S to D, we evaluate the probability of D = d given S = s, as concretely shown in the formula below:

$$P\big(D = d\big|do(S = s)\big) = P(D = d|S = s) \tag{2.1}$$

With the given value of , we evaluate the probability of C = c given D = d, as expressed in the following formula (all backdoor path from to has been blocked):

$$P\big(C = c\big|do(D = d)\big) = \sum_{s} P(C = c|D = d, S = s)P(S = s) \tag{2.2}$$

DOI: 10.54254/2753-8818/44/20240424

The overall causal effect of S on:

$$P\left(C = c\middle|do(S = s)\right) = \sum_{d} P\left(C = c\middle|do(D = d)\right)P\left(D = d\middle|do(S = s)\right) \tag{2.3}$$

The right-hand part of equation (2.3) can be evaluated by replacing *do* operation with equation (2.1) and equation (2.2). The final expression without do-operation is:

$$P\left(C = c\middle|do(S = s)\right) = \sum_{d}\sum_{s'} P\left(C = c\middle|D = d, S = s'\right)P\left(S = s'\right)P(D = d|S = s) \tag{2.4}$$

Here is the normalized result (Table 2):

Table 2. $P(S = s|C = c)$ using front-door adjustment

chronic lung disease				
	$(C = 1\|do(S))$	$(C = 2\|do(S))$	$(C = 3\|do(S))$	
$S = 1$	0.470323	0.315736	0.213941	$P > 0.5$
$S = 2$	0.525727	0.326040	0.148233	
$S = 3$	0.509840	0.322384	0.167775	$0.4 < P < 0.5$
$S = 4$	0.468463	0.173802	0.357735	
$S = 5$	0.621599	0.202381	0.176020	$0.3 < P < 0.4$
$S = 6$	0.391589	0.304205	0.304205	
$S = 7$	0.270143	0.285136	0.444721	$0.2 < P < 0.3$
$S = 8$	0.251246	0.310587	0.438167	
				$0.1 < P < 0.2$

5. Counterfactual

5.1. Background Knowledge

We have researched the causal relationship between smoking and lung cancer by using backdoor and front-door adjustment in the former part. These two methods investigate the relationship at a group level. Another dimension is counterfactual, which focuses on individuals and targets at every specific person.

Counterfactual involves asking "if." For instance, "if" I had smoked more, to what extent will my lung cancer level change. It aims to compare the results of different hypothetical conditions. Notably, there is a fundamental difference between counterfactual and intervention. The former is bound to occur in "two different worlds": the proposed hypothetical conditions are necessarily disparate from those in the real world so as to ascertain what would be the result if someone had changed his or her behaviour while other confounding variables remain unchanged. To be specific, one matter of a world happens first, then we estimate the issue in the other world based on the result, and they cannot happen in the same world at the same time. But the latter does not refer to whatsoever to another world but only aims to determine the causality between two variables for the entire research sample. Thus, there exists some kinds of problems that can only be resolved by counterfactual instead of intervention. The inquiry which is at an individual level like "what would my lung cancer level alter if I have smoked more" can only be analyzed through using counterfactual.

DOI: 10.54254/2753-8818/44/20240424

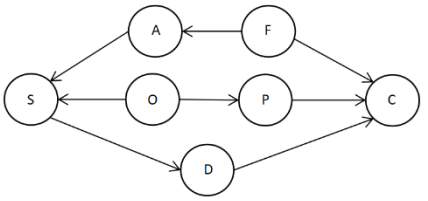

Figure 3. Causal Graph

Our work intends to search and verify the causal relationship between smoking ("S" in figure 3) and lung cancer ("C" in figure 3). In our data set, every person has an attribution representing their smoking level (S) ranging from one to eight, the greater the number is, the more they smoke. For an individual (smoking level s_0, lung cancer level c_0), we need to calculate how lung cancer level (C) changes if he or she had smoked more or less. Significant variation of lung cancer level can indicate that their truly have causal relationship. On the contrary, if lung cancer level remains constant or changes exceedingly slight, it would be apparent that they have no or extraordinarily inconspicuous extraordinarily. To better delineate the problem, we assume that when one's smoking level changed to s_1, the lung cancer level is c_1. Then we can write what we aim to compute in the following expression:

$$E\left(C_{s=s_1} \middle| C_{s=s_0} = c_0, s = s_0\right)$$

We use $C_{s=s1}$ and $C_{s=s0}$ to distinguish the actual level and the result on hypothetical condition to avoid confusion.

Our work is divided into four steps, regression analysis, abduction, action, and prediction, which will be stated thoroughly in the following paragraphs [20]. And before our work commences, we are supposed to assume all exogenous variables to be independent and that counterfactual on smoking does not affect the value of all exogenous variables.

5.2. Regression Analysis

First, we should do regression analysis. For every endogenous variable, we need to obtain function relationship of it if there existed a directed edge pointing to it from other variable. In the research, what we ought to regress are as follows:

$$S = f_S(A, O) + U_S$$
$$A = f_A(F) + U_A$$
$$P = f_P(O) + U_P$$
$$D = f_D(S) + U_D$$
$$C = f_C(F, P, D) + U_C$$

U_S, U_A, U_P, U_D, U_C are exogenous variables. For each endogenous variable in our graph, there is an exogenous variable affecting its value. We use "U" to describe all the exogenous variables, like U_S to describe endogenous variable S's exogenous variable. Exogenous variables are not drawn in our causal graph. Every assignment $U = u$ to the exogenous variables stands for an individual for the reason that each $U = u$ confirms the value of endogenous variables uniquely. So the value of U depends on every single person, different U suggesting individual difference, which means each U corresponds to an individual in the data.

Before regressing, we need to preprocess the data. For example, given that lung cancer level is categorical features, which cannot be used directly, we transform them into numerical feature by label encoding, 1 representing low, 2 representing medium, 3 representing high.

We utilize multilayer perceptron to achieve the predicted value $f_S(A, O)$, $f_A(F)$, $f_P(O)$, $f_D(S)$, $f_C(F, P, D)$.

DOI: 10.54254/2753-8818/44/20240424

5.3. Abduction

The evidence $C = c_0$ is used for determining the value of U related with every individual through plugging in the values of the known variables. In other words, we account for past U by the evidence $C = c_0$. The value of exogenous variable is obtained by subtracting true value from predicted value.

$$U_S = S - f_S(A, O)$$
$$U_A = A - f_A(F)$$
$$U_P = P - f_P(O)$$
$$U_D = D - f_D(S)$$
$$U_C = C - f_C(F, P, D)$$

Take $U_S = S - f_S(A, O)$ as an example. For a certain individual, we surely know the true value S since it is in the data set. As for the predicted value $f_S(A, O)$, we have already used multilayer perceptron to solve it in the former step. So the value of exogenous variable can be calculated.

Take patient P1 as an example. By inputting all his data ($S = 3$, $A = 4$, $F = 3$, $O = 4$, $P = 2$, $D = 3$, $C = 1$) and predicted value into the above function expression, we can gain the value of his exogenous variable ($U_S = 0.068$, $U_A = -0.835$, $U_P = 0.089$, $U_D = -1.233$, $U_C = -0.002$).

5.4. Action

After the first two procedures, we get a model with total certainty, knowing all the exogenous variables' value and the functional relationship of the endogenous variables (figure 4). To evaluate how smoking can affect lung cancer while the other variables stay unaltered, we are supposed to revise the model through removing the structural equations for the variables smoking and replacing them with the functions $S = s_1$. After that, the modified model is achieved (figure 5).

5.5. Prediction

Finally, we apply the fundamental law of Counterfactual to compute the result of lung cancer level. Assume M_x as the modified version of M, with the value of X replaced by $X = x$. The definition of the counterfactual $Y_x(u)$ is $Y_x(u) = Y_{M_x}(u)$.

For an individual in the data set, if we replace his smoking level S by $S = s_1$, his lung cancer level is predicted to transfer to $C = c_1$. We change $S = s_0$ to $S = s_1$ while keeping the other variables immobile, and use the functional expressions we solve in the former step to calculate the counterfactual result of lung cancer level. Original and modified model are at figure 4 and figure 5.

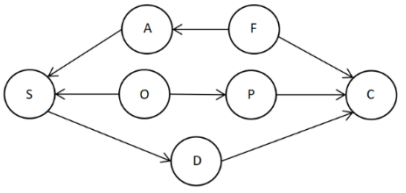

Figure 4. Original model

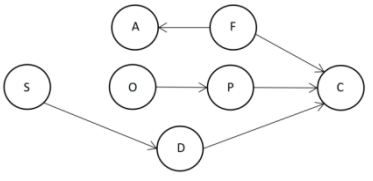

Figure 5. Modified model

As for patient P1, if his smoking doubled, which means his smoking changes from $s_0 = 3$ to $s_1 = 6$, it can be calculated that his lung cancer level is predicted to change from $C_0 = E(C_{s_0=3}) = 1$ to

DOI: 10.54254/2753-8818/44/20240424

$C_1 = E(C_{s_1=6} | C_{s=s_0} = C_0, s = s_0) = 1.876$, increasing dramatically.

After analyzing other patients in the same methodology, we finally draw the conclusion that there is a very strong positive correlation between smoking and lung cancer and that if smoking level increased while other factors stay unchanging, lung cancer level is predicted to grow as well.

6. Conclusion

The backdoor adjustment approach estimates causal impact through indirect routes. Adjusting for two backdoor paths, we conclude that there is strong positive relationship between smoking and lung cancer, yet the effect in unclear for $S = 5$and reversed for $S = 6$. We attribute the local dissonance to limited data set size, the causal assumptions in our proposed causal graph and unobserved confounders.

The front-door adjustment evaluates causation while allowing for the presence of unobserved confounders. We use an observable intermediate variable chronic lung disease. The results showcase a pattern similar to the one derived using backdoor adjustment and manifests itself clearly as a positive relationship between smoking and lung cancer. The effect is unclear for $S = 6$. We attribute it to limited data set size and the causal assumptions we made.

Our research changes from population data to individual behavior. Counterfactual, unlike backdoor adjustment and front-door adjustment, concentrates on individual level instead of group level, where we can better find out the causal relationship between smoking and lung cancer and help assess individual risks to make personal recommendations. The obtained results are experimentally unachievable with traditional methodology since one cannot have two smoking level simultaneously. We derived the functional relationships between variables using data set and calculated the exogenous variables for individuals, with which we conducted hypothetical experiments and estimated the consequence of changing smoking habits for individual patients. It further supports the conclusion that there is strong, positive causal relationship between smoking and lung cancer.

Acknowledgement

The order has no bearings on individual contributions. Authors have joint first authorship. Ziyang Qiu and Louis Cao are correspondence authors.

References

[1] Ozlü, T., & Bülbül, Y. (2005). Smoking and lung cancer. Tuberk Toraks, 53(2), 200-9.

[2] Loeb, L. A., Emster, V. L., Warner, K. E., Abbotts, J., & Laszlo, J. (1984). Smoking and lung cancer: an overview. Cancer research, 44(12_Part_1), 5940-5958.

[3] Marshall, H. (2012). Genetic and epigenetic factors in development of lung cancer. The Lancet Oncology, 13(12), 1188.

[4] Yoshino, I., & Maehara, Y. (2007). Impact of smoking status on the biological behavior of lung cancer. Surgery today, 37, 725-734.

[5] Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., & Boyle, P. (2008). Tobacco smoking and cancer: a meta‐analysis. International journal of cancer, 122(1), 155-164.

[6] Raghu, V. K., Zhao, W., Pu, J., Leader, J. K., Wang, R., Herman, J., ... & Wilson, D. O. (2019). Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models. Thorax, 74(7), 643-649.

[7] Lung Cancer Database. (2017). Available at https://data.world/cancerdatahp/lung-cancer-data [accessed July 31, 2023].

[8] Ahmad, A. S., & Mayya, A. M. (2020). A new tool to predict lung cancer based on risk factors. Heliyon, 6(2).

[9] Wagner, L. I., & Cella, D. (2004). Fatigue and cancer: causes, prevalence and treatment approaches. British journal of cancer, 91(5), 822-828.

DOI: 10.54254/2753-8818/44/20240424

[10] Obeid, S., Akel, M., Haddad, C., Fares, K., Sacre, H., Salameh, P., & Hallit, S. (2020). Factors associated with alcohol use disorder: the role of depression, anxiety, stress, alexithymia and work fatigue-a population study in Lebanon. BMC public health, 20(1), 1-11.
[11] DiFranza, J. R., & Guerrera, M. P. (1990). Alcoholism and smoking. Journal of studies on alcohol, 51(2), 130-135.
[12] McKee, S. A., Krishnan-Sarin, S., Shi, J., Mase, T., & O'Malley, S. S. (2006). Modeling the effect of alcohol on smoking lapse behavior. Psychopharmacology, 189, 201-210.
[13] Howard, J. (2004). Smoking is an occupational hazard. American journal of industrial medicine, 46(2), 161-169.
[14] Kim, Y. J. (2016). Impact of work environments and occupational hazards on smoking intensity in Korean workers. Workplace Health & Safety, 64(3), 103-113.
[15] Trichopoulos, D., Kalandidi, A., Sparros, L., & Macmahon, B. (1981). Lung cancer and passive smoking. International journal of cancer, 27(1), 1-4.
[16] Durham, A. L., & Adcock, I. M. (2015). The relationship between COPD and lung cancer. Lung cancer, 90(2), 121-127.
[17] Tanoue, L. T., Tanner, N. T., Gould, M. K., & Silvestri, G. A. (2015). Lung cancer screening. American journal of respiratory and critical care medicine, 191(1), 19-33.
[18] Denholm, R., Schüz, J., Straif, K., Stücker, I., Jöckel, K. H., Brenner, D. R., ... & C. Olsson, A. (2014). Is previous respiratory disease a risk factor for lung cancer?. American journal of respiratory and critical care medicine, 190(5), 549-559.
[19] Daniels, C. E., & Jett, J. R. (2005). Does interstitial lung disease predispose to lung cancer?. Current opinion in pulmonary medicine, 11(5), 431-437.
[20] Pearl, J., Glymour, M., & Jewell, N. P. (2016). Causal inference in statistics: A primer. John Wiley & Sons. pp. 61.