Final Project: Estimate long-term cost of poor life style choice

Yu Hui

June 2024

1 Introduction

1. Research Question: To what extent will having bad lifestyle increase the medical cost (and total avoidable cost) through having chronic illness?

2. Background: Chronic diseases such as diabetes, cardiovascular conditions, and respiratory disorders not only reduce the quality of life for millions but also impose substantial financial burdens on individuals and healthcare systems. These conditions often result from lifestyle choices, which may not have an immediate impact on an individual's health. In many cases, chronic illnesses are mild in their early stages, leading people to overlook the long-term consequences of their lifestyle decisions. This lack of immediate, visible impact can make the future medical costs associated with poor lifestyle choices less foreseeable, causing individuals to underestimate the financial burden of their habits.

   The challenge of assessing the long-term medical costs of lifestyle choices highlights a critical issue in health economics. Understanding how poor lifestyle choices today can lead to increased medical costs in the future through the development of chronic illnesses is essential for developing effective public health strategies and policies. **This study aims to address this gap by investigating whether having a bad lifestyle increases medical costs through the intermediary of chronic illness.** By examining this relationship, we seek to provide insights into the long-term economic impact of lifestyle choices and to emphasize the importance of early intervention and lifestyle modification to reduce future healthcare costs.

   In this research, we focus on analyzing the cumulative effects of lifestyle factors such as alcohol abuse, drug abuse, and obesity on the prevalence of chronic illnesses and the associated medical expenditures. By exploring the pathways through which these lifestyle choices affect health and financial outcomes, we hope to shed light on the hidden costs of poor health behaviors and the potential benefits of preventive measures. The findings from this study will contribute to the broader understanding of the economic implications of lifestyle choices, offering valuable evidence for policymakers, healthcare providers, and individuals aiming to improve public health and manage healthcare expenditures effectively. (However, as we have less information on the source of dataset, it should be note that our conclusion may loss external validity when being applied to other cases)

2 Methodology

1. Two-Stage Least Squares Regression: This study employs a two-stage least squares (2SLS) approach to explore the causal relationship between lifestyle factors and medical costs, mediated by chronic illnesses. In the first stage, we examine the impact of lifestyle choices on the probability of developing chronic illnesses, controlling for demographic variables such as age, gender, and race. In the second stage, we assess how the presence of chronic illnesses, influenced by lifestyle factors, translates into higher medical costs. The reason we choose chronic illnesses as the intermediary is that this variable captures the long-term, less observable outcomes brought about by lifestyle decisions. If we were to directly choose medical cost as the dependent variable without an intermediary, we might not properly separate the long-term, less foreseeable outcomes from the short-term disease outcomes brought about by lifestyle decisions, which is not our research interest.

2. Regression Model specification:

1. **First Stage Regression**:

$$X_i = \alpha_0 + \alpha_1 \text{lifestyle}_i + \sum_{j=1}^{k} \alpha_{2j} \text{controls}_j + \epsilon_i \tag{1}$$

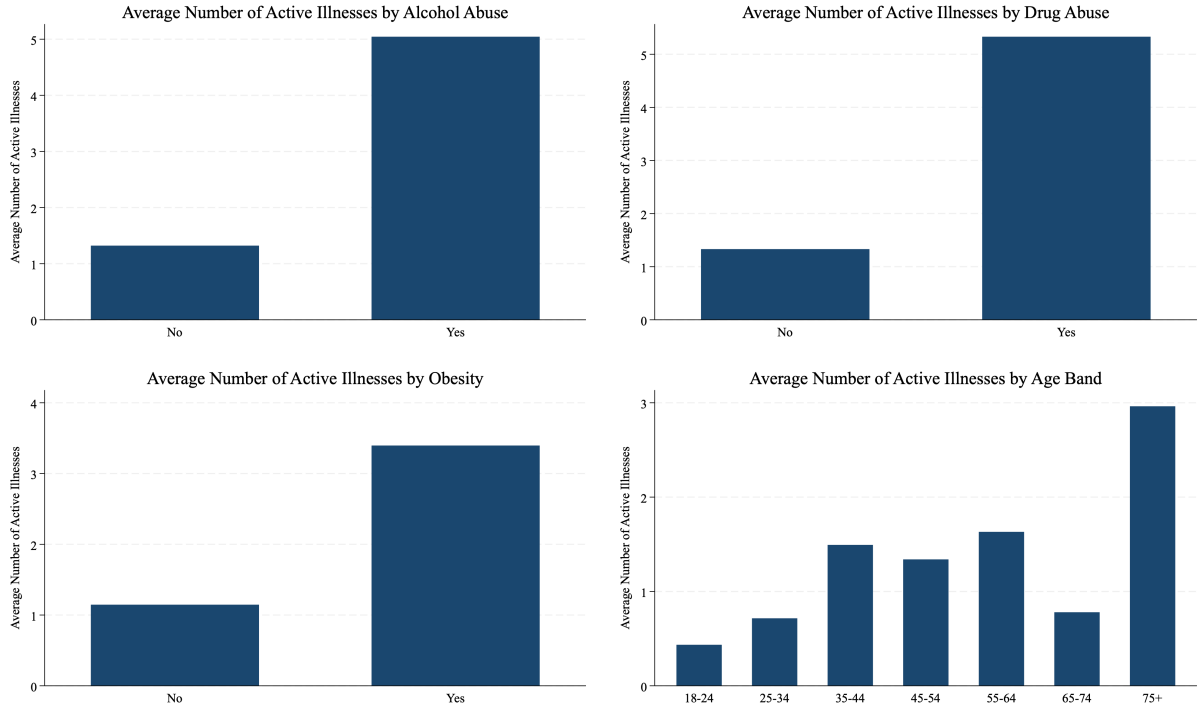
Here, X is the number of chronic illnesses, lifestyle variables are the variables of interest, including indicator of alcohol and drug abuse and obesity indicator. Controls are demographic and other relevant factors include age, gender and race, that may have impact on the possibility of having more chronic illness. For instance, as age going up, the potential of having illness also goes up, thus we need to control for this variable for lower variance and more accurate estimate.

2. **Second Stage Regression**:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \sum_{j=1}^{k} \beta_{2j} \text{controls}_j + \eta_i \tag{2}$$

Here, Y is the total medical cost, $\hat{X}$ are the predicted values of chronic illnesses from the first stage, and controls are the same set of variables used in the first stage.

3 Descriptive Analysis



In the first three charts, we can observe a significant disparity in the average number of active illnesses between individuals who abuse alcohol, abuse drugs, or have obesity and those who do not. Specifically, individuals with a history of alcohol abuse, drug abuse, or obesity exhibit a markedly higher average number of active illnesses compared to their counterparts who do not engage in these behaviors. This suggests a strong association between poor lifestyle choices and the prevalence of multiple chronic illness conditions.

In the last chart at the bottom right corner, we can see that generally increasing age is associated with a greater number of chronic illnesses, with the exception of the age group 65-74. This may be because some individuals in the previous age group with multiple chronic illnesses may have died, simultaneously decreasing both the numerator and the denominator, thus possibly lowering the overall percentage.

4 Regression Model Result

The second stage results of the model (Table 1 (2), (4), (6)) show that, controlling for covariates, poor lifestyle choices contribute to an increase of approximately $3,000 to 4{,}000$ in medical costs through the

channel of increased probability of chronic illness. This can be seen from the coefficient of the "Fitted" variable, which can be considered as the instrument created from the first stage regression.
Heterogenity Impact:: We can examine the groups of young people, middle-aged people, and elderly people separately to see if there is heterogeneity in the estimated cost of lifestyle choices after controlling for demographic covariates. We classify people aged 18-34 as young, people aged 35-64 as middle-aged, and people aged 65-75 as elderly. As shown in Tables 2 to 4, alcohol abuse has significantly higher long-term medical costs through the development of chronic illness for elderly people compared to young people. In contrast, drug abuse and obesity have significantly higher long-term medical costs through chronic illness for young people compared to the elderly.

5 Conclusion

This study focuses on measuring the long-term, less foreseeable medical costs brought about by poor lifestyle choices such as alcohol abuse, drug abuse, and obesity (excessive intake of sugar and carbohydrates). The results show that poor lifestyle choices significantly increase an individual's potential to develop chronic illnesses, which in turn raises medical costs by around $3,000. This study helps to provide a measure for the less foreseeable costs of poor lifestyle decisions, which is not only useful for policymakers in allocating health aid expenditure but can also be used in designing policy interventions as valuable information to alter people's beliefs about their lifestyle choices (as an early intervention method).

However, as we know little about the sample source of this dataset, we may need to be concerned about the external validity of our conclusion—we need information about the sample's time period, region, and other characteristics before the results can be applied to other cases. Moreover, it should be noted that depression may cause poor lifestyle choices and simultaneously increase the number of chronic illnesses for the patient, which implies that depression may confound our estimates. As we are still not sure if depression is included in the chronic illness category in this dataset, we did not control for the depression indicator in our regression, but we should be careful in dealing with this indicator if we obtain more detailed information on this dataset.

Table 1: Regression Results

	(1) Number of active chronic illnesses	(2) Total medical expenditures	(3) Number of active chronic illnesses	(4) Total medical expenditures	(5) Number of active chronic illnesses	(6) Total medical expenditures
Indicator for alcohol abuse	3.316***					
	(0.085)					
age between 18-24	-0.303***	711.462	-0.326***	1343.152**	-0.294***	676.916
	(0.060)	(578.246)	(0.060)	(583.501)	(0.057)	(574.829)
age between 25-34	-0.208***	1507.944***	-0.222***	1934.999***	-0.233***	1484.589***
	(0.053)	(508.980)	(0.053)	(513.428)	(0.051)	(507.382)
age between 35-44	-0.060	257.810	-0.074	386.811	-0.129***	250.755
	(0.051)	(488.872)	(0.051)	(492.922)	(0.049)	(489.087)
age between 45-54	0.261***	1669.473***	0.257***	1135.858**	0.173***	1698.656***
	(0.051)	(491.237)	(0.051)	(495.697)	(0.049)	(488.371)
age between 55-64	0.795***	345.583	0.803***	-1259.990**	0.666***	433.390
	(0.052)	(531.577)	(0.052)	(539.414)	(0.049)	(503.983)
age between 65-74	1.496***	-556.136	1.513***	-3567.442***	1.419***	-391.452
	(0.052)	(623.189)	(0.053)	(638.651)	(0.050)	(535.577)
age 75+	2.014***	386.579	2.020***	-3668.925***	2.014***	608.369
	(0.056)	(731.608)	(0.056)	(753.552)	(0.053)	(592.311)
race is Black	0.778***	-550.999*	0.772***	-2153.846***	0.565***	-463.342*
	(0.025)	(310.508)	(0.025)	(318.932)	(0.024)	(260.057)
female gender	-0.092***	1345.724***	-0.102***	1556.250***	-0.084***	1334.211***
	(0.017)	(159.993)	(0.017)	(161.509)	(0.016)	(158.559)
Fitted		2858.707***		4838.651***		2750.426***
		(244.582)		(257.814)		(125.505)
Indicator for drug abuse			3.786***			
			(0.102)			
Indicator for obesity					2.114***	
					(0.027)	
Observations	48784	48784	48784	48784	48784	48784

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2: 2SLS Regression Results by Age Group:Independent variable: alcohol abuse indicator

	(1) First Stage (Young)	(2) 2SLS (Young)	(3) First Stage (Middle-aged)	(4) 2SLS (Middle-aged)	(5) First Stage (Elderly)	(6) 2SLS (Elderly)
alcohol abuse	2.985***		3.581***		3.561***	
	(0.183)		(0.251)		(0.109)	
race is Black	0.916***	-553.959	0.360***	-1720.620**	0.821***	-165.397
	(0.070)	(736.918)	(0.046)	(612.404)	(0.031)	(390.322)
female gender	-0.302***	915.171*	0.128***	1606.329***	-0.100***	1458.079***
	(0.045)	(403.586)	(0.030)	(363.445)	(0.021)	(200.033)
Fitted		2523.841***		2843.049***		2852.421***
		(499.330)		(813.460)		(294.362)
Constant	2.556***	3213.166*	0.427***	3504.729***	1.100***	3059.720***
	(0.036)	(1337.824)	(0.025)	(457.092)	(0.017)	(370.181)
Observations	10357	10357	6904	6904	30604	30604

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: 2SLS Regression Results by Age Group: Independent variable: drug abuse indicator

	(1) First Stage (Young)	(2) 2SLS (Young)	(3) First Stage (Middle-aged)	(4) 2SLS (Middle-aged)	(5) First Stage (Elderly)	(6) 2SLS (Elderly)
Indicator for drug abuse	3.626***		3.875***		3.550***	
	(0.199)		(0.127)		(0.259)	
race is Black	0.357***	-2821.683***	0.823***	-2391.316***	0.887***	-356.701
	(0.045)	(585.619)	(0.031)	(406.805)	(0.070)	(795.904)
female gender	0.113***	1244.323***	-0.107***	1750.875***	-0.315***	845.785*
	(0.030)	(359.897)	(0.021)	(203.214)	(0.045)	(417.640)
Fitted		5758.015***		5435.390***		2312.644***
		(646.080)		(317.796)		(592.954)
Constant	0.430***	2225.391***	1.110***	132.718	2.586***	3765.437*
	(0.024)	(403.335)	(0.017)	(395.200)	(0.036)	(1577.772)
Observations	6904	6904	30604	30604	10357	10357

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: 2SLS Regression Results by Age Group: Independent variable: obesity indicator

	(1) First Stage (Young)	(2) 2SLS (Young)	(3) First Stage (Middle-aged)	(4) 2SLS (Middle-aged)	(5) First Stage (Elderly)	(6) 2SLS (Elderly)
Indicator for obesity	1.882***		2.151***		2.168***	
	(0.060)		(0.031)		(0.075)	
race is Black	0.266***	-1849.623***	0.592***	-163.794	0.650***	-361.699
	(0.043)	(548.111)	(0.030)	(322.422)	(0.069)	(632.542)
female gender	0.122***	1563.915***	-0.083***	1457.868***	-0.296***	847.543*
	(0.029)	(351.450)	(0.019)	(197.932)	(0.044)	(381.497)
Fitted		3184.573***		2850.561***		2317.996***
		(386.940)		(146.556)		(290.067)
Constant	0.347***	3354.839***	0.924***	3061.827***	2.426***	3751.442***
	(0.024)	(331.450)	(0.016)	(230.972)	(0.035)	(812.711)
Observations	6904	6904	30604	30604	10357	10357

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$