Evaluation on Public Policy Intervention Using Interrupted with Time Series Regression

Kahang Ngau
School of Engineering and Applied Science
George Washington University
800 22nd St NW, Washington, DC 20052
ngaukahang@gmail.com

***Abstract*—Secondhand smoke (SHS) can cause numerous serious health problems among adults and children. According to the statistics from CDC, there are more than 7,300 lung cancer deaths among nonsmokers each year caused by SHS [2]. It is important to know the effectiveness of public policy intervention which was designed to help alleviate people exposed by SHS. One of the powerful analytical approaches will be interrupted with time series regression analysis. The interrupted time series is a useful method and study to evaluate population-level health interventions, especially in the field of public health legislation. In this paper, we will introduce the use of interrupt with time series regression method on a real-world example of public intervention. We first talk about the motivation for conducting this experimental project. Then we highlight the causal inference of the topic in the project, including the proposed cause and the proposed effect. It is also critical to introduce the pilot data source that is used in this study and its accessibility as well. We then discuss the sampling strategies concerning the external validity of our pilot data. The measurement from the construct validity of the study is clearly defined before the quasi-experimental design is implemented. Finally, we describe the main method which is the interrupted with time series analysis, and interpret the meaning from the results.**

Keywords—Secondhand smoke, population-level, intervention, interrupted with time series, regression

I. INTRODUCTION

Racial/ethnic and socioeconomic disparities play a critical role in stark health disparities observed in the U.S. The prior study suggests secondhand smoke (SHS) is one of the greatest factors contributing to diseases and harming human bodies. CDC reports that there is a 20-30% increase for non-smokers to develop lung cancer if they are exposed to SHS [2]. Therefore, policymakers need to know the importance to establish certain public policies to help decrease people's level on SHS exposure. In the meantime, discovering an innovative and efficient method to evaluate the effectiveness of the public policy intervention will be critical as well. One of the gold standard experimental designs to efficiently evaluate the effeteness of public intervention is by conducting randomized assignments. However, the use of the randomized assignment approach is not possible in this project because of the observation data we have. It is also particularly difficult for projects that involve health policies and legislations that focus on population-level studies, according to the paper from Victora [1]. Moreover, the practice in this paper will be useful because it can be applied to other similar problems or studies that have the need or goal in evaluating the effeteness of public intervention. In recent decades, interrupted with time series had been used widely to solve different problems, especially in the field of public health interventions. It is a appreciated approach because it is considered as a quasi-experimental design with a target population level over a clear period time that is well defined. There are still some of the limitations of the use of interrupted time series listed from prior research papers. It is critical to understand well the limitations and possible confounds when using this approach. The goal of this project is to develop an interrupted time series analysis to evaluate a public policy intervention that will affect people's SHS exposure level. And we also want to study how people with different characteristics (poor vs rich, and smokers vs non-smokers) are affected by SHS under the policy intervention. This leads to the questions that we

try to answer in this project: How does the implementation of raising the federal tax rate for cigarettes affect the level of secondhand smoke exposure among smokers and nonsmokers? With the policy, will people with different income levels have different effects in regards to their SHS exposure. Here, we develop a hypothesis about the intervention's impact on the outcome and apply the regression model to analyze the results. The programming environment for this study is conducted in R with the statistical packages.

II. Literature review

There is one research paper from Richter about analyzing smoker exposure using the NHANES data that helps me better explore the dataset in this project [3]. The paper examines the smoke and secondhand smoke exposure and the level of 12 urinary metals on participants from the NHANES dataset. The authors believe urine measurements are a useful noninvasive approach in biomonitoring research of exposures to metals and other environmental pollutants. The year range of the dataset is from 1999 to 2004. Many influential predictors were included in the examination, including age, race, gender, poverty level, cadmium, and lead concentration. This paper first uses statistical approaches to examine the significant difference level of different urinary metals across smokers and nonsmokers. It finds that some metals like cadmium and lead, and antimony and barium are increased in smokers, while others like mercury, beryllium platinum, and thallium are lower or unchanged (tungsten). An inverse relationship was discovered between smoking and levels of important nutrients among the participants. And it also finds that children, who are classified as age below 12, are a group of the population particularly vulnerable to secondhand smoke exposure. The statistical results show that the toxic effects of lead for children who have low levels of exposure have a higher level of urine lead than children without SHS exposure. The findings from this paper provide a better understanding of interpreting the pilot data source, NHANES.

One important measurement of serum cotinine concentration is conducted when comparing the differences between smokers and nonsmokers. This measurement is very important to my capstone project since it is the key variable that I will investigate. Knowing how to handle and interpret this variable will be helpful to make predictions about it in my project. Also, the ways to handle null values and steps to classify nonsmokers will be helpful to my project since I will also conduct statistical analysis from the NHANES dataset.

III. Methodology

In this section, we introduce in detail the methodologies that are applied in this study. There are a total of three parts are included: the sampling strategies in the preparing steps of external validity, the measurement in the construct validity, the experimental design (interrupted with time series), and the analysis for the conclusion validity. These are all some of the important steps that are needed to be completed before a meaningful conclusion can be drawn. After these steps are all explained, the paper will move to the next section presenting some of the preliminary results from this study.

A. Sampling Strategy

The pilot data source of this project is from National Health and Nutrition Examination Survey (NHANES) NHANES is a national survey that combines both interviews and physical examinations. The interview questions include demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. The reason we choose to use NHANES as the pilot data source of the project is that it is a national program that was designed to access and study on the health and nutrition status among children and adults in the United States [5]. It is also part of the National Center for Health Statistics Program which aims to produce vital and health statistic for the nation. The sampling strategy that was used in NHANES is a complex and probability sampling that involves multistage in the process to select survey participants representative of the US population. The NHANES sampling procedure consists of four stages. The first stage includes dividing all the counties in the United States into 15 groups based on their characteristics. Most of them are single counties, with only a few of them being groups of contiguous counties [6]. One is selected from each group and together they form the 15 counties in the NHANES survey for the year. The sampled counties are divided up into segments or neighborhoods. A random sampling within the dwelling units or households with each selected neighborhood in stage 3. The final stages will consist of the random selection of individuals that are chosen to participate in the NHANES survey. According to the NHANES website, about 2 sample persons on average is selected from each of that eligible household [6]. These detailed procedures have applied to all previous cycles of the NHANES survey, but it is also subjected to change for some of the details over time.

The unit of analysis in this study is the individual survey participant that has conducted the blood examination. The accessible study population is all the people who are exposed to

secondhand smoke participate in NHANES and examine blood tests. The NHANES data source is public and can be downloaded from the website. The final data set after cleaning will consist of 44,667 observations, across the survey year from the year 1999 to the year 2016. There is a possible threat to the external validity that we still need to consider since there is no control for knowing the geographical locations.

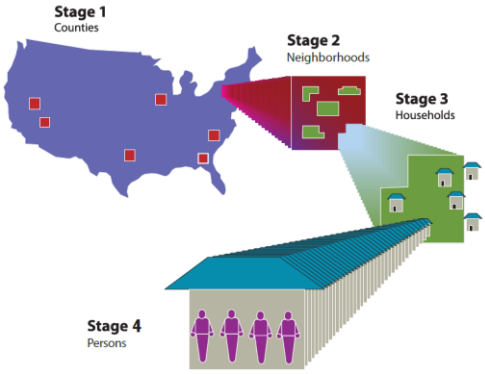

Figure 1 Four Stages of NHANES Sampling Procedure [6]

B. *Measurement*

Since the goal of this study is to evaluate the impact of a public intervention on people's exposure level of SHS among groups of people with different characteristics, there are two main groups of constructs that we need to measure. The first construct is the measurement of SHS. Through conducting prior researches, secondhand smoke is mainly measured by serum cotinine (ng/mL), which is recorded in NHANES for those who examine blood tests. The next important construct is the measurement of the Federal Poverty Income Ratio (PIR). The PIR is measured by family total income divided by the family poverty threshold. If the total family income is less than the poverty family threshold which is less than 1, that family and everyone in it is considered to be in poverty. Otherwise, if the PIR is equal or greater than 1, that family is not considered to be in poverty. These two constructs of measurement are both relatively reliable since the serum cotinine level is measured through blood tests. The PIR is also an accurate measurement since it is calculated by the family income threshold. Nonsmokers in this study will be defined as people who have not smoked 100 cigarettes in their lives. This is presented in one of the NHANES questions which survey participants respond with yes or no. Therefore, participants who answer yes are considered smokers, and those who answer no are considered nonsmokers.

C. *Experiemntal Design*

As discussed early in this paper, it is not feasible to conduct randomized assignments in experimental design in the project. Rather, a quasi-experimental design will be performed in evaluating the effectiveness of the public intervention. According to Bernal's paper, an interrupted time series study has a continuous sequent of data points over a defined population and it depends on both the nature of the intervention and the desired outcome, including the available data types [7]. The paper also introduces that a time series will be better defined when a particular desired outcome is used to develop an underlying trend. And that trend is interrupted by intervention at a particular known point of time. Since one of the project goals is to divide the data into groups of people with different characterizes and then to compare how each of them will be affected differently with the intervention. There are four different groups and which they are: poor smokers, poor nonsmokers, rich smokers, and rich nonsmokers.

The Public Intervention

An interrupted time series analysis needs a well-defined point of time to distinguish between the pre-intervention period and the post-intervention period. On February 4, 2009, the Children's Health Insurance Program Reauthorization Act of 2009 was signed into law, which raised the federal tax rate for cigarettes on April 1, 2009, from $0.39 per pack to $1.01 per pack [8]. The increase was to help cover the cost of increased coverage under the State Children's Health Insurance Program. The intervention in this study was well clearly defined with a raise in federal tax on cigarettes on Apr 1, 2009.

D. *Data Preparation and Analysis Plan*

After determining the intervention for the study of interrupted with time series, the next step is to make a hypothesis about the level of impact or how the intervention would affect the outcome which is the serum cotinine. The level of impact of the change can be varied depending on the situation after the intervention. The change can be the level, the slope, or both. There can also be a lag before any of the changes take place. It is important to make the hypothesis about the level of change or the trend at the early stage of the study. And the decision will be made based on the existing knowledge and from prior researches. It is reasonable to assume that the level change in the smokers' group will be smaller than in the nonsmoker groups. Before conducting the regression analysis, it is important to structure the data set and present some descriptive analysis from the original data.

Table 1. Example sub-dataset from the poor smokers

Year	Serum Cotinine (Y)	Time elapsed (T)	Intervention (X)
1999	97.20099	1	0
2000	135.76759	2	0
2001	112.72398	3	0
2002	147.59942	4	0
2003	98.62894	5	0
2004	170.55575	6	0
2005	128.65313	7	0
2006	157.72888	8	0
2007	138.16273	9	0
2008	202.88882	10	0
2009	142.27833	11	1
2010	162.77148	12	1
2011	131.86633	13	1
2012	154.58668	14	1
2013	167.53139	15	1
2014	188.53178	16	1
2015	113.32427	17	1
2016	162.61228	18	1

Intervention: 0, intervention not in place; 1, intervention in place.

After the data is cleaned and structured with desired variables, the regression of the interrupted with time series analysis is developed. The following are the variables we have:

- Y: the outcome which is the serum cotinine concentration
- T: the time elapsed from the beginning of this study which is starting from 1999 to 2016
- X: the dummy variable represents the pre-intervention period with 0 and the post-intervention period with 1.

The regression model we have for the interrupted with time series analysis, as in:

$$\gamma = \beta 0 + \beta 1 * T + \beta 2 * X + \varepsilon \quad (1)$$

The b0 in the regression model indicates the baseline when the time T equals to 0. The b1 is representing the change in the predicted outcome which is the Y as it is connected with the time unit changes. Furthermore, b2 is indicating the level of change according to the intervention. In this study, variables Y, X, and T are presented in table 1 and they are all important indicators in an interrupted time series regression model. And because the outcome of the regression model is a continuous figure (which is the serum cotinine concentration), the family type of the General Linear Model (GLM) will be used as Gaussian. It is important to understand the data type first before constructing the model. For instance, if the outcome of the model is a count, then it is suitable to use the family type as Poisson in a GLM regression. Or if the outcome is considered to be a binary type of value, the family type of the GLM will be binomial. Figure 2 below presents the distribution of serum cotinine across the four different groups. There is a clear drop right at the intervention for four plots. Then we can see from table 2 there is a decreasing trend for both rich and poor nonsmokers after the intervention.

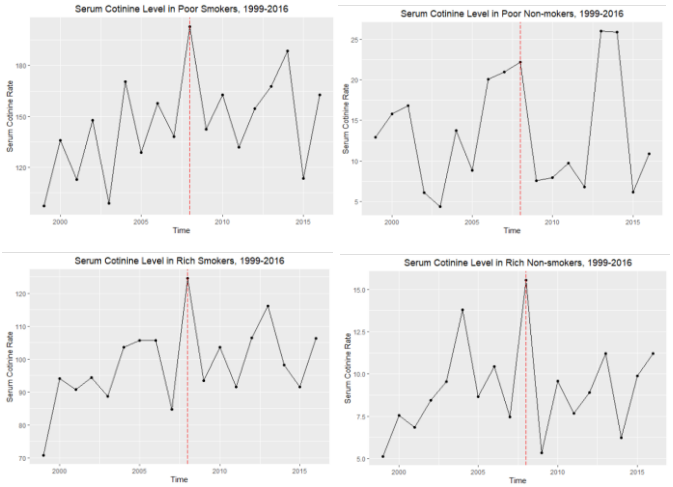

Figure 2 Line plots of the four different groups with year on the x-axis and serum cotinine on the y-axis. The red vertical line indicates the year when the intervention occurs. Plots at the top left: Poor smokers; top right: Poor nonsmokers; bottom left: Rich smokers; bottom right: Rich nonsmokers

Table 2 Mean of serum cotinine for the four groups before and after the intervention

	Poor Smokers	***Poor Nonsmokers***	***Rich Smokers***	***Rich Nonsmokers***
Pre-intervention	139.0	14.169	96.30	9.342
Post-intervention	152.9	12.604	100.93	8.754

IV. PRELIMINARY RESULTS

In our case, we divide our dataset into four different groups of people who have their characteristics. A GLM regression is implemented in each of the groups and the models are complemented in R code. The first model, in figure 3, is the poor smokers and the result suggests that there is little to no evidence of change in serum cotinine associated with the cigarette tax raised intervention. The P-value of the intervention X is 0.2016 (> 0.05) which we fail to reject the null hypothesis that there is no impact of intervention at the significance level of 0.05. However, there is an increase in serum cotinine as time elapsed increases, with a P-value < 0.05 indicating the factor Time has a positive impact on serum cotinine level at the significance level of 0.05.

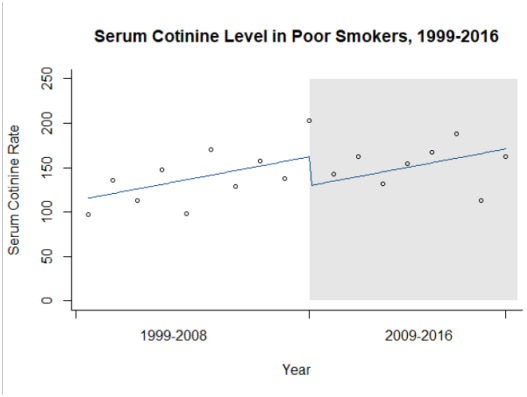

Figure 3 Interrupted time series regression model for poor smokers. Blue line: predicted trend based on the regression model. Pre-intervention period is in the white background. Post intervention period is in the grey background.

By visualizing the regression plot from figure 4, there is a bigger level change detected after the intervention for poor nonsmokers, compared with the level change in poor smokers. However, when interpreting the model result for the group of poor smokers, both the Time and Intervention factor have a P-value greater than 0.05 and it indicates we fail to reject the null hypothesis at a significance level of 0.05.

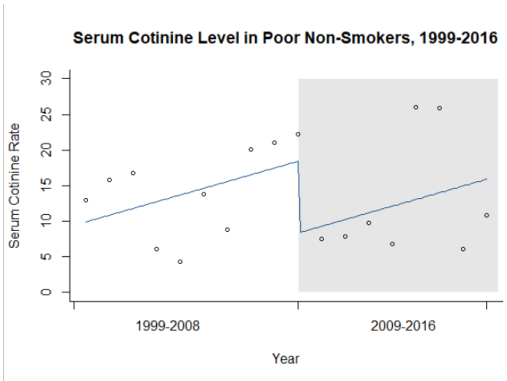

Figure 4 Interrupted time series regression model for poor nonsmokers.

After conducting the results from poor smokers and poor nonsmokers, the interrupted time series models are developed for rich smokers and rich nonsmokers in figure 5. As we see from both plots, there is also a clear level change difference between rich smokers and nonsmokers after the intervention. The model results for the rich smokers present a P-value of the intervention X is 0.1084 (> 0.05) which we fail to reject the null hypothesis that there is no impact of intervention at the significance level of 0.05. There is an increase in serum cotinine as time elapsed increases. With the P-value < 0.05 of the factor Time, indicating a positive impact on serum cotinine level at the significance level of 0.05. For the model results from rich nonsmokers, the P-value of the intervention X is smaller than 0 (P < 0.001) which we reject the null hypothesis in favor of the alternative hypothesis that there is an impact of the intervention. There is an increase in serum cotinine as time elapsed increases. The P-value is less than 0.05 for factor Time, indicating there is an impact on serum cotinine level at the significance level of 0.05.

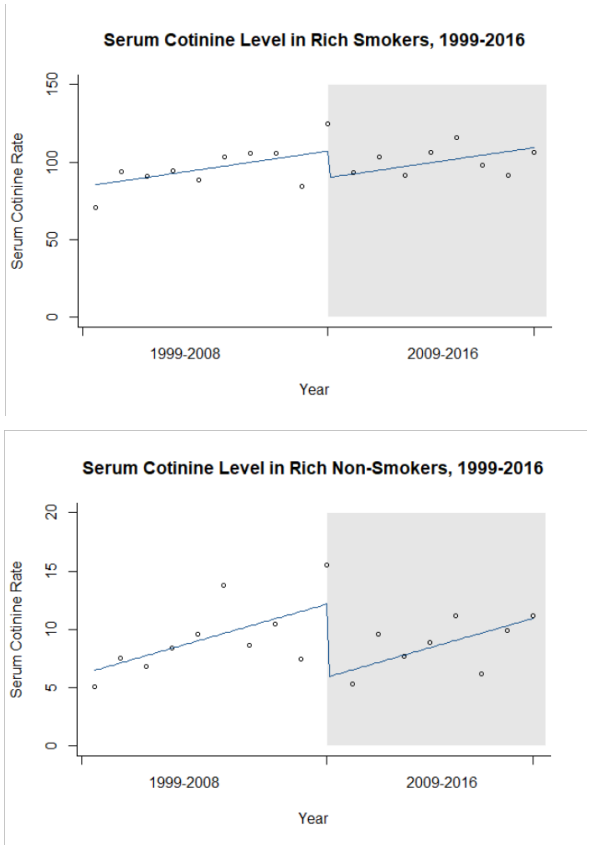

Figure 5 Interrupted time series regression model for rich smokers and rich nonsmokers.

V. DISCUSSION

According to Bernal's paper about diagnostic analysis on the interrupted with time series model, the over-dispersion issue does not apply to the analysis using continuous data for outcome in the Gaussian model [7]. Therefore, this issue of a great variance leading to inaccurate estimation of standard error does not apply to this study. In addition, some more diagnostic tests are needed to preform for the rich nonsmokers' group. The first assumption in the standard regression model is to have independent observations. Although this assumption is usually violated when dealing with interrupted time series models, the Durbin-Watson statistic in the rich nonsmokers' model is equal to 1.88456 which is close to 2 indicating very little to no presence of auto-correlation. Again, the individual P-value of explanatory variables is small enough to indicate we can reject the null hypothesis in favor of the alternative hypothesis that at least one coefficient is not equal to zero. From figure 6, the histogram of residual shows a normal distribution on the residuals of the rich nonsmokers' regression model. For the Residuals Versus Fitted Value scatters plot, there is no apparent heteroscedasticity that can be found on the graph. It means there is evidence to support the residuals have a constant variance. The Residuals Versus Order graph shows that the points are relatively chaotic. It indicates that the residuals are not infected individually distributed.

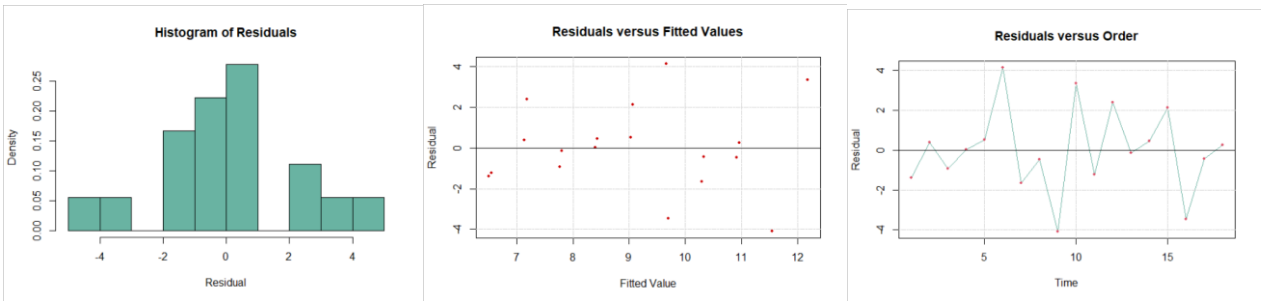

Figure 6 Plots for residual analysis of the rich nonsmokers' regression model

Base on the preliminary results, there is clear evidence indicating that rich nonsmokers are the only group that gets away from SHS exposure with the policy intervention. And there is evidence showing the public policy intervention does not have an impact on smokers, regardless of the rich or poor groups. It is not surprising to notice that the federal tax raised on cigarettes does not affect smokers' serum cotinine level since smokers tend not to quit smoking easily. For smokers in the rich and poor groups, there is a positive impact of Time in serum cotinine concentration. It indicates that smokers' serum cotinine level has increased over time. Since rich nonsmokers are the only group in this study that has a significant impact on both intervention and time, it answers the project's question of which the public policy intervention had significantly decreased rich nonsmokers' serum cotinine level. It also further illustrates that poverty may increase the chance to be exposed by SHS. However, there are some limitations and important threats to the validity of this interrupted time series study. One of the biggest limitations is the lack of statistical power in the regression analysis. Because of the nature of the public survey, there are many other useful variables (e.g., information about month and location, etc..) that are restricted and not accessible. Therefore, the data points that are used in this study are limited since we only use the serum cotinine level at the year level. Also, there is potential history confound in the project that can affect the outcomes. History confound can be any of the events that happened during the period time of the intervention, such as the implementation of smoke-free policy or financial crisis in 2008, etc. Furthermore, experimental mortality can also be concerned when there are survey participants become inactive or decide not to respond to the survey questions. This might happen because they are sensitive to the questions and not willing to expose their health conditions or behaviors. And finally, there can be instrumentation confound when the methods or instruments to examine blood test changes over time.

In this article, we have demonstrated some of the key steps in building up an interrupted with time series model. The interrupted with time series model is extremely useful when applying to a longitudinal study of the impact with an intervention. Moreover, it is also a great quasi-experimental approach when randomization of assignment is not possible. One of the great challenges is to make correct conclusions about the intervention effeteness while overcoming some of the issues relate to time-varying confounds. Future work of this project will include continuing exploring with the NHANES dataset to look for more valuable variables, adding one or more control

groups which will not be affected by the intervention to the regression model, and running the multiple baseline models with different locations. By doing this, there will be a need to request restricted data from NHANES to gather those variables. It will be a great way to access more data if we can make linkage with other data sources (e.g., American Housing Survey (AHS) or Department of Housing and Urban Development (HUD)) to investigate how the assisted housing program can help alleviate people's SHS exposure. Some interesting factors can include housing type, home type, and building year built.

ACKNOWLEDGMENT

We would like to thank Prof. David Broniatowski, Dr MyDzung Chu, and Dr Ami Zota for providing great support and the data used in this project.

REFERENCES

[1] Victora CG, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J Public Health 2004;94:400–05.

[2] "Health Effects of Secondhand Smoke" Fed. 27, 2020. Accessed on: Dec. 13,2021.[Online].Available:https://www.cdc.gov/tobacco/data_statistics/fact_sheets/secondhand_smoke/health_effects/index.htm

[3] Richter PA, Bishop EE, Wang J, Swahn MH. Tobacco Smoke Exposure and Levels of Urinary Metals in the U.S. Youth and Adult Population: The National Health and Nutrition Examination Survey (NHANES) 1999–2004. International Journal of Environmental Research and Public Health. 2009; 6(7):1930-1946.

[4] Gaurang P. Nazar, MSc, John Tayu Lee, PhD, Monika Arora, PhD, Christopher Millett, PhD, Socioeconomic Inequalities in Secondhand Smoke Exposure at Home and at Work in 15 Low- and Middle-Income Countries, Nicotine & Tobacco Research, Volume 18, Issue 5, May 2016, Pages 1230–1239

[5] "About the National Health and Nutrition Examination Survey" Sep. 15, 2017. Accessed on: Dec. 13, 2021. [Online]. Available: https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

[6] "Module 2: Sample Design" Sep. 15, 2017. Accessed on: Dec. 13, 2021. [Online].Available:https://wwwn.cdc.gov/nchs/nhanes/tutorials/module2.aspx

[7] Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2017 Feb 1;46(1):348-355. doi: 10.1093/ije/dyw098. Erratum in: Int J Epidemiol. 2020 Aug 1;49(4):1414. PMID: 27283160; PMCID: PMC5407170.

[8] Bush T, Zbikowski S, Mahoney L, Deprey M, Mowery PD, Magnusson B. The 2009 US federal cigarette tax increase and quitline utilization in 16 states. J Environ Public Health. 2012;2012:314740. doi: 10.1155/2012/314740. Epub 2012 May 8. PMID: 22649463; PMCID: PMC3356941.