The Effect of Participating in Physical Activity on Health for Americans: Evidence from 1990-2010

MARWAN BADRAN
8946590

Abstract

Using cross-sectional drawn data from the Behavioral Risk Factor Surveillance System (BRFSS) for the years 1990-2010 in the United States, I examine the causal impact of participating in physical activity on body mass index. I employ an Instrumental Variable framework as a solution to the endogeneity of physical activity, in which I employ the respondents' answers to the survey question: "Are you using physical activity or exercise to lose weight or maintain your current weight?", as the instrument. Consistent with the findings of previous studies, the OLS results suggest that participation in any level of physical activity is negatively associated with the body mass indices of individuals. Holding all other factors constant, participating in physical activity is associated with a 5.0% decrease an individuals' body mass index, while engaging in irregular, regular, and active levels of physical activity is associated with decrease of 3.6%, 5.5%, and 8.7% respectively. Therefore, their statistical effects increase in magnitude as the physical activity levels increase. The attributes of being a female or a smoker are associated with decreases in BMI, while age is positively associated with BMI. Unfortunately, I was unable to obtain causal effects due to employing a weak instrument in the instrumental variables regression. Finally, I conclude that further research is needed to find additional strong and relevant instrument and to acquire a more consistent dataset with well-defined variables in order to obtain more reliable and complete results.

1. Introduction

Physical activity is widely considered to be one of the enhancers of an individual's health, while being physically inactive is associated with significantly deteriorated health status. According to the World Health Organization, the lack of physical activity has been recognized as the fourth-ranking factor of global mortality, with an estimated 3.2 million premature deaths around the world. Moreover, the lack of physical activity is estimated to be the main factor that causes 21-25% of breast and colon cancers, 27% of diabetes and approximately 30% of ischaemic heart disease burden (World Health Organization, 2017). Furthermore, individuals who are physically active tend to live longer, healthier lives. This is due to the fact that being physically active lowers the risk for various diseases such as stroke, mental depression, and heart-related problems (Centers for Disease Control and Prevention, 2014). Physical activity can also help with mental health status, attaining weight loss, as well as strengthening bones and muscles (Centers for Disease Control and Prevention, 2014).

The impact of physical activity on health outcomes is a widely discussed topic among health economists; therefore, various studies have been conducted in order to estimate the magnitude of this effect. This topic is crucial for governments and policy makers in order to shape policies that encourage individuals to participate in physical activity, because a healthier population will always be more productive and have a lower overall mortality rate. It is also essential for individuals to understand the benefits that derive from participating in physical activity, whether in terms of short-term benefits, such as physical appearance, or in terms of long-term physical and mental health benefits.

Most research has shown that physical activity is negatively associated with numerous health risks, such as obesity, or with developing various diseases, such as chronic and heart diseases. However, a common limitation to most of the available literature on this topic is that

they do not account for the potential endogeneity of participating in physical activity. The act of engaging in physical activity is correlated with unobserved factors in the error term, such as the opportunity cost of time, leisure time preference, and whether individuals actually enjoy participating in physical activity. The endogeneity issue could be worsened by the potential reverse causality between participating in physical activity and the body mass index, as well as by the high measurement error in the regression model. Therefore, any obtained econometric estimates will be biased and inconsistent if these issues are not accounted for.

Different methods could be conducted to solve for the endogeneity of physical activity such as randomizing individuals into different levels of physical activity. However, such an experiment would be extremely costly and might also be viewed as unethical, especially if it includes a large sample size. Another, feasible, method involves using an instrumental variable (IV) approach. Instrumental variables are variables that are correlated with the endogenous explanatory variable but do not affect the dependent variable except through the channel of that explanatory variable. In other words, these instruments should be uncorrelated with any omitted variable included in the regression and, therefore, not correlated with my dependent variable, the body mass index, except through the channel of physical activity.

The substantial benefits of physical activity inspire researchers to estimate the causal impact of participating in physical activity on health. In this paper, I aim to estimate the causal impact of participating in physical activity the body mass index of Americans.[1] In order to obtain these causal inferences, I use the respondents' answers to a survey question that asks if individuals use physical activity or exercise in order to lose or maintain their weight as my instrumental variable in a two-stage least squares estimation. I argue that the individuals' motive to participate in physical activity is definitely correlated with physical activity and does

[1] Using data from the Behavioral Risk Factor Surveillance System (BRFSS) for the years 1990-2010.

not impact their health except through the act of engaging in physical activity. Therefore, it should create a degree of exogenous variation in physical activity, which will allow me to obtain causal impacts.

The findings of this study are aligned with the results of previous studies that found negative effects for physical activity and a positive effect for age on health outcomes (Parsons et al., 2005; Tao Zhang 2017). The results of the Ordinary Least Squares (OLS) regressions show that participating in physical activity is negatively associated with an individuals' body mass index. Relative to being physically inactive, engaging in irregular, regular, or active levels of physical activity appear to have negative relationships with BMI (Humphreys et al., 2014; Parsons et al., 2005). They also show that higher levels of education and income are associated with lower BMI values, holding all other factors constant (Humphreys et al., 2014; Tao Zhang 2017). Moreover, the attributes of being a smoker or a female appear to be negatively associated with BMI, all other factors held constant.

The instrumental variable regressions and, specifically, finding a good instrument, were problematic. Despite being similar to instrumental variables used in previous studies, the instrument that I used proved to be unsuitable and I was unable to obtain consistent and unbiased causal effects. I discuss the difficulties faced and the proposed solutions in the following sections of the paper.

Finally, I estimate a different regression model in order to measure the strength and sensitivity of my results. Using a binary health outcome, diabetes, I employ a probability linear model to examine the effect of participating in physical activity on the probability of diabetes. The results supported the findings of this paper and were aligned with the findings of previous studies.

The remaining sections of the paper are set out as follows, section 2 reviews the related literature. Section 3 describes the data set that was used in this paper. In section 4, I state the identification strategy used and illustrate the dependent, instrumental, and explanatory variables used. In section 5, I present the obtained results. In section 6, I estimate a different regression model in order to test for the strength and sensitivity of my results, and section 7 concludes the paper.

2. Literature Review

Research has shown that physical activity is associated with lower risks of being overweight and obese, developing chronic diseases, and the probability of premature mortality (Sarma et al., 2015; Humphreys et al., 2014). A review of the available literature shows that there exist different methods of estimating the impact of physical activity on the health of individuals. The difference is mainly attributed to using different definitions of physical activity and/or different variables that reflect health outcomes. In this section, I will discuss previous studies that examined this relationship.

The relationship between physical activity and health outcomes has been first introduced in a theoretical article by Michael Grossman's (1972) human capital model of the demand for health. The model was the first of its kind to allow for the interaction of non-medical inputs, such as physical activity, and health outcomes. Grossman argued that, contrary to previous studies, health differs from other forms of human capital. He stated that "a person's stock of knowledge affects his market and nonmarket productivity, while his stock of health determines the total amount of time he can spend producing money earnings and commodities". Put differently, consumers demand health for two reasons: consumption and investment. Consumption of health directly enters an individuals' utility function, thereby, increasing his/her utility. On the other hand, investing in health indirectly affects an individual's utility function by increasing the time available for individuals to participate in market and nonmarket

activities. Therefore, a healthy individual will have more time to increase his/her utility by consuming a different product.

Grossman stated that an individual initially inherits a stock of health, which depreciates over time as the individual grows older. An individual's stock of health can be increased by investment. An individual determines his optimal level of health capital investment by equating the marginal benefits of investing in his/her health to its marginal cost. The marginal costs are divided into two categories; the depreciation rate of health over time, and the opportunity cost of capital invested. Grossman argued that the quantity of health capital demanded decreases as age increases due to an increase in the cost of health investments as a result of higher depreciation of health as age increases. He also stated that the income and education levels are positively correlated with the quantity of health capital demanded. Higher education levels imply more efficient health investments, and higher income levels imply a decrease in relative costs of health investments. Therefore, individuals are healthier as their income and education level increase as a result of an increase in their optimal level of health.

Primarily, there are empirical papers that studied the effect of physical activity on general health risks. Grossman's (1972) human capital model of the demand for health served as a foundation, for Humphreys et al. (2014), to examine the impact of different levels of physical activity on health outcomes. They employed an instrumental variable approach as well as a probit model to obtain causal interpretations to their estimates. Their probit model contained six dependent variables that represent different health outcomes, such as self-reported health, diabetes, high blood pressure, heart disease, arthritis, as well as asthma. They established that participating in physical activity has a clear effect on different health outcomes. They found that being physically active decreases the probability of reporting poor health as well as decreases the probability of having different diseases such as arthritis, high blood pressure, and diabetes. Finally, they argued that there are diminishing marginal returns to physical activity,

as the effect on health outcomes is higher when moving from no physical activity participation to some engagement in physical activity, compared to the effect of moving from moderate to vigorous physical activity participation.

Sarma et al. (2015) examined the effect of leisure-time physical activity on obesity and other health outcomes. They employed both an instrumental variable approach and a probit model to obtain causal effects rather than just statistical associations. They used the average monthly temperatures as their instrumental variable for physical activity. Moreover, their probit model contained five different dichotomous health outcomes, such as having diabetes, high blood pressure, and heart disease, as well as, obesity and being overweight. They argued that not controlling for work-related physical activity in the model would lead to specification bias. This is due to the fact that people who work in physically demanding jobs tend to have lower leisure-related physical activity as a result of being physical exhausted from their job. Furthermore, they found that individuals who perform active levels of leisure-time physical activity enjoy a 5 percentage point decrease in their probability of being obese. The probability of being obese decreases substantially by 11 percentage points if the individuals also perform some work-related physical activity. Furthermore, opposite of the findings of Humphreys et al., Sarma et al. did not find any causal effects of leisure-time physical activity on other health outcomes, such as heart-related diseases and diabetes. They concluded that the disparity is a result of using different instruments, and that their own results are stronger due to the inclusion of work-related physical activity in their model.

Other, more relevant, papers focused on examining the effect of physical activity on the weight of individuals. Tao Zhang (2017) studied the effect of physical activity on obesity in China. Specifically, he wanted to examine the causal effect between the two variables. He found that participating in physical activity can significantly reduce an individual's body mass index (BMI). Moreover, he raised the question that there might be reverse causality, as obesity

might influence the individuals' ability to participate in physical activity; however, he did not find any clear evidence that obesity has an impact on being physically active. Similarly, Parsons et al. (2005) examined whether physical activity has an effect on the body mass index (BMI) from adolescence to mid-adulthood in Britain. They found that BMI is positively correlated with the age of individuals. Moreover, they found that participating in physical activity decreases age-related gains in BMI; however, the magnitude of the decrease varies with age. They concluded that, on average, males are more physically active than females, and that any decrease in the level of physical activity is associated with adverse effects on BMI. One major concern for those two studies is that they did not directly address the endogeneity problem of physical activity.

Meyer et al. (2016) also studied the effect of physical activity on the body mass index, using longitudinal American data from the Coronary Artery Risk Development in Young Adults study (CARDIA). However, unlike Tao Zhang (2017) and Parsons et al. (2005), they addressed the endogeneity problem of physical activity by implementing an instrumental variable approach. They used the distance between physical activity facilities and the residences of respondents, as well as the distance between food stores and the residences of each respondent as their two instrumental variables for physical activity. As expected, they found that the body mass index is negatively associated with physical activity and smoking, and positively associated with fast food consumption. Finally, they argued that fixed-effects or random-effects regressions might be preferable to use in comparison with instrumental variable approaches in the presence of weak instruments. This is due to the fact that instrumental variable regressions, in general, are sensitive to weak instruments, therefore, including a weak instrument in the regression increases the potential of severe bias in the estimates rather than decreasing it. In addition, it drastically decreases the precision of the obtained parameter estimates.

Furthermore, the impact of physical activity has been widely studied by using physical education as a proxy for physical activity. Sabia et al. (2017) and Cawley et al. (2007) examined the effect of physical education requirements on youth body weight. Both papers exploited the variation in state requirements, taken as exogenous, for in-school physical activity as a natural experiment in order to obtain causal impacts. They both found similar positive effects of physical education requirements on the time individuals spend in being physically active. Surprisingly, both studies found no evidence that physical education decreases the probability of being overweight or obese. This might be due to the fact that the increases in physical education requirements do not lead to a large enough increase in physical activity time in order to affect weight. Nonetheless, their results are aligned with those reported in the previous literature that studied this relationship without addressing the potential endogeneity of physical activity.

Cawley et al. (2013) employed an instrumental variable approach to examine the causal effect of physical education on obesity for school children. They conducted their study using American data from the Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K) for 1998-2004. They used the states' compulsory amount of physical activity minutes per week as their instrumental variable. In contrast with the aforementioned findings of Sabia et al. (2017) and Cawley et al. (2007), they found evidence of a causal effect of physical education on the obesity of school children. They found that a compulsory increase of 60 minutes per week in physical education results in a 14% decrease in BMI levels of 5^{th} graders. Moreover, they found that the effects are mainly for boys, which might be a result of boys being more physically active during physical education activities relative to girls. These findings were among the first to obtain causal effects of physical education on obesity.

Moreover, in order to assess the importance of work-related physical activity and work hours, and whether they have a role in the endogeneity of physical activity, I look at a recent

study by Cook and Gazmararian (2018). They examined the associations between obesity, leisure-time physical activity, and work-related physical activity. Work-related physical activity is any movement or movements of an individuals' body parts, during work, that requires expenditure of energy. They found no evidence of associations between hours of work and leisure-time physical activity or obesity. This finding contradicts the general pattern of findings that leisure-time physical activity should be negatively associated with working hours, as individuals who work long hours would not have the time to participate in physical activity outside of work. They also found evidence that individuals in different occupational groups face different incidences of being obese due to different levels of physical activity demanded on the job. Therefore, the results are somewhat inconclusive. We cannot determine if work-related physical activity and hours of work have a direct effect on the endogeneity of physical activity. However, it is better to control for them to avoid any potential bias in the estimates.

Finally, on a related topic, some studies examined the effects of sport participation on the life satisfaction of individuals. Dolan et al. (2014) and Ruseski et al. (2014) studied the casual impacts of sports participation and physical activity on the outcome of life satisfaction, with both papers implementing an instrumental variable approach to obtain casual effects. Dolan et al. (2014) used the respondents' answers to a survey question of their perceived benefits of participation in physical activity as their instrumental variable. Similarly, Ruseski et al. (2014) used the respondents' beliefs about the importance of participating in physical activity and the distance between each respondents' home and the nearest sports facility as their instruments. Both studies found positive correlations between participating in physical activity and the subjective well-being of individuals, with respondents' who participate in sports or physical activity reporting higher levels of happiness relative to respondents who do not participate. Moreover, previous studies on the association effects between happiness and health suggest that happier individuals face higher rates of recovery and survival (Lamers et al., 2012).

Therefore, even though these papers do not necessarily examine the impact of participating in physical activity on health, they still support the findings of health-related studies, as previous researchers insinuate that health and happiness are positively correlated.

Despite the efforts of many researchers to estimate the effect of physical activity on different health outcomes, the causal effect still remains unclear. Thus, this paper contributes to the existing literature by attempting to obtain causal effects rather than correlation effects. Other than Humphereys et al. (2014), Sarma et al. (2015), and Meyer et al. (2016), a limited number of studies have employed an instrumental variable strategy to solve for the endogeneity of physical activity. Furthermore, most of the existing literature does not take into account the frequency, intensity and duration of physical activity, perhaps due to the lack of available data on these variables. Failure to include variables that correspond to the frequency, intensity, duration of physical activity in the regression equation would lead to omitted variable bias, as the two conditions for omitted variable bias are: the omitted variable must be correlated with one of the main explanatory variables and determinants of the dependent variable. This in turn might lead to an over-estimation of the impact of participating in physical activity, which means that the true impact is actually lower that the obtained coefficients. Therefore, in this paper I use the frequency, intensity, and duration statistics that are available in the dataset, in one of my estimating equations, in order to obtain more efficient estimates.

3. Data

The data set that I am using is the Behavioral Risk Factor Surveillance System (BRFSS) for the years 1990-2010 in the United States. The BRFSS is the leading system of health-related surveys that collects data on adult U.S. residents using telephone surveys. Currently, data are collected each month on all U.S. states regarding risk-related behaviours associated with the health of American individuals. The population of interest is American civilians between 18

and 75 years of age. The unit of observation is the individuals who completed the BRFSS survey. The analysis is conducted using cross-sectional data over a number of years.[2] The sample size is 4,892,576 observations with the lowest number of observations equal to 81,577 observations collected in 1990, and it increases with each year. There are two final samples, that consist of 3,450,273 and 595,201 observations, respectively.[3] The data are collected using non-random sampling process by conducting a questionnaire across all American states. The population was divided into categories based on education level, states and income group. The main problem with the BRFSS is the frequent change in the questions and variable names in the data. This, consequently, increases the measurement error in the regression model and, thereby, results in biased estimates and misleading patterns in the data.

The BRFSS contains data on the self-reported body mass index (BMI) calculated as kilogram of weight per meter-squared of height, which is my main outcome variable that represents the health status of each individual. Throughout the 21 years of data, the body mass index is reported in four different ways. It would be straightforward to combine all variables into one variable that represents BMI for all individuals across all years, however, the problem arises from the different definitions of BMI across the years. For the years 1990-1991 and 1996-2000, the range of reported values was from 1 to 999, with values over 300 corresponding to obese individuals, values under 300 and over 250 corresponding to overweight individuals and values under 250 corresponding to individuals that are neither obese nor overweight. Moreover, the years 1992-1995 had the same definition of obese, overweight and neither, but the values lay between 1 and 1000 instead of ending at 999. On the other hand, for the years 2002 to 2010, the BMI values were rescaled by a factor of 10. In other words, the BMI values

[2] Panel data would be ideal for this study; however, existing longitudinal data are not large enough in sample size.

[3] The difference in sample size is due to the exclusion of the instrumental variable and the frequency, duration and intensity variables from some OLS specifications. The reasons and results of their exclusion will be explained in the following sections.

started at 1 and ended at 9999, instead of 999. For these years, values over 3000, between 2500 and 3000, and under 2500 correspond to individuals who are obese, overweight, and neither obese or overweight respectively. It is worth noting that the year 2001 had an entirely different definition of the BMI, where the values lay between 1 and 999999. Thus, for that year, values over 300000, between 250000 and 300000, and under 250000 correspond to individuals who are obese, overweight, and neither obese or overweight respectively.

I converted all these BMI variables into one variable that is defined from 1 to 9999 by simply multiplying or dividing by a number in order to reach the chosen consistent range of values across all years. [4] Thus, in the newly constructed BMI variable, an overweight individual has a BMI value that lies between 2500 and 3000, an obese individual has a BMI value over 3000, and finally an individual who is neither obese nor overweight has a BMI less than 2500. Finally, respondents' who did not report their BMI or respondents with missing values for BMI were dropped; thus, only 4 observations in total were dropped from the sample.

The survey question for physical activity in the BRFSS is "During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?". Thus, physical activity is a dummy variable that indicates if each individual participated in any physical activity in the month preceding the survey. Physical activity is equal to 1 if the individual engaged in physical activity and 0 otherwise. Respondents' who refused or did not know the answer to the aforementioned question were dropped, and as a result 423,812 observations were dropped.[5] These dropped observations account to 8.66% of the overall sample size.

[4] For years where BMI ranged from 1 to 999, I multiplied by 10.009009. For the years 1992-1995, I multiplied by 9.999. Finally, for 2001, I divided by 100.009901. Therefore, the values were just rescaled to provide a consistent definition for BMI.

[5] It is worth noting that the large number of dropped observations could lead to less reliable estimates as the sample is not very representative of the overall population.

Moreover, according to previous studies, physical activity should not be measured solely as a discrete choice, but also based on the intensity, frequency and duration of that activity. As discussed before, the BRFSS has different variables that define levels of physical activity across the sample years. Fortunately, the BRFSS contains calculated variables that contain already calculated levels of physical activity that are divided into categories based on their intensity, frequency and duration. However, the number of categories changed throughout the interval of data availability. Before 2001 there were four different categories; physically inactive, irregular activity, regular activity, and regular activity combined with vigorous activity. On the other hand, after 2001, they added a new category for vigorous activity alone. Therefore, given the limited resources and the constrained definition of the variables, I concorded vigorous activity to regular and vigorous activity in later years. In other words, I merged the new vigorous activity variable into the regular and vigorous activity variable that already existed in order to obtain consistent and comparable variables.

After compiling the data, I constructed four different dummy variables that indicate the intensity, frequency and duration of the physical activity that was exerted. These variables are **Active**, **Regular**, **Irregular**, and **Inactive**, and they represent the level of physical activity that the respondents' have engaged in. In my final sample, an individual is considered to be active if he/she engages in vigorous activity for at least 3 days per week with a minimum duration of 20 minutes for each time, an individual is considered regular if he/she engages in moderate physical activity for at least 5 days per week with a minimum duration of 30 minutes for each time, an individual is considered irregular if he/she does not meet the requirement for either moderate or vigorous physical activity but does engage in some physical activity, and an individual is considered inactive if he/she has not participated in any physical activity in the month preceding the survey. Another issue was that the BRFSS did not contain any data on the intensity, frequency and duration of physical activity for the years 2002, 2001, 2004, 2006,

2008, and 2010. As a result, 2,092,856 observations were dropped from the sample which includes them in the regression model.[6]

Another important variable related to my study is smoking activity, as many of the previous studies, such as Meyer et al. (2016), emphasized that smoking is negatively correlated with the body mass index. Therefore, omitting it from my regression equation would definitely lead to omitted variable bias. The BRFSS contains data on the smoking status of each individual and, in most years, it was divided into four categories as follows: current smokers who smoke every day, current smokers who smoke somedays, former smokers, and never smoked. However, for the years 1994 and 1995, the BRFSS divided current smoker into four categories instead of just two. The final six categories were as follows: current smokers who smoked on all of the past 30 days, current smokers who smoked on 1-29 days of the past 30 days, current smokers who smoked 0 days in the past 30 days, current smoker with an unknown number of smoking days in the past 30 days, former smokers, and never smoked. Therefore, it was straightforward for me to construct an indicator variable, **Smoker**, that indicates the smoking status of each individual. **Smoker** is equal to one if the individual falls under any of the current smoker categories, whether he smokes every day or just a few days, and is equal to 0 if the individual is a former smoker but does not smoke now or has never smoked. Finally, 6,553 observations were dropped as a result of missing values or respondents refusing to declare their smoking status.

Furthermore, the BRFSS does not specify the number of years of education for each level of education, however, it consists of six categories that define each of level of education. The categories are: never attended school or only kindergarten, elementary school (grades 1-8), some high school (grades 9-11), high school graduate (grade 12), some college or technical

[6] The high number of observations dropped could potentially increase the measurement bias in the parameter estimates, which will be discussed in later sections.

school, and college graduate. Therefore, I divided the level of education into four different categories. The first one consists of individuals who did not attend school or just kindergarten. The second one consists of individuals who attended elementary school or did not complete high school. The third one consists of high school graduates or those who did not complete college or technical school. The final category consists of individuals who graduated from college. Therefore, the categorization scheme is weighted towards degree attainment, and anyone who did not complete a certain degree was placed in the lower category. Only one observation was dropped as a result of missing values for education.

Similarly, the BRFSS contains data on the annual income levels of each individual, and they are categorized as follows: individuals who earn less than $10000 per year, more than $10000 but less than $15000, more than $15000 but less than $20000, more than $20000 but less than $25000, more than $25000 but less than $35000, more than $35000 but less than $50000, and individuals who earn more than $50000. However, an additional category was added after 1994 that consisted of individuals who earn more than $75000. Therefore, I constructed eight dummy variables corresponding to each income group.[7]

The BRFSS also contains more than six different variables that describe the race of each individual. I chose to use the variable that has the least number of missing values and the most consistent reporting across the years. The race variable is also divided as follows: White Non-Hispanic, Black Non-Hispanic, White Hispanic, Black Hispanic, Other Hispanic, Asian or Pacific Islander, and Other. Therefore, I divided the race of the individuals into four different categories: **White Non-Hispanic**, **Black Non-Hispanic**, **Hispanic**, and **Other**, and I constructed four separate indicator variables to indicate the race of each individual

[7] It is worth noting that there are probably some individuals in the years 1990-1993 that earned more than $75000 but are not included in that category due to the level of aggregation of the data provided. This might lead to slight measurement error; however, this is merely a second order bias and it will not drastically affect the results.

respectively. In some years additional race categories were introduced, but I assigned them under **Other** in my categorization.[8]

Moreover, the BRFSS includes eight different categories that represent the employment status of each individual. These categories are divided as follows: employed, self-employed, out of work for more than one year, out of work for less than one year, homemaker, student, retire, and unable to work. I am only interested in individuals who are currently working: thus, I constructed an indicator variable that is equal to 1 if the individual is either employed or self-employed and 0 otherwise.

Additionally, I included other control variables that are correlated with physical activity to reduce the degree of omitted variable bias, such the marital status, age, and gender. The marital status dummy variables indicate if the individual is married, single, or divorced, respectively. Gender is indicated by a dummy variable that is equal to 1 if the individual is a female and 0 if the individual is male. Age is reported in numbers rather than in categories. Moreover, previous studies suggest that the body mass index might not be a good predictor of health for older adults. After a certain age, the need to maintain muscle mass increases in order to prevent or counter frailty. Therefore, I dropped individuals that are older than 75 years old. I also dropped individuals that are younger than 18 years old at the time of the survey.[9] Therefore, 226,825 observations were dropped due to these age restrictions.

Finally, an instrumental variable is introduced to solve for the endogeneity problem of my main explanatory variable, physical activity. My estimation strategy aims to extract the exogenous variation in physical activity to obtain unbiased causal effects. The instrumental variable is the respondents' answers to a survey question that explicitly asks, "Are you using

[8] For example, new races such as Native Hawaiian and American Indian were included in some years but are still considered as Other in my categorization.

[9] The sample contains 1,062 and 12,551 observations for 7 years old and 9 years old respondents respectively. These individuals are unsuitable for this study as they are too young to engage in physical activity or report true information regarding our variables of interest. Therefore, dropping them makes the sample more efficient and avoids having any outliers.

physical activity or exercise to lose weight or maintain your current weight?". However, the years 2004, 2006, 2007, 2008, and 2009 do not include any data on this variable, thus, causing 1,269,156 observations to be dropped. The dropped observations also include individuals who refused or failed to respond to the question of interest.

Columns (1) and (2) in Table (1) shows the weighted means and standard deviations (in brackets) of all the variables, in the samples of both specifications. Column (1) displays those of the first specification, which excludes the frequency, duration and intensity variables from the regression. The average body mass index in the sample is 2883, which means that on average the individuals in the sample just below the threshold for being overweight. Table 1 also shows that 74.6% of the sample participated in physical activity. 50.3% of the sample consists of females, and the average age in the final sample is 43 years. Furthermore, individuals that did not attend any school or only kindergarten account for 1.1% of the sample, and those who reached elementary school or some high school account for 17.0% of the sample. Individuals who graduated from high school or did not complete their postsecondary degree are the largest group in the sample with a share of 47.5%. Finally, individuals graduated from college account for 30.4% of the sample.

Moreover, 27.1% of the sample are smokers, and 67.2% are currently employed. The sample consists of a majority of White non-Hispanic individuals that account for 71.9% of the total sample. The rest of the sample consists of 9.9%, 12.8%, and 5.5% of Black non-Hispanics, Hispanic, and other races, respectively. Furthermore, the sample consists of 61.3% of married individuals, 12.0% of divorced individuals, and 4.0% of widowed individuals. Table 1 also shows the percentage of people in each income group, where richer income groups have a higher share of people. Individuals in the lowest income group, which consists of individuals with an annual income that is less than $10000, account for 7.4%. On the other hand, individuals in the highest income group account for 19.3% of the sample, which is the largest

share in the sample, and it consists of individuals with an annual income that is larger than $50,000 but lower than $75,000. Finally, 24.5% of the individuals in the sample reported that they have diabetes.

Column (2) displays the weighted means and standard deviations of the variables in the sample which includes the frequency, duration, and intensity variables. Column (2) shows that the average body mass index in this sample is 2929, which is just slightly higher than the average in the first sample. Moreover, 21.4% of the sample reported that they were inactive in the month preceding the conduction of the survey. 31.0% of the sample participated in irregular activity, which means that they participated in some kind of physical activity, but it was insufficient to meet the moderate or vigorous recommendations. 30.2% of the sample participated in regular activity, which means that they engaged in 30 minutes of moderate physical activity at least 5 times per week. Finally, 17.4% of the sample were active, which means that they engaged in 20 minutes of vigorous physical activity at least 3 times per week. The average age in the sample is 43 years, and 55.7% are females.

Furthermore, 1.0%, 15.1%, 50.7%, and 29.4% of the sample have no education, low levels of education, medium levels of education, and high levels of education, respectively. Individuals with the attributes of being employed, smokers, married, divorced, and widowed account for 68.4%, 25.1%, 62.6%, 12.3%, and 4.3%, respectively. This sample consists of 75.5%, 9.7%, 10.6%, and 4.2% of white non-hispanics, black non-hispanics, hispanics, and other races, respectively. Individuals with the lowest and highest annual income account for 7.9% and 13.7%, respectively. Moreover, individuals with annual income levels that lay between $35000 and $50000 make up the largest income group, with a share of 19.4% of the sample. Finally, 61.6% of the individuals in the sample reported that they participated in physical activity in order to lose or maintain their current weight, while 21.6% reported that they have diabetes.

Despite the large disparity in the number of observations between the two samples, the descriptive statistics did not alter much. The average body mass index slightly increased from 2883 to 2929 when the sample size was reduced. However, it remains just below the threshold for being overweight in both samples. The average age is identical in both samples with a mean of 43 years. On the other hand, the share of females increased from 50.3% to 55.7% after the reduction of the sample size. Furthermore, the weighted means of the remaining variables remained mostly similar in both samples. Despite this similarity, the huge reduction in the number of observations raises a concern for the reproducibility and the reliability of the survey's results.

4. Identification Strategy

The econometric model takes the following form:

$$lBMI_{it} = \beta_0 + \beta_1 Phys_{it} + X_{it}\gamma + \delta_s + \lambda_t + \varepsilon_{it} \quad (1)$$

where $lBMI_{it}$ is the log of body mass index (BMI) for each individual i at time t, $Phys_{it}$ is a dummy variable indicating whether individual i participated in physical activity at time t, and X_{it} is a vector of explanatory variables and demographic controls for each individual i at time t that allow us to reduce the degree of omitted variable bias and thus increase the reliability of my estimates. I use the log of body mass index rather than the level values of body mass index, as I am more interested in obtaining the relative change effects rather than the absolute change effects of participating in physical activity on the body mass index.[10] X_{it} also includes dummy

[10] Using the log of the dependent variable can facilitate the analysis of patterns in the data by normalizing the data. In this study, for example, if an individual has a BMI value of 1000, an increase of 100 points in BMI is significant and equivalent to a 10% increase in BMI. On the other hand, if an individual has a BMI value of 6000, an increase of 100 points is equivalent to a 1.7% increase in BMI. Therefore, the relative change gives more insight about the changes in BMI compared to the absolute change.

variables that indicate the age, gender, race, education level, income level, smoking status, and marital status of each individual *i* at time *t*. Furthermore, in some specifications, the vector X_{it} includes four indicator variables that reflect the intensity, frequency, and duration of the aforementioned physical activity. Finally, δ_s and λ_t are the parameters for the state and time fixed effects respectively, which include state, month and year dummy variables. The month and year fixed effects will pick up any variation in the outcome variable that is specific to a certain time, while the state fixed effects will pick up any variation in the outcome that is specific to a certain state. Finally, ε_{it} is the error term for each individual *i* at time *t*. The coefficients of equation (1) do not have any causal interpretations unless physical activity is taken as exogenous.

In order to address the possible endogeneity of physical activity, I adopt an instrumental variable strategy to be able to examine the casual effects of participating in physical activity on the body mass index of Americans. There are two main assumptions for a valid instrument, namely, instrument relevance and instrument exogeneity. Instrument relevance implies that the instrument should to be correlated with the endogenous explanatory variable, while the instrument exogeneity, or the exclusion restriction, implies that the instrument should not affect the outcome except indirectly through the channel of the endogenous explanatory variable.

$$Phys_{it} = \beta_0 + \beta_1 Post_{it} + X_{it}\gamma + \delta_s + \lambda_t + \varepsilon_{it} \qquad (2)$$

equation (2) is the first stage of the 2-stage least squares regression, in which I predict physical activity using the instrumental variable, $Post_{it}$. $Post_{it}$ is a dummy variable that indicates the explicitly stated motive behind participating in physical activity for each individual i at time t. It assumes a value of 1 if losing or maintaining their weight was their motive behind

participating in physical activity and 0 if that was not their motive behind engaging in physical activity. X_{it}, δ_s and λ_t are the same controls and fixed effects included in equation (1).

$$lBMI_{it} = \beta_0 + \beta_1 \widehat{Phys}_{it} + X_{it}\gamma + \delta_s + \lambda_t + \varepsilon_{it} \quad (3)$$

equation (3) is the second stage of the 2SLS regression, in which I regress the log of body mass index of the predicted value of physical activity $\widehat{Phys}_{it}$ and the same set of controls and fixed effects. Therefore, β_1 from (3) represents the treatment effect of participating in physical activity on the body mass index of each individual. Regressions are conducted before and after adding state and time fixed effects. The frequency, intensity, and duration variables are not included in the second stage IV regressions due to the existence of multicollinearity between them and my main explanatory variable, physical activity. Therefore, adding them to the IV regression without adding additional instruments leads to biased and unreliable estimates.

5. Results – Dependent Variable: Ln(BMI)

The main goal of regression analysis is to estimate the effect of each independent variable on the dependent variable, holding all other independent variables constant. The key to obtaining reliable estimates is based on the models' ability to isolate the association effect of each covariate on the outcome variable. In other words, it is crucial to be able to independently estimate the relationships between the dependent variable and each independent variable, respectively. However, significantly high correlations between independent variables makes it difficult to estimate these relationships independently. Highly correlated covariates tend to move together, a change in one of them is associated with changes in the other correlated variable. Consequently, serious problems of collinearity or multicollinearity can arise in the regression analysis.

According to previous research, the simplest way to test for multicollinearity is to check if the correlation between any of the suspected explanatory variables is higher than 0.8 (Gujarati, 2004). Generally, the rule of thumb for testing for the existence of collinearity or multicollinearity between variables is correlation coefficients that lie between 0.7 and 0.9 (Yoo et al., 2014). Another way to test for multicollinearity is using the variance inflation factor (VIF), which identifies the strength of the correlation effects. VIF values that are greater than 5 imply that there exist severe levels of multicollinearity, while values that lie between 1 and 5 indicate moderate levels of multicollinearity between the variables (Yoo et al., 2014). The literature also indicates that the existence of multicollinearity results in an increase in the variance of the estimated coefficients, thus, decreasing the probability of rejecting the null hypothesis and obtaining reliable statistically significant estimates. This, in turn, reduces the precision and reliability of the obtained estimated coefficients in the regression. Moreover, severe collinearity could also lead to changes in the estimated direction of the association effects of these estimates and increases their sensitivity to small changes in the model or in the sample.

In my model, correlation tests show that at least one of the frequency, duration and intensity variables is highly correlated with physical activity. For example; the correlation between my main physical activity variable and **Inactive** is -0.83, which implies that there is high potential of multicollinearity if they are included together in the regression equation. Moreover, the VIF test indicates moderate multicollinearity between physical activity and the variables that correspond to irregular, regular, and active levels of physical activity, with VIF values equal to 1.79, 1.76, and 1.60, respectively. A VIF value of 1.79 means that the variance of the estimated coefficient for irregular physical activity is 79% higher than it would be if there was no multicollinearity Therefore, based on the findings of the literature, it is safe to say that there exists moderate to high multicollinearity between the frequency, duration and intensity

variables and my main explanatory variable, physical activity. The inclusion of all these problematic variables in the same regression would definitely lead to loss of power in the model, as well as unreliable estimates. Therefore, I decided to separate the frequency, duration and intensity variables from my main explanatory variable, physical activity, in the regression analysis. Thus, in order to account for this multicollinearity issue, I estimated two separate OLS regressions by including the physical activity variable or the frequency, duration, and intensity variables in separate regressions.

The results of the regressions are shown in Table (2), where each column represents a different specification. Columns (1) and (2) display the Ordinary Least Squares (OLS) regression results, when excluding the frequency, duration and intensity variables. Column (1) shows that, without including the state and time fixed effects, participating in physical activity is statistically associated with a 5.0% decrease in an individuals' body mass index, holding all other factors constant. Ceteris paribus, an additional year of age is associated with a 0.2% increase in BMI. The attribute of being a female is associated with a 4.4% decrease in BMI relative to being a male. Relative to individuals with no education, all levels of education are positively associated with BMI, with a decrease in the magnitude of the association effect as individuals obtain higher levels of education. Low, medium, and high levels of education are associated with increases of 5.7%, 7.2%, and 2.7% in BMI, respectively. The attribute of being a smoker is associated with a 4.6% decrease in BMI levels, holding all other factors constant.

Furthermore, being employed, married, or divorced are statistically associated with positive increases in BMI, with increases of 1.4%, 2.7%, and 1.4%, respectively. On the other hand, the attribute of being widowed is associated with a 0.1% decrease in BMI, holding all other factors constant. Moreover, the estimates of the racial categories show that, relative to white non-hispanics, black non-hispanics and hispanics are associated with increases of 6.6% and 7.0% in BMI, respectively, whereas other races are associated with a 2.0% decrease in BMI. Finally,

holding all other factors constant, income levels are negatively associated with BMI relative to the lowest income group. The second lowest level of income is associated with a 1.5% decrease in BMI, while the highest level of income is associated with a 3.5% decrease in BMI. Generally, the negative association effects of income on BMI increase as income increases. All of the estimated coefficients are statistically significant at the 1% level, except the estimated coefficient of **Widowed**.

Column (2) shows the results of the OLS regression after including the state and time fixed effects but excluding the frequency, duration, and intensity variables. It shows that participating in physical activity is associated with a 5.0% decrease in BMI, holding all other factors constant. Holding all other factors constant, the attributes of being a female or a smoker are associated with decreases of 0.4% and 4.5% in BMI, respectively. Age is associated with an increase of 0.2% in BMI, with each additional year. Employment, marriage and divorce are associated with positive increases of 1.6%, 2.9%, and 1.4% in BMI, while being widowed in associated with a 0.01% decrease in BMI.

Moreover, relative to having no education, education levels are associated with positive effects on BMI, with a decrease in the degree of these effects as education levels increase. These magnitudes of these effects are 4.3%, 5.1%, and 1.1% for low, medium, and high levels of education, respectively. On the other hand, income levels are associated with negative effects on BMI, with increasing negative effects as income levels increase. Finally, relative to white non-hispanics, black non-hispanics and hispanics are positively associated with BMI, whereas other races are negatively associated with BMI. The statistical significance of the estimated coefficients did not change after the inclusion of the time and state fixed effects.

Columns (3) and (4) display the OLS regression results after excluding the primary measure of physical activity and including the frequency, duration, and intensity variables, where they

display the results before and after including the state and time fixed effects, respectively. As aforementioned, the variable physical activity is omitted from both regressions due to its high collinearity with the frequency, duration, and intensity variables. Column (3) shows that, relative to being physically inactive, participating in any level of physical activity is associated with a decrease in an individuals' body mass index. Holding all other factors constant, engaging in irregular physical activity is associated with a 2.9% decrease in an individuals' body mass index in comparison with individuals that are inactive. Engaging in regular physical activity is associated with a 5.4% decrease in an individuals' body mass index, and engaging in active physical activity is associated with a 7.8% decrease in an individuals' body mass index. Therefore, there is an increasing gradient in the magnitude of the association effects as the individual moves from low to higher intensity levels of physical activity.

Column (3) also shows that, on average, females have lower body mass indices compared to men, holding all other factors constant. This difference in BMI between men and women is estimated to be 1.4%. Age is positively associated with BMI, with an increase of 0.2% for each additional year. Moreover, holding all other factors constant, being a smoker is associated with a 2.2% decrease in BMI. Relative to white non-Hispanics, black non-hispanics and hispanics are associated with increases of 6.9% and 3.9% in BMI, respectively, while individuals from other races are associated with a decrease of 3.2% in BMI. Being employed, married, divorced, or widowed are associated with increases of 1.1%, 2.5%, 1.5% and 0.4% in BMI, respectively. Income is associated with negative and increasing, in magnitude, effects on BMI as an individual becomes richer. Finally, education seems to have mixed effects, as low and medium levels of education are associated with increases of 3.1% and 0.1% in BMI, relative to having no education. On the hand, high levels of education are associated with a 2.5% decrease in BMI. The statistical significance levels of the estimated coefficients remain unchanged after the alteration of the variables in the regression equation.

Column (4) displays the estimated results of the second specification after including the state and time fixed effects in the model. Column (4) shows that, holding all other factors constant, engaging in irregular, regular, and active levels of physical activity are statistically associated with decreases of 3.6%, 5.5%, and 8.7% in BMI, respectively, relative to being physically inactive. Holding all other factors constant, being a smoker is associated with a 2.4% decrease in BMI. Age remains positively associated with BMI, with an increase of 0.2% with each additional year. Relative to being a male, being a female is associated with a 1.5% decrease in BMI, holding all other factors constant. Moreover, the estimated coefficients of the remaining control variables did not vary much after the inclusion of fixed effects in the model. Other than the estimated coefficient of the variable that corresponds to being a widow, which changed signs from negative to positive, the signs of all of the other estimates remained unchanged after including the fixed effects in the model. However, the magnitude of these estimates varied slightly. Similarly, there was little change in the significance levels of the estimated coefficients. The estimated coefficient of the variable the corresponds to medium level of education became statistically insignificant at the 1% level. On the other hand, the statistical significance levels of the other variables remained unchanged.

The results of the four OLS specifications are consistent with most of the findings of previous studies, despite not yet accounting for the endogeneity of physical activity. Participating in physical activity and the frequency, duration, and intensity variables are negatively associated with an individuals' body mass index. Holding all other independent variables constant, participating in physical activity is statistically associated with decreases of 5.04% and 5.00% in BMI, before and after including state and time fixed effects, respectively. Relative to physical inactivity, engaging in irregular, regular, and active levels of physical are associated with decreases 2.9%, 5.4%, and 7.8% in BMI, before the inclusion of the fixed effects. After including the fixed effects in the model, these association effects increased to

3.6%, 5.5%, and 8.7% for each level of physical activity, respectively. Similarly, the majority of previous studies found that physical activity has a negative association effect on BMI (Humphreys et al., 2014; Parsons et al., 2005; Tao Zhang 2017; Meyer et al., 2016). For example, Humpherys et al. (2014) found that the marginal impact of participating in physical activity increases in magnitude as an individual increases his/her frequency, duration, and intensity levels. Moreover, the attribute of being a smoker has a negative association effect on BMI, holding all other factors constant (Meyer et al., 2016). Furthermore, higher levels of income and education are associated with lower levels of BMI, while age is positively associated with BMI, holding all other factors constant (Humphreys et al., 2014; Tao Zhang 2017).

After including the state and time fixed effects, the direction of the association effects on BMI, for my main explanatory variables of interest, remained unchanged. However, the magnitude of these effects varied after the inclusion of the fixed effects in the regression. The negative association effect of participating in physical activity on BMI slightly decreased from 5.04% to 5.00%, on the other hand, the negative association effects of the frequency, duration, and intensity variables increased in magnitude. The magnitude of these association effects increased from -2.9% to -3.6%, from -5.4% to -5.5%, and from -7.8% to -8.7% for irregular, regular, and active levels of physical activity, respectively. Hence, the OLS and FE-OLS estimates show that including the state and time fixed effects in the equations had two main effects: a decrease in the association effect of participating in physical activity and an increase in the association effects of the frequency, duration, and intensity variables. This demonstrates that there is a slight bias in the estimated coefficients when the fixed effects are excluded. The impact of physical activity was overestimated, or biased-upwards, while the impacts of the frequency, duration, and intensity variables were underestimated, or biased-downwards, before

their inclusion. Thus, outlining the importance of the fixed effects in obtaining more efficient and reliable estimates.

The aforementioned estimated coefficients of the OLS and FE-OLS regressions still contain a considerable amount of bias as a result of not accounting for the endogeneity of physical activity. The endogeneity of physical activity could be due to the exclusion of variables that are correlated with both physical activity and the dependent health outcome, BMI. For example, urbanization is correlated with both physical activity and BMI. Urbanization might decrease an individuals' ability to participate in physical activity due to the decrease of the amount of green space available, while it can affect an individuals' health as a result of the increase in pollution. Unobserved heterogeneity, such as unique individual characteristics like an individuals' rate of leisure time preference, his/her level of enjoyment when participating in physical activity, as well as the opportunity cost of time, could also lead to the endogeneity of physical activity. Controlling for any of these variables, if they are a source of the endogeneity of physical activity, is essential in order to obtain consistent parameter estimates.

Another major source of bias in the estimates is the potential presence of reverse causality between participating in physical activity and an individuals' body mass index. Participating in physical activity might cause a reduction in BMI levels, however, it could also be the case that individuals choose to participate in physical activity due to their BMI levels. In other words, the decision about participating in physical activity could be driven, in part, by an individuals' BMI. Additionally, there are various sources of measurement error, such as the frequent change of the survey questions and variable definitions in the BRFSS, throughout the years in the sample. More importantly, the enormous number of dropped observations, due to the exclusion of the main explanatory variables in some years, which leads to less random variables. These factors lead to high measurement error and, therefore, increase the bias in the

estimates. Thus, I decided to employ an instrumental variable approach in order to account for these issues, and thereby, obtain causal effects.

An instrumental variable analysis employs a new variable, the instrumental variable, in order to extract exogenous variation in the endogenous variable that is unrelated to the outcome variable and hence the disturbance term. Therefore, allowing the estimates to have causal interpretations. There are two main assumptions for a strong instrument: its validity and its relevance. Instrument validity entails that the instrument should not be correlated with the error term in the regression, while instrument relevance means that the instrument should be correlated with the endogenous variable. In other words, strong instruments need to be correlated with physical activity and uncorrelated to BMI, except through the channel of participating in physical activity.

Finding instrumental variables that satisfy the aforementioned conditions for a valid instrument proved to be troublesome. Previous studies have used the temperature of each state as an instrumental variable for physical activity, therefore, I tried using the monthly average temperature in each state as an instrumental variable in my study (Sarma et al., 2014). However, the correlation between temperature and my endogenous physical activity variables was too weak for it to be considered a valid instrument. Other studies used neighborhood-level measures as their instrumental variable. These measures included the presence of physical activity facilities and restaurants around an individuals' habitat, as well as the distance between the nearest physical activity facility and an individuals' home. However, available data for similar instruments proved to be difficult to find.

In this study, I employed the individuals self-reported answers to the survey question: "Are you using physical activity or exercise to lose weight or maintain your current weight?", as my instrumental variable. The decision behind employing this instrument was a previous study that

used a similar instrument, which was the answers of respondents on whether they believed that physical activity was important (Ruseski et al., 2014). The motive behind participating in physical activity is definitely correlated with physical activity, thus, implying that the instrument relevance condition is satisfied. However, if there is indeed reverse causality between participating in physical activity and BMI, then the instrument validity condition might not hold. Individuals might choose to engage in physical activity in order to lose or maintain their weight as a result of their BMI. Moreover, the first-stage F-statistics are 127,820 and 126,354, before and after including the state and time fixed effects, respectively. These F-statistics imply that the instrument used has almost identical series to those of the endogenous variable, physical activity. In other words, the correlation between the instrument and physical activity seems to be mechanical. This implies that the instrumental variable used is weak and the estimated coefficients using this instrument would be biased and unreliable due to its inability to correct for the aforementioned consequences of endogeneity. Hence, unfortunately, this study was unable to obtain causal effects of participating in physical activity on BMI.

Finally, I conduct the Durbin-Wu-Hausman test to assess whether the IV estimates are significantly different from the OLS estimates. In other words, the test examines whether physical activity is actually endogenous. The test produced p-values equal to 0.00 for both specifications. Therefore, the null hypothesis, which states that physical activity is exogenous, is rejected.

It is important to keep in mind that when using an instrumental variables approach, the estimates only show the local average treatment effects (LATEs) for compliers. In this case, compliers are individuals who change their physical activity status as a result of using physical activity as a tool to lose or maintain their current weight. Therefore, if the instrument used was strong, it would identify the causal effect of participating in physical activity on BMI for those

who participated in physical activity due to the motivational effects of losing or maintaining their current weight.

6. Results – Dependent Variable: Binary Diabetes Variable

In this section, I aim to check the strength of my results by examining if they are sensitive to changes. I introduce a new outcome variable that represents an individuals' health, using a linear probability model. The variable that I introduced is diabetes, which is a dummy variable that indicates whether a respondent suffers or does not suffer from diabetes. It is equal to 1 if the individual suffers from diabetes and 0 otherwise. Similar to the previous OLS regressions, I estimate two regressions that include the physical activity variable and the frequency, duration, and intensity variables, separately, due to the aforementioned issue of multicollinearity. Moreover, due to the importance of the fixed effects that I discussed in the preceding analysis, I include the state and time fixed effects in both specifications.

Column (1) in Table (3) displays the estimated results of the OLS regressions, when physical activity is included as the main explanatory variable. Column (1) shows that, holding all other factors constant, participating in physical activity is associated with a decrease of 1.4 percentage points in the probability of having diabetes. Age is positively associated with diabetes, with an increase of 0.1 percentage points in the probability of having diabetes for each additional year. Relative to men, being a female is associated with a 1.0 percentage point increase in the probability of suffering from diabetes, holding all other factors constant. Furthermore, ceteris paribus, the attribute of being a smoker is associated with an increase of 3.9 percentage points in the probability of having diabetes. The attributes of being married, divorced, or widowed are positively associated with diabetes, while being employed is negatively associated with diabetes.

Moreover, all levels of education are negatively associated with diabetes, with decreases of 19.4, 43.1, and 35.6 percentage points, in the probability of having diabetes, for low,

medium, and high levels of education, respectively. Relative to white non-hispanics, being a black non-hispanic is positively associated with the probability of having diabetes; on the other hand, the attributes of being hispanic and from other races are negatively associated with the probability of having diabetes. Finally, income levels are associated with negative and increasing effects on the probability of having diabetes as levels of income increase. On the contrary, relative to the lowest income group, individuals in the second lowest income group seem to be positively associated with the probability of having diabetes; however, the estimated coefficient of this variable is statistically insignificant at the 1% level. All of the other estimated coefficients are significant at the 1% level.

Column (2) in Table (3) displays the OLS regression estimates in which the frequency, duration, intensity variables are the main explanatory variables of interest. Column (2) shows that engaging in any level of physical activity is negatively associated with the probability of having diabetes, relative to being physically inactive. Engaging in irregular, regular, and active levels of physical activity are associated with decreases of 1.0, 1.1, and 2.3 percentage points in the probability of having diabetes. Age remains positively associated with diabetes, with an increase of 0.02 percentage points in the probability of having diabetes for each additional year. Holding all other factors constant, being a smoker is associated with a 0.8 percentage point decrease in the probability of having diabetes.

Moreover, there is a positive relationship between all races and diabetes, relative to white non-hispanics. Being married, divorced, or widowed are associated with 0.8, 0.4, and 0.1 percentage points increases in the probability of having diabetes, respectively, while being employed is associated with a 2.6 percentage point decrease in the probability of having diabetes. Furthermore, medium and high levels of education are negatively associated with the probability of having diabetes, with decreases of 1.2 and 1.9 percentage points, respectively; on the other hand, low levels of education are associated with a 0.4 percentage point increase

in the probability of having diabetes, relative to individuals with no education. Finally, negative relationships between income levels and diabetes persist in this specification, with higher association effects as income levels increase, holding all other factors constant. All of the coefficients are significant at the 1% level, except for the estimated coefficients of the variables that correspond to low levels of education and being a female.

The estimated results of both specifications outline, to some degree, the robustness of the findings of this study. Aligned with my findings, as well as the findings of previous studies, participating in physical activity and engaging in different levels of physical activity have a negative relationship with the probability of having diabetes (Humphreys et al., 2014; Sarma, et al., 2014). The negative relationships between health outcomes and levels of education and income persist, with higher income and education levels being associated with better health outcomes for individuals. Moreover, health outcomes deteriorate as individuals grow older, holding all other factors constant. The results show that the attribute of being a smoker has mixed statistical effects on the probability of diabetes. Smoking has a negative relationship with diabetes in the second specification. However, contrary to the literature and my previous results, being a smoker seems to have a positive relationship with the probability of having diabetes in the first specification. This raises the question that there might be reverse causality between participating in physical activity and smoking. Individuals might decide to engage in physical activity due to the fact that they smoke. However, this issue is not very clear and needs further investigation.

Despite the similarity in results, the estimated results remain unreliable. The inability to fully address the issues of endogeneity and reverse causality causes bias and reduces the precision of the obtained estimates. Additionally, the frequent changes in the survey questions and the definition of variables in the BRFSS increases the probability of having measurement error in the estimates. Consequently, this leads to obtaining inconsistent estimates. In other

words, despite having a large sample size, the obtained parameter estimates will not converge to their true values.

Finally, I re-estimate the FE-OLS regression with BMI as my dependent health outcome variable, using both the physical activity and the frequency, duration, and intensity variables as the main explanatory variables in the equation, in order to examine if the multicollinearity issue will have drastic effects on the estimated coefficients. The results of this regression show that the direction of the statistical effects did not change after the inclusion of all the variables in the equation. However, the estimated coefficient of one the main variables of interest, irregular physical activity, became statistically insignificant. Multicollinearity reduces the statistical power of the regression model and makes it harder to identify the statistical significance of the independent variables. Therefore, this test confirms that there is multicollinearity between the variables and that the issue needs to be fully addressed before including them together in the regression model.

7. Conclusion

The goal of this paper was to estimate the causal effects of participating in physical activity on the body mass index (BMI) for Americans. The dataset used is the Behavioral Risk Factor Surveillance System (BRFSS) for the years 1990-2010. I use an Instrumental Variable (IV) framework, with the respondents' answers to the survey question "Are you using physical activity or exercise to lose weight or maintain your current weight?", as the instrumental variable for my main variable of interest, physical activity, in order to address the endogeneity issue.

The OLS and FE-OLS regression results suggest that, holding all other factors constant, participating in physical activity is negatively associated with body mass index, with decreases of 5.04% and 5.00% in BMI for each specification, respectively. The results also show that, holding all other factors constant, age is positively associated with BMI in both specifications.

Those two results are consistent with the results of previous studies that found negative effects for physical activity and a positive effect for age (Tao Zhang, 2017; Parsons et al., 2005). Moreover, consistent with the literature, the results of both regressions suggest that engaging in irregular, regular, and active levels of physical activity have negative relationships with BMI, relative to being inactive (Humphreys et al., 2014; Parsons et al., 2005). Ceteris paribus, the OLS regression shows that engaging in irregular, regular, and active levels of physical activity are associated with 2.9%, 5.4%, and 7.8% decreases in BMI, relative to physical inactivity. On the other hand, after including the state and time fixed effects, the negative association effects increased in magnitude to be 3.6%, 5.5%, and 8.7%, respectively. The increase in the magnitude of the association effects of engaging in different levels of physical activity and the decrease in the magnitude of participating in physical activity, outline the important role of the state and time fixed effects in reducing, to a certain degree, the chance that changes in the dependent variable occur due to omitted variables rather than the independent variables of interest that are included in the regression model.

Furthermore, consistent with the previous studies, the results show that smoking has a negative relationship with BMI, holding all other factors constant (Meyer et al., 2016). Education and income levels appear to be negatively associated with BMI, holding all other factors constant (Humphreys et al., 2014; Tao Zhang 2017). These association effects increase as an individual becomes richer or more educated. Moreover, relative to white non-hispanics, black non-hispanics and hispanics are associated with positive effects on BMI, on the other hand, other races have a negative relationship with BMI. Finally, being employed, married, or divorced are positively associated with BMI, holding all other factors constant.

Ordinary Least Squares estimates do not have any causal inferences due to not addressing the endogeneity of physical activity as well as other potential issues. There is a strong case for reverse causality between the participation in physical activity and the body mass index. The

individuals' decision to engage in physical activity could be driven, in part, by his/her body mass index. Overweight and obese individuals might be motivated to participate in physically active sports due to their high body mass indices. Moreover, the estimated variables contain a great amount of measurement error. The survey questions and the definition of variables are extremely inconsistent in the BRFSS, which increases the degree of measurement error in the regression model. Self-reported data are also a potential source of measurement error, as literature suggests that people tend to report higher height values and lower weight values, compared to the true figures. All these issues add to the endogeneity of physical activity and increase the bias in the parameter estimates.

An instrumental variable regression model that employs a strong instrument can extract enough exogenous variation in the physical activity variable that is unrelated with the health outcome variable, in order to obtain causal effects. Unfortunately, finding a valid strong instrument proved to be extremely problematic. Despite its similarity to other instruments used in previous studies, the instrument used in this study was unable to extract enough exogenous variation in physical activity. Therefore, the regression model was unsuccessful in obtaining the causal effects of participating in physical activity and the aforementioned issues still persist.

Finally, I use the linear probability model in order to estimate the statistical effects of my main variables of interest on the probability of having diabetes. The estimated coefficients suggest that participating in any level of physical activity is negatively associated with the probability of diabetes. The results are consistent with the literature and supported the findings of this study, as they partially confirm that engaging in physical activity is associated with better health outcomes.

In conclusion, this study was unable to obtain causal effects of participating in physical activity the body mass index of individuals. Further research is needed in order to obtain strong and relevant instruments to be able to correct for the endogeneity and reverse causality in the

main explanatory variables. A more consistent dataset with well-defined variables is necessary in order to decrease the degree of measurement error in the estimates, which in turn, would go a long way in obtaining better, unbiased, and reliable estimates. Finally, a dynamic panel model would be more suitable and can yield more solid results for similar future studies; however, currently available longitudinal data are relatively small in sample sizes.

References

Cawley, J., Frisvold, D., and Meyerhoefer, C.: 2013. The Impact of Physical Education on Obesity Among Elementary School Children. Journal of *Health Economics* 32, 743-755.

Cawley, J., Meyerhoefer, C., and Newhouse, D.: 2007. The Impact of State Physical Education Requirements on Youth Physical Activity and Overweight. *Journal of Health Economics* 16, 1287-1301.

Centers for Disease Control and Prevention.: 2018. The Benefits of Physical Activity.

Cook, M. A., and Gazmararian, J.: 2018. The Association Between Long Work Hours and Lesure-Time Physical Activity and Obesity. *Preventive Medicine Reports* 10, 271-277.

Dolan, P., Kavetsos, G. and Vlaev, I.: 2014. The Happiness Workout. *Social Indicators Research* 119, 1363-1377.

Grossman, M.: 1972. On the Concept of Health Capital and the Demand for Health. *The Journal of Political Economy* 80(2), 223-55.

Gujarati, D. N.: 2003. Basic Econometrics. *Berkeley, CA: Osborne McGraw-Hill*.

Humphreys, B. R., Mcleod, L. and Ruseski, J. E.: 2014. Physical Activity and Health Outcomes: Evidence from Canada. *Journal of Health Economics* 23(1), 33-54.

Lamers, S. M. A., Bolier, L., Westerhof, G. J., Smit, F., and Bohlmeijer, E. T.: 2012. The Impact of Emotional Well-being on Long-term Recovery and Survival in Physical Illness: A Meta-analysis. *Journal of Behavioral Medicine* 35(5), 538-547.

Meyer, K. A., Guilkey, D. K., Tien, H., Kiefe, C. I., Popkin, B. M., and Gordon-Larsen, P.: 2016. Instrumental-Variables Simultaneous Equations Model of Physical Activity and Body Mass Index: The Coronary Artery Risk Development in Young Adults (CARDIA) Study. *American Journal of Epidemiology* 184(6), 465–476.

Parsons, T. J., Manor, O., and Power, C.: 2006. Physical Activity and Change in Body Mass

Index from Adolescence to Mid-adulthood in the 1958 British Cohort. *International Journal of Epidemiology* 35(1), 197-204.

Ruseski, J. E., Humphreys, B. R., Hallman, K., Wicker, P., and Breuer, C.: 2014. Sport Participation and Subjective Well-Being: Instrumental Variable Results from German Survey Data. *Journal of Physical Activity and Health* 11(2), 396–403.

Sabia, J. J., Nguyen, T. T., and Rosenberg, O.: 2017. High School Physical Education Requirements and Youth Body Weight: New Evidence from the YRBS. *Journal of Health Economics 26*(10), 1291-1306.

Sarma, S., Devlin, R. A., Gilliland, J., Campbell, M. K., and Zaric, G. S.: 2014. The Effect of Leisure-Time Physical Activity on Obesity, Diabetes, High BP and Heart Disease Among Canadians: Evidence from 2000/2001 to 2005/2006. *Journal of Health Economics* 24(12), 1531-547.

Stock, J., and Yogo, M.: 2005. Testing for Weak Instruments in Linear IV Regression. *Cambridge University Press*, 80-108.

World Health Organization: 2017. Global Strategy on Diet, Physical Activity and Health.

Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q., Lillard, J. W.: 2014. A Study of Effects of MultiCollinearity in the Multivariable Analysis. *National Institutes of Health* 4(5), 9-19.

Zhang, T.: 2017. Modeling the Effect of Physical Activity on Obesity in China: Evidence from the Longitudinal China Health and Nutrition Study 1989–2011. *International Journal of Environmental Research and Public Health* 14(8).

Table 1. Summary Statistics: Means and standard deviations in brackets

Explanatory Variables	Weighted Means	
	(1)	(2)
A. Health Outcome		
Body Mass Index (BMI)	2883 (1353)	2929 (1353)
B. Participation in Physical Activity		
Physical Activity	0.746 (0.435)	-
C. Frequency, Duration, and Intensity		
Inactive	-	0.214 (0.410)
Irregular	-	0.310 (0.462)
Regular	-	0.302 (0.459)
Active	-	0.174 (0.379)
D. Gender		
Female	0.503 (0.500)	0.557 (0.497)
E. Educational Attainment		
No High School	0.011 (0.105)	0.010 (0.098)
Elementary School (Some high)	0.170 (0.376)	0.151 (0.358)
High School (Some college)	0.475 (0.499)	0.507 (0.500)
Bachelors	0.304 (0.460)	0.294 (0.456)
F. Age of Individual		
Age	42.71 (15.10)	42.84 (14.84)
G. Marital Status		
Married	0.613 (0.487)	0.626 (0.484)
Divorced	0.120 (0.325)	0.123 (0.328)
Widowed	0.040 (0.197)	0.043 (0.202)
H. Race		
White non-Hispanic	0.719 (0.449)	0.753 (0.431)
Black non-Hispanic	0.099 (0.298)	0.097 (0.296)
Hispanic	0.128 (0.334)	0.106 (0.308)
Other	0.055 (0.227)	0.042 (0.200)
I. Employment Status		
Employed	0.672 (0.469)	0.684 (0.465)
J. Smoking Status		
Smoker	0.271 (0.444)	0.251 (0.434)
K. Income		
Income1 (< $10000)	0.074 (0.262)	0.079 (0.270)
Income2	0.064 (0.245)	0.064 (0.244)
Income3	0.081 (0.274)	0.081 (0.273)
Income4	0.098 (0.297)	0.100 (0.300)
Income5	0.146 (0.353)	0.161 (0.367)
Income6	0.175 (0.380)	0.194 (0.395)
Income7	0.168 (0.374)	0.185 (0.388)
Income8	0.193 (0.395)	0.137 (0.344)
L. Instrumental Variable		
Post (Lose or maintain weight)	-	0.616 (0.486)
L. Other Health Outcomes		
Diabetes	0.245 (0.430)	0.215 (0.411)
Observations	3,450,273	595,201

[a] All means are weighted.

Table 2. Regression Results | Dependent variable: Ln(Body mass index)

Explanatory Variables	(OLS) (1)	(FE-OLS) (2)	(OLS) (3)	(FE-OLS) (4)
Physical Activity	-0.050***	-0.050***	-	-
	(0.000)	(0.000)	-	-
Irregular	-	-	-0.029***	-0.036***
	-	-	(0.001)	(0.001)
Regular	-	-	-0.053***	-0.055***
	-	-	(0.001)	(0.001)
Active	-	-	-0.078***	-0.087***
	-	-	(0.001)	(0.001)
Age	0.002***	0.002***	0.002***	0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Female	-0.004***	-0.004***	-0.014***	-0.015***
	(0.000)	(0.000)	(0.001)	(0.001)
Low Education	0.057***	0.043***	0.031***	0.028***
	(0.001)	(0.001)	(0.002)	(0.002)
Med Education	0.070***	0.051***	0.006***	0.004*
	(0.001)	(0.001)	(0.002)	(0.002)
High Education	0.027***	0.011***	-0.025***	-0.025***
	(0.001)	(0.001)	(0.002)	(0.002)
Smoker	-0.046***	-0.045***	-0.022***	-0.024***
	(0.000)	(0.000)	(0.001)	(0.001)
Employed	0.014***	0.016***	0.011***	0.013***
	(0.000)	(0.000)	(0.001)	(0.001)
Married	0.027***	0.029***	0.025***	0.029***
	(0.000)	(0.000)	(0.001)	(0.001)
Divorced	0.014***	0.014***	0.015***	0.016***
	(0.001)	(0.001)	(0.001)	(0.001)
Widowed	-0.014	0.000	0.004*	0.005**
	(0.001)	(0.001)	(0.002)	(0.002)
Black non-Hispanic	0.066***	0.063***	0.069***	0.069***
	(0.001)	(0.001)	(0.001)	(0.001)
Hispanic	0.069***	0.077***	0.039***	0.044***
	(0.000)	(0.001)	(0.001)	(0.001)
Other	-0.020***	-0.020***	-0.032***	-0.033***
	(0.001)	(0.001)	(0.002)	(0.002)
Income2	-0.015***	-0.019***	-0.013***	-0.019***
	(0.001)	(0.001)	(0.002)	(0.002)
Income3	-0.016***	-0.023***	-0.022***	-0.031***
	(0.001)	(0.001)	(0.002)	(0.002)
Income4	-0.023***	-0.031***	-0.029***	-0.040***
	(0.001)	(0.001)	(0.002)	(0.002)
Income5	-0.032***	-0.040***	-0.034***	-0.045***
	(0.001)	(0.001)	(0.002)	(0.002)
Income6	-0.030***	-0.039***	-0.036***	-0.049***
	(0.001)	(0.001)	(0.002)	(0.002)

Table 2--Continued

Explanatory Variables	(OLS) (1)	(FE-OLS) (2)	(OLS) (3)	(FE-OLS) (4)
Income7	-0.026***	-0.037***	-0.041***	-0.057***
	(0.001)	(0.001)	(0.002)	(0.002)
Income8	-0.035***	-0.056***	-0.058***	-0.086***
	(0.001)	(0.001)	(0.002)	(0.002)
State Fixed-Effects	No	Yes	No	Yes
Time Fixed-Effects	No	Yes	No	Yes
F-statistic	7734	1804	1050	304
R^2	0.04	0.05	0.04	0.05
Observations	3,450,273	3,450,273	595,201	595,201

[a] BMI is the dependent variable in log form.

[b] *Significant at 10%, **Significant at 5%, ***Significant at 1% (all two-tailed test).

Table 3. Regression Results | Dependent variable: Binary Diabetes Variable

Explanatory Variables	(FE-OLS) (1)	(FE-OLS) (4)
Physical Activity	-0.013***	-
	(0.000)	-
Irregular	-	-0.010***
	-	(0.001)
Regular	-	-0.011***
	-	(0.001)
Active	-	-0.023***
	-	(0.001)
Age	0.001***	0.002***
	(0.000)	(0.000)
Female	0.010***	0.000
	(0.000)	(0.001)
Low Education	-0.194***	0.004**
	(0.001)	(0.002)
Med Education	-0.431***	-0.012***
	(0.001)	(0.002)
High Education	-0.356***	-0.019***
	(0.001)	(0.002)
Smoker	-0.039***	-0.008***
	(0.000)	(0.001)
Employed	-0.010***	-0.026***
	(0.000)	(0.001)
Married	0.033***	0.008***
	(0.000)	(0.001)
Divorced	0.006***	0.004***
	(0.001)	(0.001)
Widowed	0.036***	0.011***
	(0.001)	(0.002)
Black non-Hispanic	0.012***	0.032***
	(0.001)	(0.001)
Hispanic	-0.085***	0.013***
	(0.001)	(0.001)
Other	-0.040***	0.016***
	(0.001)	(0.002)
Income2	-0.002**	-0.005***
	(0.001)	(0.002)
Income3	-0.026***	-0.013***
	(0.001)	(0.002)
Income4	-0.034***	-0.015***
	(0.001)	(0.001)
Income5	-0.029***	-0.024***
	(0.001)	(0.001)
Income6	-0.053***	-0.029***
	(0.001)	(0.001)

Table 3--Continued

Explanatory Variables	(FE-OLS) (1)	(FE-OLS) (4)
Income7	-0.111***	-0.035***
	(0.001)	(0.001)
Income8	-0.193***	-0.050***
	(0.001)	(0.002)
State Fixed-Effects	Yes	Yes
Time Fixed-Effects	Yes	Yes
F-statistic	35914	13224
R^2	0.51	0.69
Observations	3,450,273	595,201

[a] Diabetes is the dependent binary variable, where it is equal to 1 if an individual has diabetes and 0 otherwise.
[b] *Significant at 10%, **Significant at 5%, ***Significant at 1% (all two-tailed test).