Smoking Status, BMI, Life Expectancy, and Age in the Sanford Data Collaborative

JANUARY 2019

T.J. Liggett

Prepared for COSC 219: Big Data in Medicine at Augustana University with Dr. Olson-Manning

Introduction

Smoking and obesity have been some of the most highlighted health concerns of the American people over the past several decades. In 1997, an article in the Journal of the American Medical Association identified smoking and sedentary lifestyle as the first and second most important causes of death in the American population (Ferucci et al. 1999). Smoking and obesity are dangerous health risks that are prominent in America. Research has shown that both smoking and obesity have an adverse effect on health and life expectancy (Stewart et al. 2009). Multiple studies have observed the effects of both obesity (Bigaard et al. 2012; Ferucci et al. 1999; Van Baal et al. 2006) and smoking (Streppel et al. 2007; Sakata et al. 2012; Van Baal et al. 2006) on life expectancy. These studies show negative correlations between life expectancy and smoking/obesity.

While Bigaard et al. (2012) included waist circumference as an added measure of obesity, BMI is generally used as a standard measure of obesity. However, some research has shown a negative correlation between BMI and smoking (Albanes et al. 1987). This correlation between BMI and smoking may have an impact when attempting to find a correlation between BMI and life expectancy when failing to segregate smokers from non-smokers. It also remains in question whether these two variables together can provide any prediction of life expectancy.

Despite all this research into the correlations of life expectancy, smoking, and BMI, we are not aware of any research attempting to predict individual life expectancy utilizing both variables. Sources that studied both variables sorted smokers into groups, rather than measuring smoking status as a spectrum. While Stewart et al. attempted to forecast the effects of obesity and smoking on U.S. life expectancy (2009), they focused on the overall trends of these variables going into the future rather than individual predictions. This study will attempt to find and quantify correlations between smoking status, BMI, and life expectancy, and use these correlations to predict the impact of BMI and smoking status on life expectancy.

Hypothesis

There is no correlation between smoking status, BMI, and life expectancy. BMI and smoking status cannot be used to predict life expectancy.

Methods

Statistical analyses were done utilizing the Sanford Data Collaborative Teaching Dataset (SDC), provided by Sanford Research. The SDC provided healthcare data for 155,143 Sanford patients. Of these patients, 151,480 were alive and 3,663 were deceased. While providing a variety of health factors, the SDC gave limited demographic information about patients to protect their privacy.

Analyses were made using four variables from the dataset (Age, Smoking Status, BMI, and Life Status). An assumption was made that the age of deceased individuals was their age at death, allowing for predictions about life expectancy. Smoking Status was a spectrum ranking from 1 to 10, with higher scores indicating higher smoking levels. Using these four given variables, two additional variables were created: Year of Birth and Average Smoking Status (per YOB). Year of Birth was generated by subtracting the age of living patients from the year 2018. As it is unclear when the dataset was submitted to Augustana and the birth dates of patients are unknown, this metric could be off by as much as a year. The mean smoking status of all patients born in each year was found to create the Average Smoking Status of a given birth year. Correlations between variables were found using Pearson's product-moment correlation.

Figure 1.0: Pearson's Product-Moment Correlation of Variables in the Sanford Data Collaborative

Population Status	X Variable	Y Variable	Correlation Coefficient	Confidence Interval(95%)	Degrees Freedom	P-Value
Deceased	Age	Smoking Status	0.222	(0.185, 0.257)	2668	<0.001
Deceased	Age	BMI	-0.018	(-0.060,0.025)	2144	0.411
Deceased	BMI	Smoking Status	-0.003	(-0.040,0.035)	2734	0.887
Alive and Deceased	BMI	Smoking Status	0.007	(0.001, 0.012)	144900	0.012
Alive	Age	Smoking Status	0.112	(0.114, 0.124)	145770	<0.001
Alive	Year of Birth	Average Smoking Status	-0.373	(-0.557, -0.155)	70	0.001

Figure 2.0: Age at Death and Smoking Status of Dead Individuals in the SDC

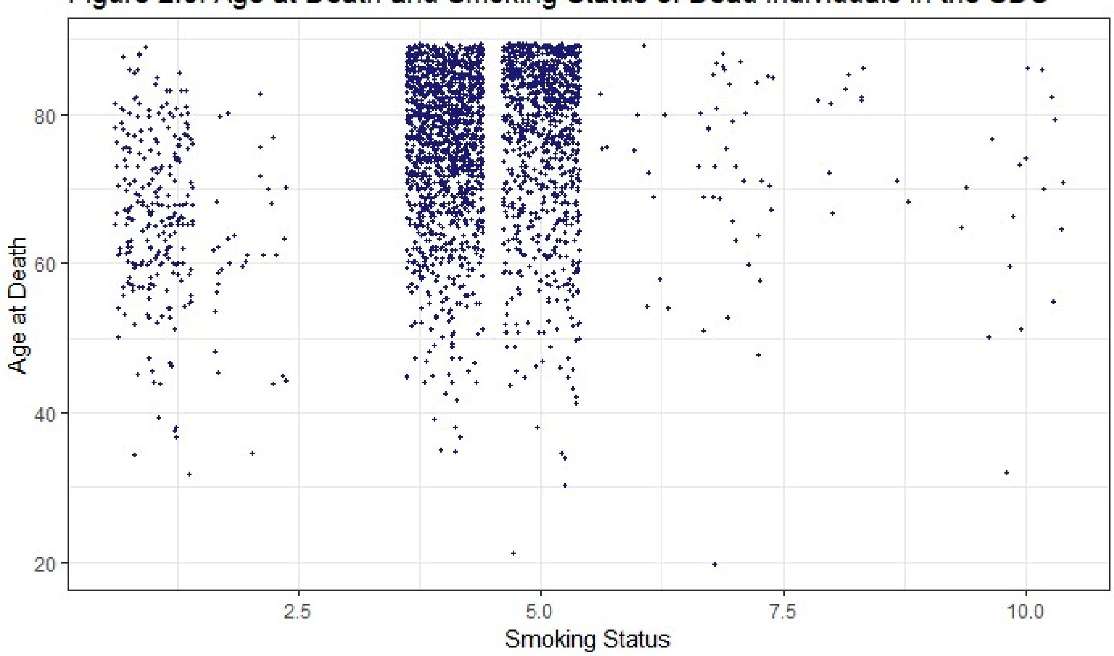


Figure 3.0: Frequency of Age of Individuals in SDC

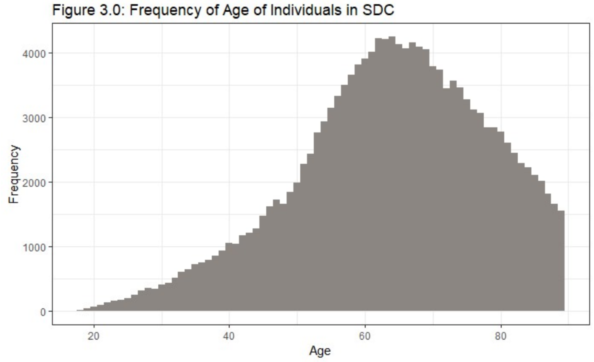


Figure 3.1: Frequency of Age of Deceased Individuals in SDC

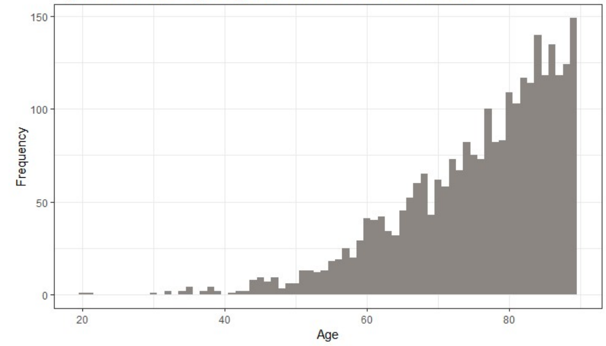


Figure 3.2: Frequency of Age of Living Individuals in SDC

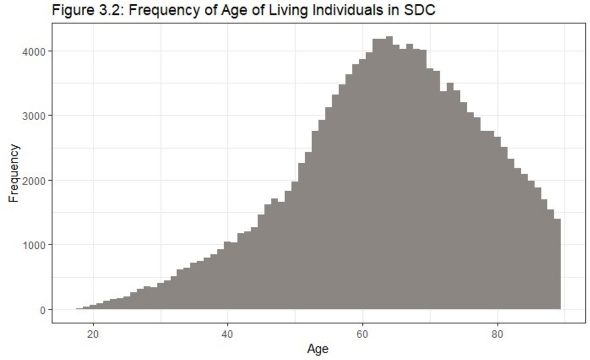


Figure 3.3: Frequency of Year of Birth of Living Individuals in SDC

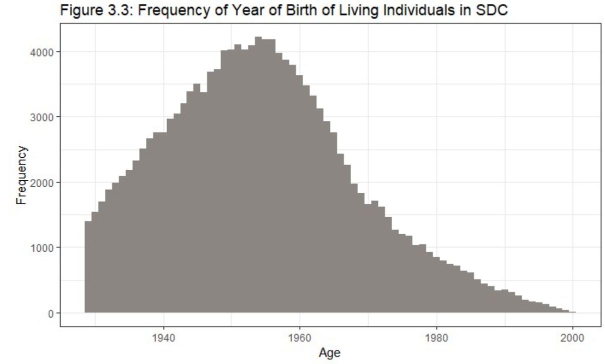
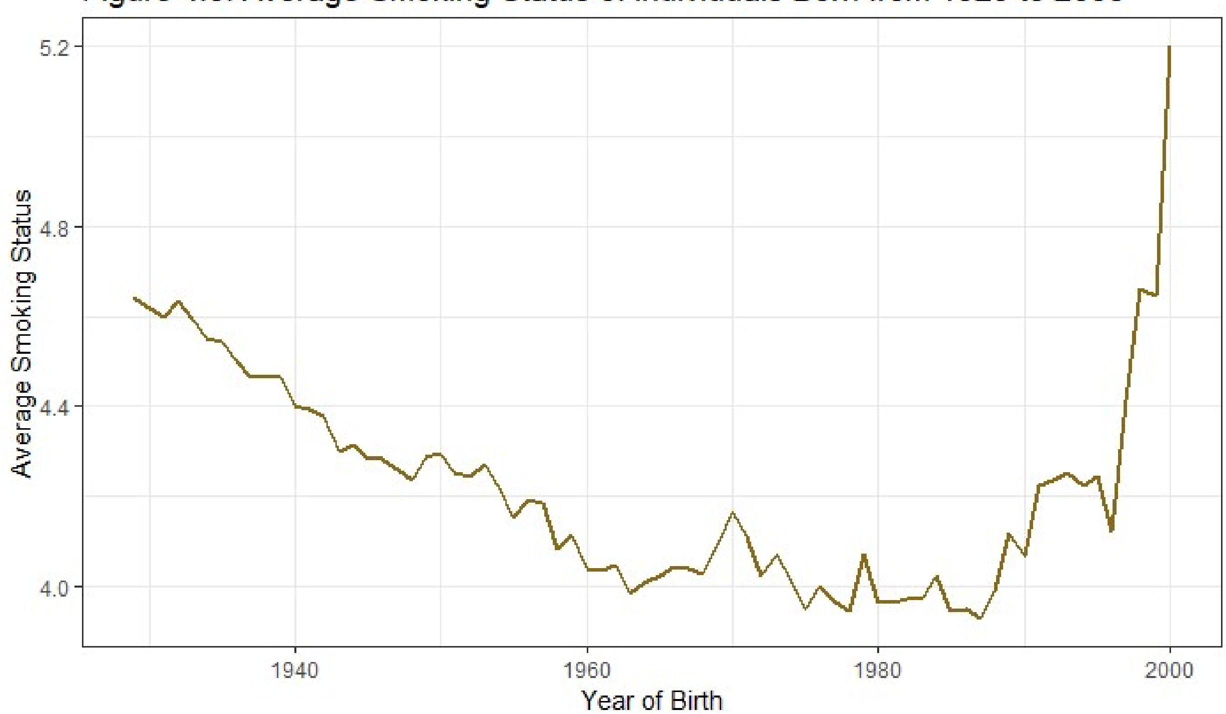


Figure 4.0: Average Smoking Status of Individuals Born from 1929 to 2000



Results

There was no correlation found between BMI and either age of death or smoking status among deceased individuals. A miniscule correlation was found between BMI and smoking status among the general population, but this correlation was so small across so many data points that no further research was done into the matter, and BMI was not considered to be at risk of being a confounding variable. A correlation of 0.222 was found between smoking status and age of death (see Figure 1.0). This positive correlation was concerning as a majority of research suggested a negative correlation between these variables (Streppel et al. 2007; Sakata et al. 2012; Van Baal et al. 2006).

To check what was causing these surprising results, a correlation test was performed between age and smoking status among living members of the population. A correlation of 0.112 among living patients revealed a correlation not only between age of death and smoking status but also smoking status and age. Distribution frequencies of the age variable among the population revealed a strong negative skew in both alive and deceased members of the population (see Figures 3.0-3.2). This skew was especially present among dead individuals in the population, possibly explaining the larger correlation. To compare historical smoking rates, an average smoking status was computed for each year of birth among living patients. Average smoking status saw a steady decline in individuals born from 1929 to about 1985 (see Figure 4.0), before seeing increases which could be caused by a lack of data. Average smoking status and year born had a negative correlation of -0.373 among living individuals of the SDC.

Discussion

The lack of correlation of BMI with other factors may suggest that this metric is inaccurate, as suggested by Bigaard et al. in 2012. BMI does not accurately portray a person's obesity or health, and other measurements might form a more complete picture of a person's obesity status. These other measurements would be essential in further research into a correlation between obesity and life expectancy.

The generational decline in smoking observed in Figure 4.0 aligns with reports of declining smoking rates (Ny et al., 2014). This decline might explain the odd correlation between smoking status and life expectancy. Older individuals grew up with less awareness of the negative effects of smoking, and their views on the subject might differ from younger individuals. Thus, older patients might be more likely to have smoked for much of their lives due to a lack of educational awareness about the dangers of smoking.

Further research must account for generational differences in smoking rates as well as the negative skew in the SDC. To account for these effects, a dataset must be used with more than 3,663 deceased individuals, preferably one with a more even distribution of age of death. A perfect distribution is likely unattainable, as humans can die 60 years before the median age of death, but will not live 60 years past. However, the information of more individuals who died earlier than expected would be helpful.

References

Albanes, D., Jones, D.Y., Micozzi, M.S., Mattson, M.E. (1987). Associations between smoking and body weight in the US population: analysis of NHANES II. American Journal of Public Health, 77(4), 439-444.

Bigaard, J., Tjonneland, A., Thomsen, B.L., Overvad, K., Heitmann, B.L., & Sorensen, T.I.A. (2012). Waist Circumference, BMI, Smoking, and Mortality in Middle-Aged Men and Women. Wiley Obesity Research, 11(7).

Ferrucci, L., Izmirlian, G., Leveille, S., Phillips, C.L., Corti, M., Brock, D.B., Guralnik, J.M. (1999). Smoking, Physical Activity, and Active Life Expectancy. American Journal of Epidemiology, 149(7), 645-653.

Ng, M., Freeman, M.K., Fleming, T.D., et al. (2014). Smoking Prevalence and Cigarette Consumption in 187 countries, 1980-2012. Journal of the American Medical Association, 311(2), 183-192.

Sakata, R., McGale, P., Grant, E.J., Ozasa, K., Peto, R., & Darby, S.C. (2012). Impact of smoking on morality and life expectancy in Japanese smokers: a prospective cohort study. BMJ.

Stewart, S.T., Cutler, D.M., Allison, B.R. (2009). Forecasting the Effects of Obesity and Smoking on U.S. Life Expectancy. The New England Journal of Medicine, 361, 2252-2260.

Streppel, M.T., Boshuizen, H.C., Ocke, M.C., Kok, F.J., Kromhout, D. (2007). Mortality and life expectancy in relation to long-term cigarette, cigar and pipe smoking: The Zutphen Study. Tobacco Control, 16(2).

Van Baal, P.H.M., Hoogenveen, R.T., De Wit, A.G., Boshuizen, H.C. (2006). Estimating health-adjusted life expectancy conditional on risk factors: results for smoking and obesity. Population Health Metrics, 4(14).