Pre-Data Science Capstone: Diabetes Prevalence in the United States and Puerto Rico

Naomi Hernandez
March 4, 2019

Introduction

The United States is currently experiencing a wide-spread diabetes epidemic. The Centers for Disease Control (CDC) at the United States Department of Health and Human Services estimate that as many as one in three Americans is on their way to developing diabetes.[1] A metabolic illness with no cure and serious health risks, diabetes puts people's lives and well-being at risk, and puts strain on states' employers and health systems. To that end, I analyzed a national dataset to understand the national disease burden.

The diabetes prevalence data used in this analysis comes from the Behavioral Risk Factor Surveillance System (BRFSS), a national telephone survey. Established by the CDC in 1985, BRFSS is administered annually to collect data on chronic disease and behaviors that impact health and chronic disease outcomes. Topics include alcohol, diabetes, chronic obstructive pulmonary disease, nutrition and physical activity, cancer, and more. I used the dataset included in a larger multi-source dataset called the United States Chronic Disease Indicators (CDI) dataset, available on data.gov: https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi. This most recent dataset included BRFSS data from 2011 to 2016. Records include data from all 50 states, Washington DC, and three United States territories—Puerto Rico, Guam, and the Virgin Islands.

The mortality data presented in this analysis comes from the National Vital Statistics System (NVSS), and was also included in the CDC Chronic Disease Indicators dataset. NVSS is housed in the National Center for Health Statistics in the CDC, and disseminates mortality and birth data to the United States government and public.[2] The CDI dataset includes data from all 50 states and Washington DC.

Methods and Data Preparation

The Virgin Islands have been excluded entirely from the analysis due to insufficient data. There was also a record in the original dataset representing the United States averages that has also been excluded; it represents duplicate information having been calculated from the state averages. To see my code for data wrangling, calculations, and plots, please see my notebook: https://github.com/nrhernan/Thinkful/blob/master/Thinkful_PreDS_Capstone_WranglingVisualizations.ipynb.

[1] "About Diabetes." Centers for Disease Control and Prevention. June 1, 2017. Accessed March 02, 2019. https://www.cdc.gov/diabetes/basics/diabetes.html.

[2] "About the National Vital Statistics System." Centers for Disease Control and Prevention. January 04, 2016. Accessed March 02, 2019. https://www.cdc.gov/nchs/nvss/about_nvss.htm.

Question 1: What is the current prevalence of diabetes in the United States?

Before we can take meaningful action against the diabetes epidemic, we first need to understand what the current prevalence is across the United States, as well as recent prevalence trends. The following charts illustrate age-adjusted state/ territory level prevalence 2011 and 2016:

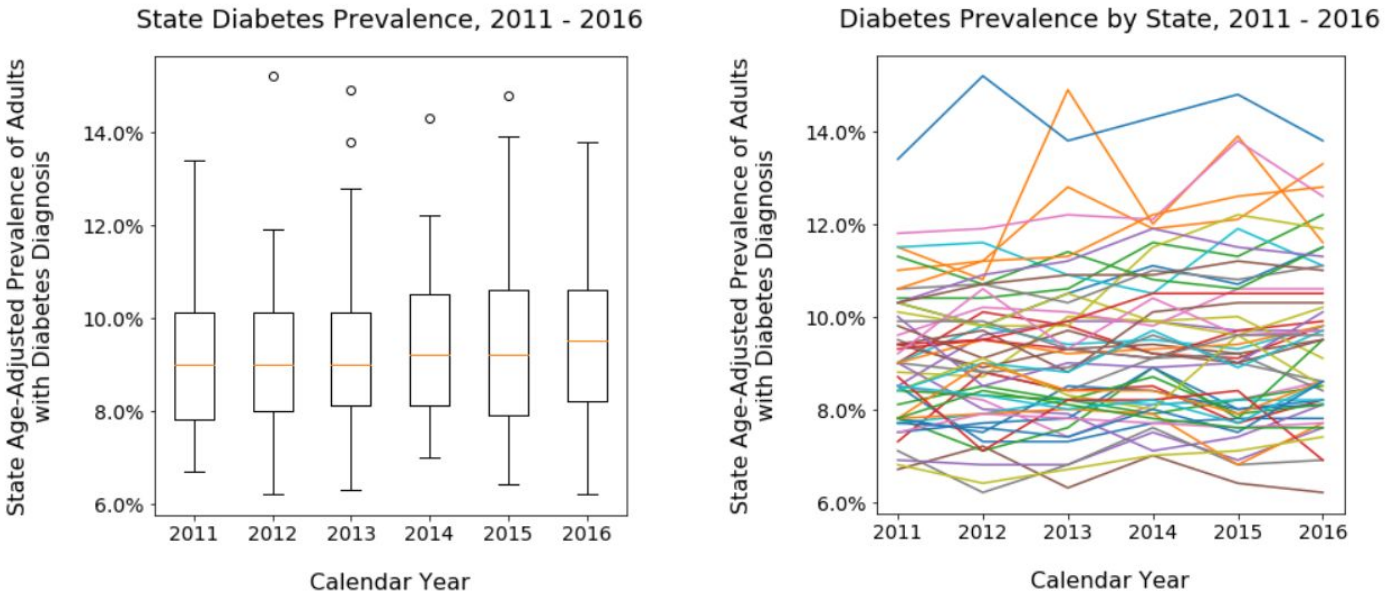


Figure 1: Age-adjusted diabetes prevalence by state/ territory

The box plot illustrates a small increase in the median—from 9.0% in 2011 to 9.5% in 2016—as well as the lower limit of the second quartile and the ceiling of the third quartile. Taken together, these indicate a genuine movement of the central tendency. The line graph, however, displays a much more ambiguous picture, with the prevalence of diabetes in some states appearing to decrease. To test the significance of this change I conducted a paired t-test, comparing the 2011 state prevalences to their 2016 counterparts. This resulted in a p-value of 0.0002—far below my pre-selected alpha of 0.05. We can therefore conclude with reasonable certainty that state prevalences in general increased during the study period.

Question 2: What is driving the increased prevalence in diabetes?

Given the epidemiology of diabetes, the increase in diabetes prevalence between 2011 and 2016 is likely to be caused by one or more of the following:

- Increased diabetes incidence. If the rate at which adults in the United States are being diagnosed with diabetes is increasing then we would expect the prevalence to also increase, unless diabetes-related mortalities also increased.
- A proportionally older population. Age is a risk factor for diabetes. While prevalences were age-adjusted within survey years for comparison between states, they were not adjusted across survey years. It would be possible for the prevalence of diabetes within age groups to remain constant and still see an increase in diabetes prevalence if those age groups increased relative to other age groups throughout the United States.
- Decreased diabetes mortality. Unless the population became proportionally younger or the incidence of diabetes decreased, an decrease in mortality among diabetics would also result in an increased diabetes prevalence.

Because there was no data on diabetes incidence or population demographics in the CDI dataset, I explored whether a change in mortality was a likely driver. The data comes from NVSS and differs slightly in its composition from the BRFSS prevalence data:

1. NVSS reports on mortality for 2010-2014 instead of 2011-2016; and
2. NVSS excludes two territories I included in the prevalence analysis: Guam and Puerto Rico.

However, the majority of data points in both datasets overlap. It is therefore reasonable to extend conclusions from NVSS to BRFSS data. State-level mortality rates for 2010-2014 are summarized in the figure below:

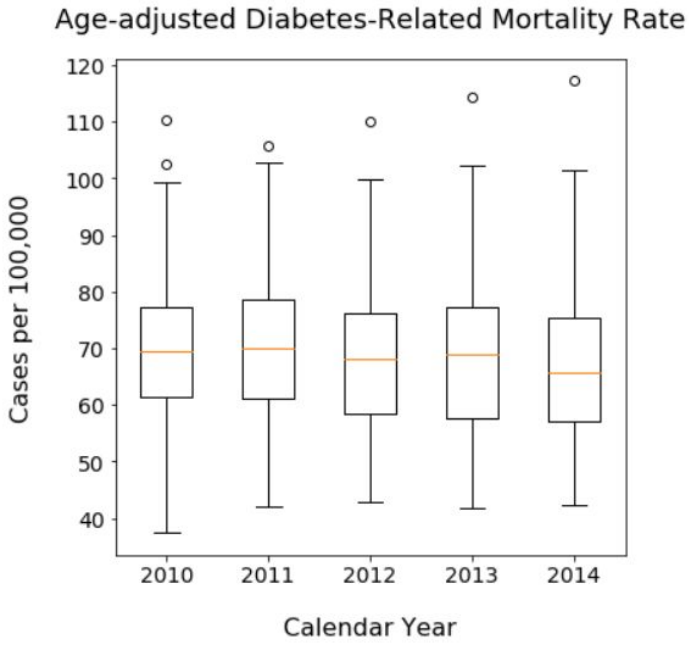
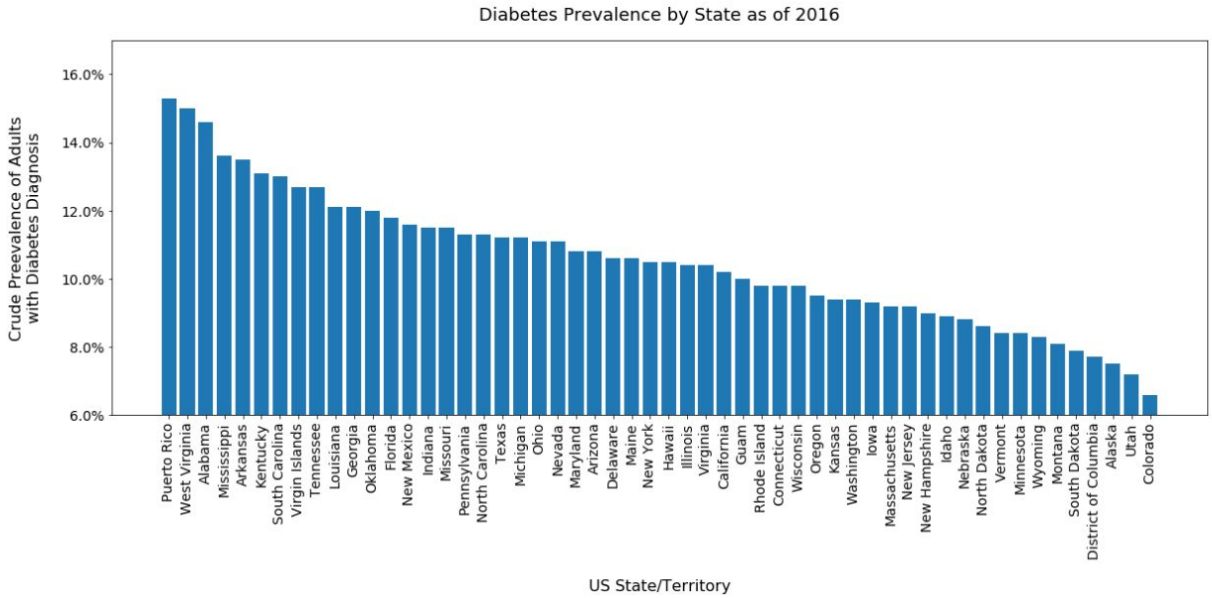


Figure 2: Age-adjusted diabetes-related mortality rate

While the median decreased from 69.3 deaths per 100,000 population in 2010 to 65.0 deaths per 100,000 population in 2014, the difference appears small. To check whether this difference was statistically significant I conducted a paired t-test. The resulting p-value was 0.003, well below my pre-selected alpha of 0.05, suggesting that the decrease in median diabetes-related mortality reflects a true national decrease.

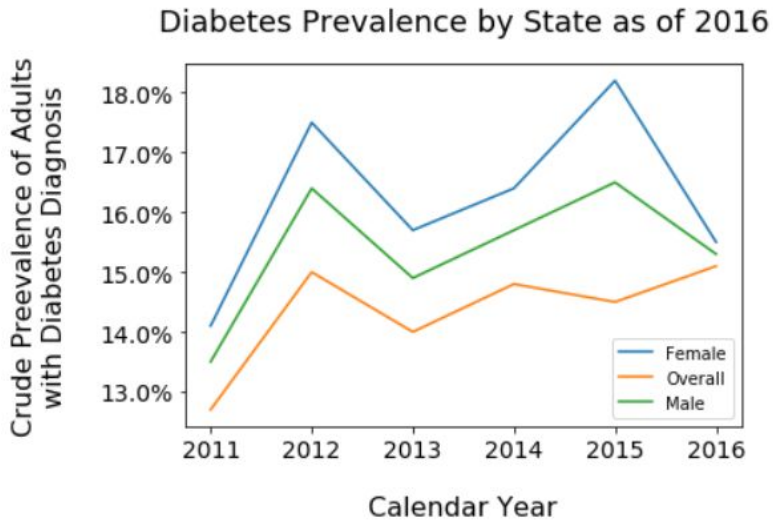
Question 3: Who bears the majority of the disease burden?

So far this analysis has focused on historical trends. The goal, however, is to improve the health and well-being of Americans it behooves us to focus our resources where the need is greatest. I therefore looked for the state/ territory with the highest prevalence of diabetes in 2016, illustrated below:



We can see that Puerto Rico had the highest overall diabetes prevalence in 2016. This time I used crude prevalence instead of age-adjusted prevalence because I was interested in finding the state/ territory with the highest absolute disease burden.

Since disease burden can fall disproportionately on some parts of a population, I further looked at the prevalence of diabetes by gender:



This graph illustrates that diabetes prevalence has risen among men and women since 2011, and that the prevalence has historically been higher among Puerto Rican women compared to Puerto Rican men. However, the 2016 data shows nearly identical prevalences, with 15.5% of adult women having diabetes compared to 15.1% of men. This suggests that public health efforts in Puerto Rico should equally engage men and women.

Opportunities for Further Research

This analysis only scratched the surface of the diabetes epidemic in the United States. More research is needed to understand why diabetes prevalence is generally increasing in Puerto Rico and elsewhere. I only explored mortality as a possible driver in increasing diabetes prevalence. Further research should explore trends in diabetes incidence, as well as whether the U.S. population has become proportionally older since 2011.

In addition, this analysis did not address any of the causes of diabetes nor where opportunities exist to prevent individuals from developing diabetes altogether. These opportunities could include general behaviors, such as amount of physical activity during recreation time, or specific interventions such as Michigan's implementation of the National Diabetes Prevention Program.[3] Finally, further research is needed to understand how best to support individuals living with diabetes. This dataset contained information on health management and outcomes for diabetics which could be correlated with prevalence trends. However since the impact of prevention behaviors manifests slowly over time, the author recommends that anyone trying to answer questions about the efficacy of these behaviors conduct longitudinal studies with diverse cohorts.

[3] "Diabetes Prevention in Michigan." Keeping Michigan Healthy. Accessed March 03, 2019. https://www.michigan.gov/mdhhs/0,5885,7-339-71550_2955_2980_3168-136877--,00.html.