**Disease Prevention: A Data Mining Approach**

We hear it in the news every week that health care costs are rising faster than the general inflation rate. Health care reform has been the focus of intense effort in the U.S. Congress for months. Americans spend a lot of money for health care, but health outcomes are not as good as they should be. Some new types of infection are testing our ability to control them. Even diseases that normally strike adults are now affecting the younger population. In short, health care is in the spotlight.

The discussion on health care reform has raised many questions. Should there be a one-payer system for health care? Should the industry be more regulated? Where does the responsibility lie for being healthy and providing health care services?

**Problem Statement**

With all of the talk on health care reform, many want to know the importance of prevention and what cost savings might result from taking preventive measures to improve the population’s health. The disease of diabetes is of particular concern due to its chronic nature. Treatment is expensive and ongoing. It can take a toll on a person’s life expectancy and quality of life. Although there are genetic elements to many diseases and, no doubt, to diabetes also, there is evidence that behavioral factors are important (http://www.technologyreview.com/blog/editors/24700/).

The United States Health Expense Think Tank (USHETH) is a fictional group that wants to know the impact of preventive measures on disease occurrence. Furthermore, it would like to know how an increase or decrease in disease prevalence would affect the amount of money spent on health care. A large study was done in the United States to collect information on individuals, their health picture and how much money is spent on their behalf for health care. The sample is representative of the population and represents a snapshot of the country and its health care costs at a point in time.

USHETH would like to introduce a number of preventive measures into the health care environment in the country but wants to pre-test these ideas mathematically before incurring marketing and other expenses associated with advertising campaigns and physician education. If the health and financial impact can be shown for these measures, it will provide valuable information and incentive for implementation.

USHETH has a number of ideas that it wants to test but wants to analyze them one at a time in order of potential impact. Following a recent meeting to set its research agenda, the Chief Research Officer (CRO) announced the decision to focus on diabetes prevention. Using the dataset USHETH wants to research three points:

1. USHETH would like to know how having diabetes influences a patient’s total health care expense. There are many estimates of the cost of diabetes care that have focused on diabetes-specific expenses, but it is known that diabetes might affect total health care expenses in many ways. There are many factors in the patient database that need to be taken into account so USHETH is looking for a model to know the difference in health care costs for a person with diabetes from the costs for a person without the disease. It acknowledges that there are many factors that need to be taken into account.
2. USHETH needs to know how many people have diabetes by age and perhaps gender. This will be critical for the analysis.
3. USHETH wants to measure the impact of its preventive measure. Although it sounds simple, the organization believes that it will have a profound impact on people’s health and total health care expenses. The measure focuses on people who have a BMI (body mass index) larger than 25 in adults. In children the threshold BMI varies by age linearly: at 5 years old, the threshold BMI is 17 and at 20 years old, the threshold BMI is the same for adults. These people will be enrolled in special programs to reduce their BMI by ten percent. USHETH then wants to know how many people would contract diabetes and the resulting reduction in total health care expenses.

A population distribution chart has been provided so that your research group will know the age and gender profile of the country.

Your team will be provided with a dataset and a codebook explaining each variable and the values they can take on. The dataset is a sample of people from across the country. Your team will need to choose the predictor variables, transform them as necessary and build a model for the target variable. From the health data that you have been provided, build a model to predict incidence of diabetes.

A significant portion of health care expenses are paid by the government programs Medicare and Medicaid. In addition to the reduction in total health care expenses, USHETH would like to know what portion of those savings will go to Medicare and Medicaid.

Prepare your analysis and present it in a report. The report should include a brief executive summary and the results of your analysis. Please see the submission instructions for other items that you must submit.