**Big Data Analysis to Improve Lung Cancer Patient Outcomes**

**Shilpa Balan, Vrunda Shah, Nishant Shristiraj**

**California State University-Los Angeles**

**Abstract**

Analysis of cancer survival data and related outcomes is essential to assess cancer treatment options. Lung cancer is one of the leading causes of death in the United States. Several individuals are diagnosed with lung cancer every year in the United States. The associated healthcare costs to lung cancer are escalating. Historically, most cancers are diagnosed at late stages, when the chance for cure is lower. From previous research it is seen that the improvement of cancer treatment technologies has emphasized on the focus of the patient’s characteristics and disease features. Thus, the objective of this study is to describe the similarity and pattern of patient characteristics detected with lung cancer in the United States. For instance, some of the factors examined in this study are gender, age, trend of lung cancer incidences and mortality from years 1999 to 2014. The aim of this study is to improve lung cancer patient outcomes to enable better chance of survival to a patient.

Data has been increasing in every industry including the healthcare industry. It has been a challenge to manage this data. However, if managed correctly, data can be an asset. For this research, the lung cancer data is collected from CDC (Centers for Disease Control and Prevention). Additional data is collected from Twitter for the hashtags #lungcancer and #tobacco. From previous literature it is found that smoking is a major cause of lung cancer. Hence, we selected the hashtag #tobacco for data collection. Due to its openness, Twitter has generated massive amounts of data. Python scripts were written for data analysis.

The following are some of the observations made. From 2001 to 2003, there is an increase in the number of lung cancer incidences. The incidence rate is constant from 2003 onward and until 2010. There is a decline in the number of lung cancer incidences from 2010 to 2011. However, the incidences again steadily increase in 2012 and is constant until 2014. The number of youth male smokers is more than the number of youth female smokers. From the lung cancer incidences, the rate of mortality is almost the same for both male and female. These described here are some of our findings and analysis.