Lung cancer is the leading cause of cancer death and the second most diagnosed cancer in both men and women in the United States. After increasing for decades, lung cancer rates are decreasing nationally, as fewer people smoke cigarettes.

The American Cancer Society’s projected that in 2019 lung cancer in the United States are:

* Around 1,762,450 new cancer cases
* Around 606,880 cancer deaths to occur

Over the past decade of data, the cancer incidence rate (2006‐2015) was stable in women and declined by approximately 2% per year in men, whereas the cancer death rate (2007‐2016) declined annually by 1.4% and 1.8%, respectively.

**Source**

This Data Set is downloaded from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data)

**Download Link**: <https://archive.ics.uci.edu/ml/machine-learning-databases/00277/>

**Original File Name**: ThoraricSurgery.arff

### Abstract

The data is dedicated to the classification problem related to the post-operative life expectancy in the lung cancer patients: class 1 - death within one year after surgery, class 2 - survival.

**Data Set Information**

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007-2011. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

The sample data set includes 470 rows of patient’s data with 16 attributes/features as the columns. The death\_1yr attribute column can determine if a given patient lived or died. A "false" value indicates that the patient lived 1 year post surgery, while a "true" value indicates the patient died within 1 year of the surgery.

1. Problem Description:
   1. Determine post-operative life expectancy rate in lung cancer patients from patient attributes in the data set.
   2. Is there a correlation on higher death rates in lung cancer patients with certain patient attribute values and attributes? If so, do they correlate to higher mortality rate?
   3. How can we increase life expectancy rate in some of the high risk patients?
   4. Any hospitals that performs thoracic surgery for lung cancer patients can benefit from this research. Using this research health care providers and medical professionals can save patients with higher likelihood of postoperative death.
2. Solution.
   1. First, perform exploratory data analysis and cleansing exercise to check missing values, rename columns and/or manipulate data.
   2. This task consist of exploring all attributes’ values in terms of what they represent, how they relate to each other, and the quantitative aggregate values.
   3. Once data is ready, perform experimentation design and data visualization to quickly assess data patterns and overall trends.
   4. Hypothesis testing with the null being the attributes having no relations to death rate will be performed.