# Introduction

Death has always been a major fear for many and is the natural course of life. In our society we have always associated death with hospitals as it shown in popular movies like Disney’s Up. With ongoing advancements in medical science and technology a person’s life expectancy has risen compared to centuries ago.

However, there are the unfortunate cases of deaths within a hospital stay and all these records have been recorded by the National Impatient Sample (NIS) and its data collected by the Healthcare Cost and Utilization Project (HCUP). Hospitals themselves should not be looked on as the causation of death when a patient is admitted, as the NIS has a wealth of information of each inbound patient with records of their drug use `ARPDRG\_Severity`, length of stay (`LOS`), `Race`, `Age`, `Hospital Region`, if the patient was transferred in (`Tran\_In`), and operating procedure to name a few. With this vast dataset statisticians and hospitals can use these variables of interests to predict death for another inbound patient, to gauge a metric of how likely the patient is to die during their hospital stay with their incoming conditions and their personal demographics of age, race, and gender. Predicting this particular outcome of death will help to improve the efficiency of hospitals and to better improve the quality of life for those admitted into the hospital, as nurses and doctors will be able to provide the proper attention and care for those who are likely to die from their current conditions in the prediction model. The (NIS data) [https://www.hcup-us.ahrq.gov/] can be found using the link and this project will be looking into a subsample of 200,000 of all hospital patients in 2012.

# Methods

The sampled NIS dataset is a real-world dataset that is not preprocessed and there are a few missing observations that would conflict with our prediction models such as random forests and boosting models. There are a total of 13,904 observations missing and some with our main outcome of interest `Death`, we delete these missing observations and are left with 186,096 observations for our prediction model. Out dataset has 157 variables recorded for each patient and running our prediction model for a patient’s death we use selected variables of what the hospitals would have records of at the time of admission. The primary interest of our prediction model is to predict the death of an admitted patient and some variables may be recorded during their stay at the hospital and those observations would not help our model to predict death for an entirely new patient with different records before reaching the hospital. Therefore, we selected variables such as their demographics like `age`, `gender`, `race`, and what month they were admitted `amonth`, if it was the weekend or not `aweekend`, their drug severity `APRDRG\_Severity`, their current health conditions if they have `CM\_AIDS`, are alcholics `CM\_Alcohol`, have churned lung `CM\_CHRNLUNG`, are obese `CM\_OBESE`, if they elected to go to the hospital themselves `ELECTIVE`, if they had a operation procedure `ORPROC`, their method paying their hospital bill `PAY1`, and if they were transferred in from a different medical facility `TRAN\_IN` to name a short few.

I have used three different types of predictive models each with progressive performance. The first I have used was elementary logistic regression learned from basic statistics classes with a categorical outcome of whether the patient died or not and fitting all our parameters of interests with our main outcome of death to find the log likelihood of death occurring. We fit all the parameters of interest because these are readily available for each inbound patient and will give our model predictive accuracy. Then using machine learning techniques such as random forests to decorrelate trees that have positive correlated samples, to reduce variance. Since our prediction model is a classification problem of those that died and did not we consider a random subset of $\sqrt(46)$ of all our possible variables of interest.

Our final prediction model is boosting