Introduction

We explore the *colon* dataset found in the R "survival" package, containing data observed from a clinical trial where stage B/C colon cancer patients receive adjuvant chemotherapy. 929 independent patients–484 men and 445 women–were randomly assigned between two treatments and a control group: Levamisole, Levamisole and Fluorouracil, and control(denoted as *Lev*, *Lev+5FU*, and *Obs* in the dataset). Levamisole is a low-toxicity compound that was originally used to treat worm infestations in animals, while 5-FU is a moderately toxic chemotherapy agent used to treat cancer.

Patients were then observed until one of two events occurred: recurrence or death(denoted as "1" and "2" in its respective order under column *etype*). The time of occurrence, in days, was then recorded to later investigate and determine whether or not different treatments were effective in keeping the patients alive. Each patient in the dataset, identified by their *id*, has two rows for both recurrence and death. The *status* column indicates whether or not the event occurred or not("0" indicates no and "1" indicates yes). If a patient has been recorded for 3000 days for both recurrence and death and the status remains 0 for both, it signifies that they did not experience any event for 3000 days and dropped out of the study for unknown reasons. Figure 1 and 2 below represents the Kaplan-Meier Survival Curve after splitting the dataset by *etype*(recurrence and death). The convex shape of Figure 1 conveys that many recurrences occur early on while the curve for death events show that deaths in patients are gradual and consistent.

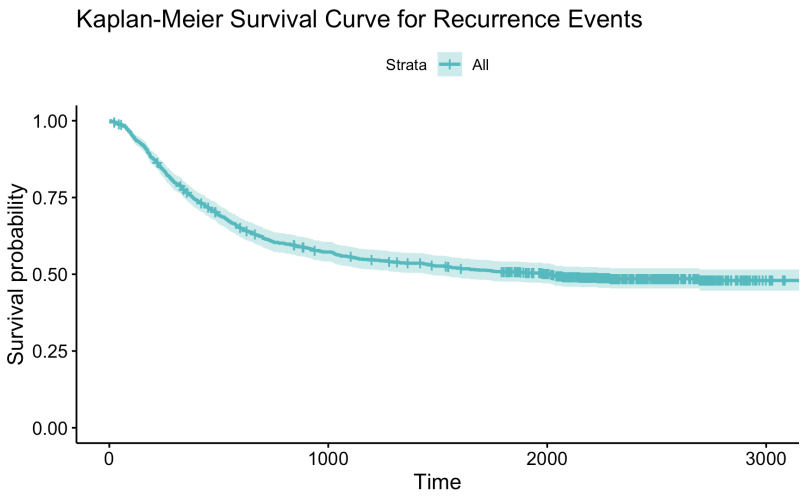

Figure 1: KM for Recurrence Event

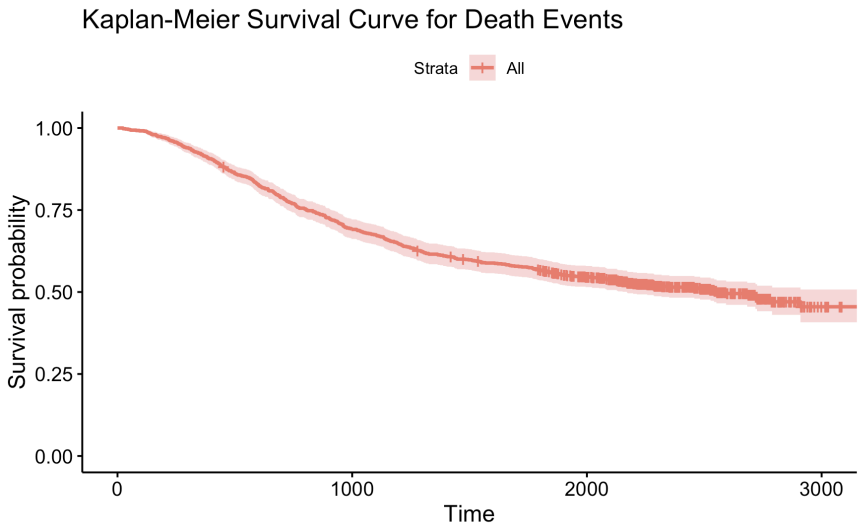

Figure 2: KM for Death Event

The dataset contains the *id, age*(in years), *sex*("1" indicates man and "0" indicates woman), *rx*(the treatment type or control), *obstruct*("1" indicates a colon obstructed by a tumor and "0" indicates no obstruction), *perfor*("1" indicates a perforated colon and "0" indicates no perforation), *adhere*("1" indicates cancer adhering to other organs and "0" indicates no adherence), *nodes*(the number of lymph nodes with colon cancer), *time*(time until event occurrence or censoring), *status*(whether or not the event occurred or not), *differ*("3" indicates quickly growing cancer, "2" indicates moderate growth, and "1" indicates slowly growing and less likely to spread), *extent*(describes the spread of the tumor and ranges from 1-4, where "1" indicates that the tumor is limited to the inner lining of the colon and "4" indicates invasion of tumor to nearby organs and tissues), *surg*("1" indicates a long time between initial surgery and registering to the study while "0" indicates a short time), *node4*("1" indicates a patient has more than four positive lymph nodes and "0" indicates four or less), and *etype*(recurrence or death event) of each patient.

Taking all of the covariates we listed above into consideration, our aim is to determine whether or not a specific treatment has a significant effect on the survival of the patient. Our secondary objective is to assess which covariate(s) have a significant effect on the hazard risk. In the course of the analysis, we omit observations with N/A values, reducing our final dataset to 888 independent patients. A five percent significance level(0.05) will be used to balance the risk of false positives and ensure adequate sensitivity to detect meaningful effects.