# Stat538 Midterm - Chronic Granulomatous Disease

**Arthur Lui**

**7 March 2014**

---

## 1 Introduction:

Chronic Granulomatous Disease (CGD), an inherited immunodeficiency disease that increases the body's susceptibility to infections caused by certain bacteria and fungi. CGD weakens an individual's immune system such that they are more susceptible to mild pathogens. CGD affects 1 in 200,000 people in the United States, with about 20 new cases diagnosed each year. The evaluation of the efficacy of a certain treatment (rIFN-g) in comparison to a placebo is of primary interest. A dataset (cgd.csv) provided by Dr. Engler consisting the variables including treatment, time to infection or end of study, and infection status is used for analysis to determine whether the treatment (rIFN-g) is effective in improving survival.

## 2 Analysis:

### 2.1 Logrank Test

An appropriate test to determine whether the treatment (rIFN-g) is effective in improving survival is the logrank test. The logrank test can be used to determine if one survival curve is significantly different from the other. To perform the test, the assumption that the two survival curves don't cross must be met. Therefore, prior to performing the logrank test, we plot the Kaplan-Meier (KM) curves for both treatment groups. That is, we plot the KM curves for both the placebo group and the rIFN-g group (see **Figure 1**).

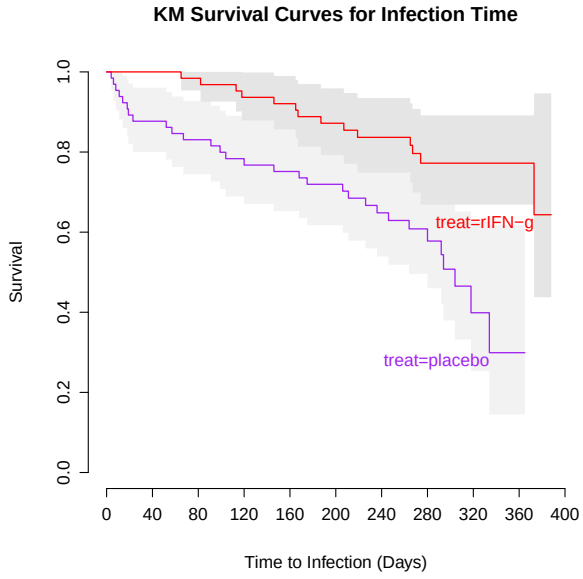



Figure 1: Kaplan-Meier Survival Curve for the Time until Infection with CGD with 95% confidence interval bands shaded in grey.

The two curves do not cross. Though, their 95% confidence interval (shaded in grey) regions do. Since the curves do not cross, we can perform the logrank test. Before proceeding with the logrank test, note that the median survival time for the rIFN-g treatment was computed to be **304** (95% confidence interval = (**248,360**)). The median survival time for the placebo could not be computed as the survival curve did not extend to 0.5.

The logrank test was performed and the results indicate that the survival of the treatment rIFN-g group was significantly (p-value = 0.00061) higher than that of the placebo group. That is, an individual from the treatment rIFN-g group is likely to catch an infection at a later time than an individual from the placebo group. The rIFN-g is significantly better than the placebo in prolonging the time to infection.

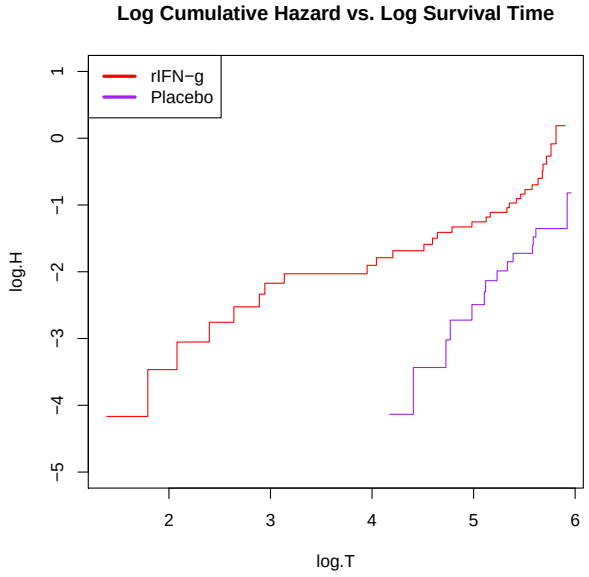
## 2.2 Cox Regression Model

The Cox regression model for proportional hazards provides a way to quantify how strongly the effects of the rIFN-g treatment correlate with the hazard function. A cox model was fit to the data. The log hazard ratio was regressed against treatment. **Table 1** summarizes the results of the regression. The estimate of $\hat{e^{\beta}} = 0.334866672397844$, where $\beta$ is the partial slope in the regression model. Therefore, we can say that the hazard rate for the rIFN-g treatment group is $\hat{e^{\beta}}$ (0.334866672397844) that of the placebo group. That is, the rIFN-g treatment group has a lower hazard function than the placebo group.

Table 1: Summary Table For Cox Model

| | exp(coef) | exp(-coef) | lower .95 | upper .95 |
|---|---|---|---|---|
| treatrIFN-g | 0.33 | 2.99 | 0.17 | 0.65 |

The log cumulative hazard function against log survival times for the two treatment groups should be parallel under the proportional hazards assumption. **Figure 2** shows that the two curves are reasonably parallel.



# 3 Conclusions

The analysis shows that the rIFN-g treatment is effective in decreasing the hazard of individuals getting infection. While in this analysis, the log hazard ratio was regressed only on treatment, future analysis can be conducted to explore different cox models to model the effect of different variables on log hazard ratio.

# 4 Appendix:

## 4.1 R Code:

```
rm(list=ls())
library(survival)
library(rms) # For: survplot

# My own function :)
se.pcnt <- function(p,km,e=.05){# p is a percentage for a percentile
  # Computes the standard error of a given percentile
  se.S <- km$std.err[which(km$surv<p)[1]]
  perc <-  km$time[which(km$surv<p)[1]]
  u <- tail(km$time[which(km$surv >= 1-p+e)],1)
  l <- km$time[which(km$surv <= 1-p-e)][1]
  S.u <- tail(km$surv[which(km$surv >= 1-p+e)],1)
  S.l <- km$surv[which(km$surv <= 1-p-e)][1]

  f <- (S.u-S.l) / (l-u)
  se <- se.S / f
  list("se"=se,"percentile"=perc)
}

# center
# treat ******
# sex
# age
# height
# weight
# inherit
# steroids
# propylac
# hos.cat
# time (days)
# status (1 = infection, 0 = censored) Want dead=infection=1 => data is good

#Main: ##############################################################################

  # Read in data:
  cgd <- read.csv("../../Data/cgd.csv")[,-c(1:2)]
  cat.i <- c(1,2,6,9)

  # Part 1: Logrank Test
  # KM Curve:
  km <- survfit(Surv(time,status) ~ treat, type="kaplan-meier",data=cgd)

  plot.km <- function(){
    survplot(km,col=c("purple","red"),lwd=1.2,lty=1,
         xlab="Time to Infection (Days)",ylab="Survival")
         title(main="KM Survival Curves for Infection Time")}

  # Find Median
```

```
plac.i <- which(cgd$treat!="placebo")
km.trmt <- survfit(Surv(time,status)~1, type="kaplan-meier",data=cgd[-plac.i,])
km.plac <- survfit(Surv(time,status)~1, type="kaplan-meier",data=cgd[plac.i,])
se.me.trmt <- se.pcnt(.5,km.trmt)
se.me.plac <- se.pcnt(.5,km.plac) #NA

CI.median.trmt <- qnorm(c(.025,.975),se.me.trmt$perc,se.me.trmt$se)
km.trmt.median <- matrix(c(se.me.trmt$perc,CI.median.trmt),1,3)
colnames(km.trmt.median) <- c("Estimate","CI.Lower","CI.Upper")

# Logrank Test:
logrank.treat <- survdiff(Surv(time,status) ~ treat, data=cgd)
p.logrank.treat <- pchisq(q=logrank.treat$chisq,df=1,lower.tail=F)
# p < .05 => treatment vs. placebo significantly different effects on survival.


# Part 2: Cox Model
# Model Selection: Stepwise Forward / Backward
cox.ful <- coxph(Surv(time,status) ~ ., data=cgd)
cox.red <- coxph(Surv(time,status) ~ 1, data=cgd)
cox.stp <- step(cox.red,scope=list(lower=cox.red,upper=cox.ful),data=cgd,
                direction="both")
# Age is not significant. So, throw out.

cox.mod <- coxph(Surv(time,status) ~ treat, data=cgd) # logrank test included
                                                      # same p value
cox.mod.coef <- summary(cox.mod)$conf.int

#predicted survival curve

# Log.h ~ Log.T: Parallel => Proportional Hazards Assumption Met
plot.log.H <- function(km,add=F,main="",col=1,type='p') {
  log.H <- log(-log(km$surv))
  log.T <- log(km$time)
  if (!add) {
    plot(log.T,log.H,col=col,pch=20,main=main,type=type,ylim=c(-5,1))
  } else {
    lines(log.T,log.H,col=col,pch=20,main=main,type=type)
  }
}

plot.model.ass <- function(){
  plot.log.H(km.trmt,col="red",type='s',
             main="Log Cumulative Hazard vs. Log Survival Time")
  plot.log.H(km.plac,col='purple',type='s',add=T)
  legend("topleft",col=c("red","purple"),lwd=3,legend=c("rIFN-g","Placebo"))
}

#survplot(km,col=c("purple","red"),loglog=T,logt=T,add=T)
```