Stat538 Midterm - Primary Central Nervous System Lymphoma

Arthur Lui

7 March 2014

---

1 Introduction

In previous studies, methotrexate-based chemotherapy has improved median survival time for patients with Primary Central Nervous System Lymphoma (PCNSL). Using a newly collected data set provided by Dr. Engler, we would like to verify these findings. The data contained many variables, including:

- DS: treatment (1 = methotrexate, 0 = cytarabine)
- OS: overall survival (Days)
- OSc: censoring indicator for overall survival
- PFS: progression-free survival (Days)
- PFSc: censoring indicator for progression-free survival

Overall survival refers to the time from treatment to death. Progression-free survival refers to the time from treatment to reappearance of cancer. We will determine, for each outcome measure, whether methotrexate-based chemotherapy improves median survival time for PCNSL patients.

2 Overall Survival

The overall survival Kaplan Meier (KM) Curves for each treatment (Methotrexate and cytarabine) were first plotted (see **Figure 1**). The KM curves do not intersect and are clearly separated. Their 95% confidence intervals do not overlap. The median survival time was computed to be **448** (95% CI: 298.34, 597.66) for patients that received the methotrexate treatment. The median survival time of the patients that received the cytarabine treatment was **1470** (95% CI: 715.69, 2224.31). Since the 95% confidence intervals for the median do not overlap, we conclude that the medians are not the same. In fact, from the survival curve, it appears that cytarabine significantly improves upon the methotrexate-patients survival at the median. The median survival time of the cytarabine group is **3.28** times that of the methotrexate group's. The results of a logrank test also indicates that the two treatments have a significant difference (p-value = 3.04e-05.

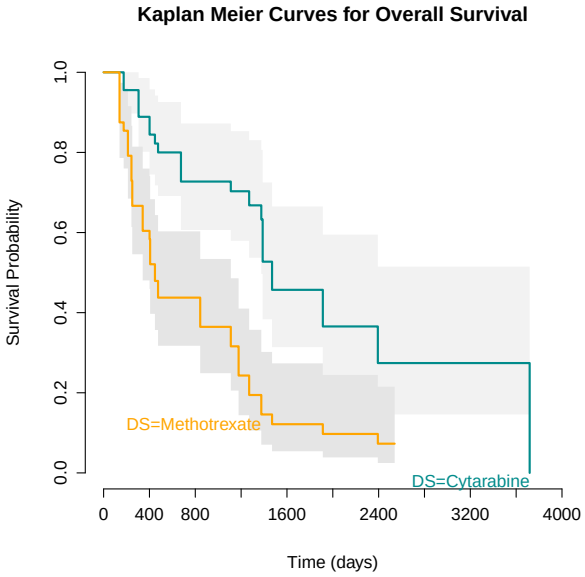

Figure 1: Overall survival KM curves for each treatment. Since the 95% confidence intervals do not overlap, and the confidence interval for cytarabine is greater than that of methotrexate, we conclude that cytarabine treatment's improvement of survival time is significantly greater than methotrexate treatment's improvement of survival time.

A cox model was fit by regressing the log hazards ratio on the treatment variable. That is,

$$log\left(\frac{h_i(t)}{h_0(t)}\right) = x\beta$$

where x is an indicator for whether the treatment received by a patient was methotrexate (1) or cytarabine (0). We obtain an estimate for the parameter $\beta = \hat{\beta}$, but interpret $e^{\hat{\beta}} = 2.82$ (95% CI: 1.7, 4.68). This means that the hazard rate of the methotrexate treatment group is 2.82 times that of the cytarabine treatment group. In other words, the methotrexate group has a higher risk of dying at any given time. A plot of the log cumulative hazard against the log time shows that the treatment groups' hazards are indeed proportional (see **Figure 2**), which is an assumption made when performing a cox regression.

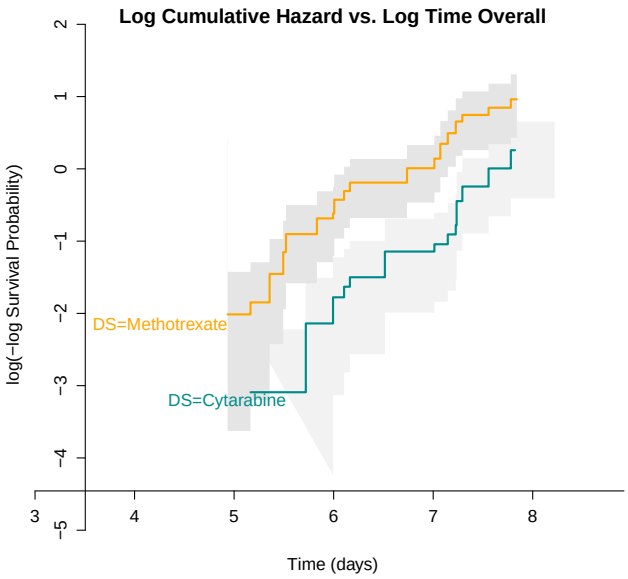

Figure 2: A plot of the log cumulative hazard against the log time shows that the treatment groups' hazards are indeed proportional.

3 Progression Free Survival

The progression-free Kaplan Meier survival curves for the two treatments cross over (**Figure 3**). So we cannot assume that the two treatment groups have proportional hazards. Moreover, we cannot model the effect of treatment with a cox model. We will, therefore, fit a parametric model to the data. Before doing so, we computed the median of the survival times to be 212 (95% CI 167.85, 256.15) for the methotrexate group, and 212 (95% CI 167.85, 256.15) for the cytarabine group. The intervals overlap, so the medians are not significantly different from each other.

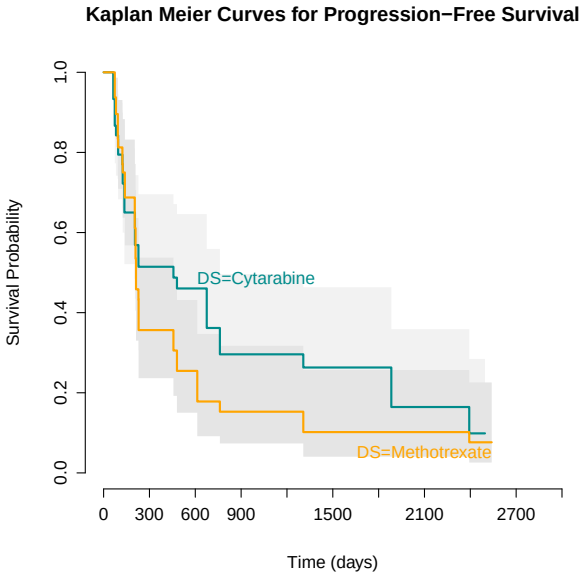

Figure 3: Progression free KM survival curve. The crossing over suggests that the hazards are not proportional. A cox model should not be used.

The KM curve and estimated parametric curves were plotted together (**Figure 4**). The lognormal (green) curve seems to follow the KM curve the closest. The probability plot for the lognormal distribution is not particularly straight. But it is reasonably straight. So we will use it as our modelling distribution. We will use the Accelerated Failure Time (AFT) model:

$$log(T_i) = \mu + \boldsymbol{X}\boldsymbol{\beta} + \sigma\epsilon_i$$

where T_i is the survival time with a lognormal distribution.

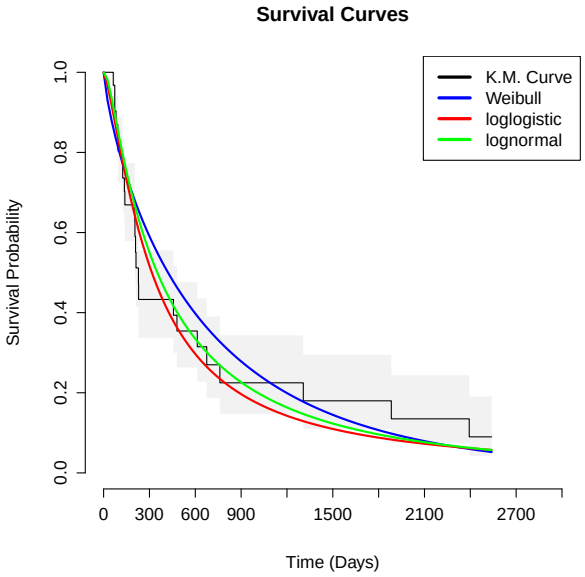

Figure 4: The lognormal (green) curve seems to follow the KM curve closest.

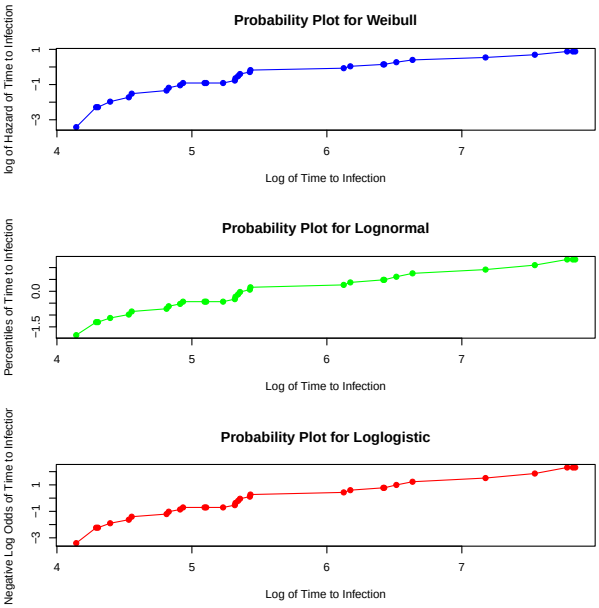

Figure 5: Probability plots seem reasonably straight. I will choose to use the lognormal distribution.

We use stepwise forward selection to obtain a model for survival time. The model selected did not contain the treatment variable, which we are most interested in. It has a p-value greater than .05. So treatment does not contribute to the model significantly in affecting the survival. Nevertheless, we can still fit a model to help us understand the relationship between survival and other variables. The model selected

includes the terms, KFS and LDH. Karnofsky performance status (KFS) is a measurement of the general well-being and activities of a cancer patient. It is used to determine further treatment for the patient. A higher score corresponds to a healthier and happier patient; a lower score corresponds to a sick and unhappy patient. CSF is a categorical variable. I dichotomized KFS to plot the relationship between KFS, CSF, and survival. **Figure 5** plots the survival for different combinations of $\boldsymbol{X}$, a matrix expressing the covariates. The parameter estimates are: $(\hat{\mu}, \hat{\sigma}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (4.49, 0.92, 0.99, 0.94, 1.22)$. So the median of the survival time for a person with high KFS is $e^{.99} \approx 2.7$ times that of a person with low KFS (<70).

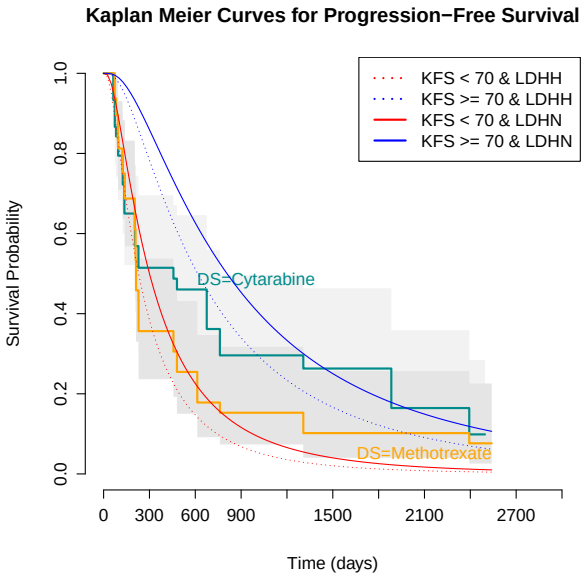

Figure 6: Parametric Survival Curves and Kaplan Meier Curves. Patients with high KFS have higher survival. Patients with LDHN also have higher survival.

The percentile-percentile plots suggest that the curves fitted may not be valid. The curve does not pass through 0, and does not appear linear. Further analysis can be done with more data to determine the validity of the model.

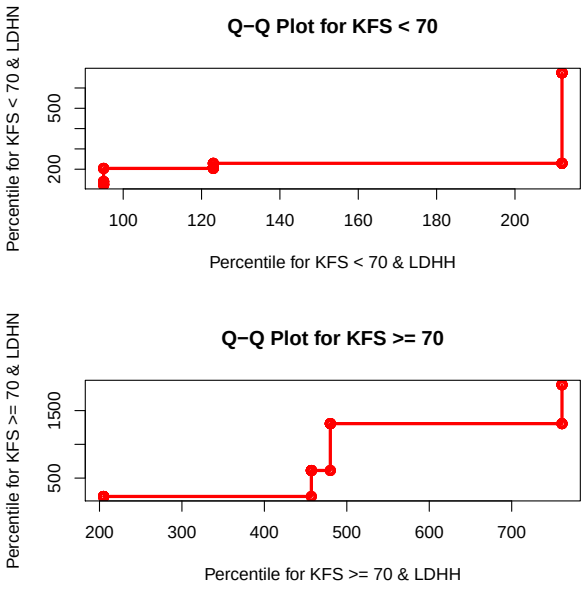

Figure 7: Percentile-percentile plots

The residuals are right skewed. But they are approximately normal. Once again, with more data, we can see the distribution of the residuals more clearly.

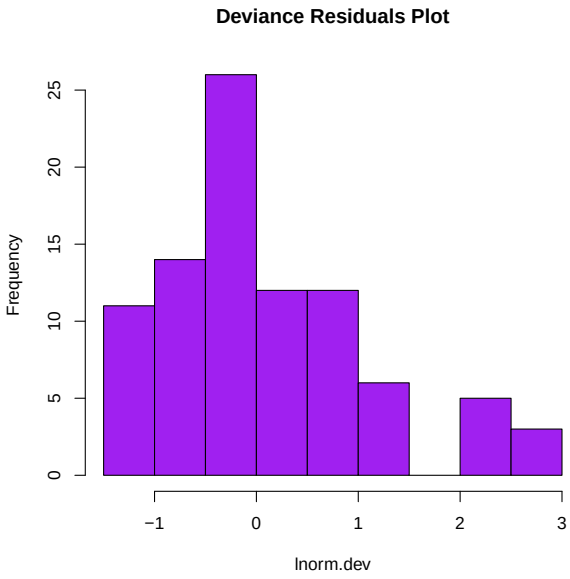

Figure 8: Plot of deviance residuals.

4 Conclusions

The median survival times for methotrexate patients is lower than that of the cytarabine patients' for both outcome measures. The median overall survival time of the cytarabine group was 3.28 times that of the methotrexate group's. The difference between overall survival times between the two treatment groups was significant.

The median progression-free survival times for the two treatment groups was not significant. But the median progression-free survival times for a person with high (≥ 70) KFS was **2.7** that of a person with low KFS (< 70). The difference in progression-free survival times between high and low KFS groups was significant.

5 Appendix:

5.1 R Code:

```
rm(list=ls())
library(survival)
library(rms)

# 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

  CI <- matrix(c(perc,qnorm(c(.025,.975),perc,se)),1,3)
  colnames(CI) <- c("Estimate","CI.Lower","CI.Upper")
  list("se"=se,"percentile"=perc,"CI"=CI)
}


# Read Data:
  pcnsl <- read.csv("../../Data/pcnsl.csv")
  pcnsl$DS <- ifelse(pcnsl$DS==1,"Methotrexate","Cytarabine")
  #pcnsl <- pcnsl[,-c(4,5,6)]
  #pcnsl <- pcnsl[which(pcnsl$Gender!="" & !is.na(pcnsl$KFS)),]


# OS:#############################################################################
  # Plot KM curves:
  km.os  <- survfit(Surv(OS,OSc) ~ DS, data=pcnsl,type="kaplan-meier")
  plot.km.os <- function(){
    survplot(km.os,xlab="Time (days)",lty=1,lwd=2,col=c("cyan4","orange")) #no cross
             title("Kaplan Meier Curves for Overall Survival")
  }

  # Cox Model Selection: Nothing really significant
    # cox.os <- coxph(Surv(OS,OSc) ~ Gender, data=pcnsl) Gender not signif.
    #(cox.os.f <- coxph(Surv(OS,OSc) ~ .^2, data=pcnsl[,-c(7,8)]))
    #(cox.os.1 <- coxph(Surv(OS,OSc) ~ 1, data=pcnsl[,-c(7,8)]))
    # cox.step <- step(cox.os.1,scope=list(lower=cox.os.1,upper=cox.os.f),
    #                 data=pcnsl,dir="both")

  # Logrank Test:
    cox.os <- coxph(Surv(OS,OSc) ~ DS, data=pcnsl)
    logrank.os.p <- 1-pchisq(cox.os$score,1) # <.05 => One treatment better
```

```
  # Plot log.H ~ log.T to Test Proportional Hazards Assumption:
    plot.os.log.H <-function() { # Parallel => Proportional Hazards Assumptions Met
    survplot(km.os,xlab="Time (days)",lty=1,lwd=2,col=c("cyan4","orange"),
           loglog=T,logt=T)
           title("Log Cumulative Hazard vs. Log Time Overall")
    }

  # Find Median:
    meth.i <- which(pcnsl$DS=="Methotrexate")
    km.os.meth <- survfit(Surv(OS,OSc)~1,type="kaplan-meier",data=pcnsl[meth.i,])
    km.os.cyba <- survfit(Surv(OS,OSc)~1,type="kaplan-meier",data=pcnsl[-meth.i,])
    se.os.meth <- se.pcnt(.5,km.os.meth)
    se.os.cyba <- se.pcnt(.5,km.os.cyba)

    CI.os.meth <- round(se.os.meth$CI,2)
    CI.os.cyba <- round(se.os.cyba$CI,2)
    # Intervals don't overlap => Significantly different

  # Parameter Estimates & CI:
    cox.os.coef <- summary(cox.os)$conf.int


# PF:#############################################################################

  # KM Plots:
  km.pf  <- survfit(Surv(PFS,PFSc) ~ DS, data=pcnsl,type="kaplan-meier")
  plot.km.pf <- function() {
    survplot(km.pf,xlab="Time (days)",lty=1,lwd=2,col=c("cyan4","orange")) # crossing
             title("Kaplan Meier Curves for Progression-Free Survival") # => AFT
  }

  cox.pf <- coxph(Surv(PFS,PFSc) ~ DS, data=pcnsl)
  logrank.pf.p <- 1-pchisq(cox.pf$score,1) # >.05 => Not better

  plot.pf.log.H <-function() { # Crossing => Proportional Hazards Assumptions Not Met
    survplot(km.pf,xlab="Time (days)",lty=1,lwd=2,col=c("cyan4","orange"),
           loglog=T,logt=T)
           title("Log Cumulative Hazard vs. Log Time Overall")
  }

  # Find Median:
    km.pf.meth <- survfit(Surv(PFS,PFSc)~1,type="kaplan-meier",data=pcnsl[meth.i,])
    km.pf.cyba <- survfit(Surv(PFS,PFSc)~1,type="kaplan-meier",data=pcnsl[-meth.i,])
    se.pf.meth <- se.pcnt(.5,km.pf.meth)
    se.pf.cyba <- se.pcnt(.5,km.pf.cyba)

    CI.pf.meth <- se.pf.meth$CI
    CI.pf.cyba <- se.pf.cyba$CI
    # Intervals overlap => NOT Significantly different
```

```
#0: Survival Functions:
sweib <- function(x,gamma,lambda) exp(-(x/lambda)^gamma)
sllog <- function(x,beta,alpha) 1-(((x/alpha)^(-beta)) + 1)^(-1)
slnorm <- function(x,mu,sigma) pnorm((log(x)-mu)/sigma,lower=F)

#1: Fit AFT Models: (Weibull, loglogistic, lognormal)
weib <- survreg(Surv(PFS,PFSc)~1,data=pcnsl,dist="weibull")
weib.lambda <- exp(weib$coef[[1]]) # lambda = scale = exp(Intercept)
weib.gamma <- 1/weib$scale         # gamma = shape = 1/scale

llog <- survreg(Surv(PFS,PFSc)~1,data=pcnsl,dist="loglogistic")
llog.lambda <- exp(llog$coef[[1]]) # lambda = scale = exp(Intercept)
llog.gamma <- 1/llog$scale         # gamma = shape = 1/scale

lnorm <- survreg(Surv(PFS,PFSc)~1,data=pcnsl,dist="lognormal")
lnorm.mu <- lnorm$coef[[1]]
lnorm.sigma <- lnorm$scale

#2: Plot Parametric and Non-Parametric (KM) Curves:
km.pf.1  <- survfit(Surv(PFS,PFSc) ~ 1, data=pcnsl,type="kaplan-meier")
plot.distributions <- function(cex=1){
  survplot(km.pf.1,xlab="Time (Days)");title("Survival Curves")
  to <- max(km.pf$time)
  curve(sweib(x,weib.gamma,weib.lambda),fr=0,to=to,add=T,col="blue",lwd=2)
  curve(sllog(x,llog.gamma,llog.lambda),fr=0,to=to,add=T,col="red",lwd=2)
  curve(slnorm(x,lnorm.mu,lnorm.sigma),fr=0,to=to,add=T,col="green",lwd=2)
  legend("topright",legend=c("K.M. Curve","Weibull","loglogistic","lognormal"),
         col=c("black","blue","red","green"),lwd=3,cex=cex)
# It appears that the lognormal curve follows the K.M. curve the closest.
}

#3: Assess Parametric Assumptions:
#summary(km.pf.1)

# Weibull
plot.weib.prob <- function()
  plot(log(km.pf.1$time), log(log(1/(km.pf.1$surv))),pch=19,
       xlab="Log of Time to Infection",
       ylab="log of Hazard of Time to Infection",
       main="Probability Plot for Weibull",col="blue",type='o')

# Lognormal
plot.lnorm.prob <- function()
  plot(log(km.pf.1$time), qnorm(1-km.pf.1$surv),pch=19,
       xlab="Log of Time to Infection",
       ylab="Percentiles of Time to Infection",
       main="Probability Plot for Lognormal",col="green",type='o')

# Loglogistic
plot.llog.prob <- function()
```

```
  plot(log(km.pf.1$time), log((1/(km.pf.1$surv))-1),pch=19,
       xlab="Log of Time to Infection",
       ylab="Negative Log Odds of Time to Infection",
       main="Probability Plot for Loglogistic",col="red",type='o')

plot.pp <- function() {
  par(mfrow=c(3,1))
  plot.weib.prob()
  plot.lnorm.prob()
  plot.llog.prob()
  par(mfrow=c(1,1))
}
# They all look bad. But go with lnorm because plot.KM appears to be closest.

#4: Fit Using Covariates:
#Parametric Model
plot.km.pf()
temp <- pcnsl[,-c(9,10)]
temp$KFS <- ifelse(temp$KFS<70,0,1)
lnorm.f  <- survreg(Surv(PFS,PFSc)~.,dat=temp,dist="lognormal")
lnorm.1  <- survreg(Surv(PFS,PFSc)~1,dat=temp,dist="lognormal")
lnorm.s  <- step(lnorm.1,scope=list(lower=lnorm.1,upper=lnorm.f),
               data=temp,direction="both")
lnorm.s <- update(lnorm.s, .~. -CSFpro)
summary(lnorm.s)

# Parameter Estimates:
m     <- lnorm.s$coef[[1]]
s     <- lnorm.s$scale
coef  <- lnorm.s$coef[-1]
en <- function(x,b=coef){
  exp(x %*% b)
}

#km.pf.s <- survfit(Surv(PFS,PFSc)~KFS+LDH,data=temp,type="kaplan-meier")
#survplot(km.pf.s)

plot.pf.s <-function(cex=1,lwd=1) {
  to <- max(c(temp$PFS,temp$PFSc))
  plot.km.pf()
  e <-en(c(0,1,0));curve(slnorm(x/e,m,s),fr=0,to=to,add=T,col="red",lty=3,lwd=lwd)
  e <-en(c(1,1,0));curve(slnorm(x/e,m,s),fr=0,to=to,add=T,col="blue",lty=3,lwd=lwd)
  e <-en(c(0,0,1));curve(slnorm(x/e,m,s),fr=0,to=to,add=T,col="red",lwd=lwd)
  e <-en(c(1,0,1));curve(slnorm(x/e,m,s),fr=0,to=to,add=T,col="blue",lwd=lwd)
  legend("topright",legend=c("KFS < 70 & LDHH","KFS >= 70 & LDHH",
                             "KFS < 70 & LDHN","KFS >= 70 & LDHN"),
         cex=cex,lwd=2,col=rep(c("red","blue"),2),lty=c(3,3,1,1))
}

#5: Model Fit
# Assess AFT Assumption: Q-Q plot
```

```
  # One for each covariate

  lh <- temp[which(temp$LDH=="H" & temp$KFS==0),]
  hh <- temp[which(temp$LDH=="H" & temp$KFS==1),]
  ln <- temp[which(temp$LDH=="N" & temp$KFS==0),]
  hn <- temp[which(temp$LDH=="N" & temp$KFS==1),]

  KM.lh <- survfit(Surv(PFS,PFSc)~1,data=lh)
  KM.hh <- survfit(Surv(PFS,PFSc)~1,data=hh)
  KM.ln <- survfit(Surv(PFS,PFSc)~1,data=ln)
  KM.hn <- survfit(Surv(PFS,PFSc)~1,data=hn)

  lh.p <- hh.p <- ln.p <- hn.p <- NA
  #p <- seq(.1,1,by=.1)
  p <- 1:100/100
  for(i in 1:length(p)) {
    lh.p[i] <- min(KM.lh$time[KM.lh$surv <= (1-p[i])])
    hh.p[i] <- min(KM.hh$time[KM.hh$surv <= (1-p[i])])
    ln.p[i] <- min(KM.ln$time[KM.ln$surv <= (1-p[i])])
    hn.p[i] <- min(KM.hn$time[KM.hn$surv <= (1-p[i])])
  }

  L.index <- min(c(sum(lh.p < Inf), sum(ln.p < Inf)))
  plot.L.qq <- function(){
    plot(lh.p[1:L.index],ln.p[1:L.index],type="o",
         main="Q-Q Plot for KFS < 70",col="red",lwd=3,
         xlab="Percentile for KFS < 70 & LDHH",
         ylab="Percentile for KFS < 70 & LDHN")
  }

  H.index <- min(c(sum(hh.p < Inf), sum(hn.p < Inf)))
  plot.H.qq <- function(){
    plot(hh.p[1:H.index],hn.p[1:H.index],type="o",
         main="Q-Q Plot for KFS >= 70",col="red",lwd=3,
         xlab="Percentile for KFS >= 70 & LDHH",
         ylab="Percentile for KFS >= 70 & LDHN")
  } # Warrants further investigation

#6: Assessing Overall Fit: Deviance Residuals
    lnorm.dev <- residuals(lnorm.s,type="deviance")
    plot.resid <- function()
      plot(lnorm.dev,pch=19,col='purple',main="Residuals Plot")
    # Centered around zero, no obvious patterns.

#Plots for PF Section: ##################################################
# Plots:
#plot.km.pf()
#plot.pf.log.H()
#plot.distributions()
#plot.weib.prob()
#plot.lnorm.prob()
```

```
#plot.llog.prob()
#plot.pf.s()
#plot.L.qq()
#plot.H.qq()
#plot.resid()


# Tables:
#cox.pf
#logrank.pf.p
#summary(lnorm.s)

# Notes: ##################################################################
# 1) No proportional hazards for second plot => use AFT for second.
# 2) This website suggests that the cytarabine + methotrexate should
#    be better than methotrexate alone:
# http://www.ncbi.nlm.nih.gov/pubmed/19767089
```