

# Praca domowa nr 1 - raport

*Gabriel Bożek, Luiza Dobosz, Anna Sikorska, Aneta Skorupska*

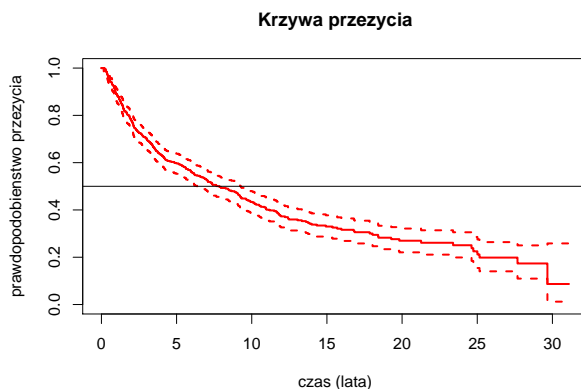
*30.03.2020*

Otrzymaliśmy dane dotyczące 517 chorych na chłoniaka leczonych radioterapią i, ewentualnie, chemioterapią. Dane zawierały następujące zmienne:

- stnum - identyfikator chorego
- age - wiek chorego w chwili diagnozy (lata)
- clinstg - stopień zaawansowania choroby (1 – I, 2 – II)
- hgb - poziom hemoglobiny (g/l)
- chemo - wskaźnik leczenia chemioterapią (0 – nie, 1 – tak)
- dftime - czas do wystąpienia nawrotu choroby lub zgonu (lata)
- dfstat - wskaźnik zdarzenia (0 – żył bez nawrotu, 1 – nawrót i/lub zgon)

Chcemy sprawdzić, które ze zmiennych charakteryzujących chorego mają wpływ na czas przeżycia bez nawrotu choroby, tzn. czas do wystąpienia nawrotu lub zgonu.

Najpierw narysowaliśmy wykres krzywej przeżycia dla całego zbioru danych z przedziałami ufności wyznaczonymi metodą log-log.

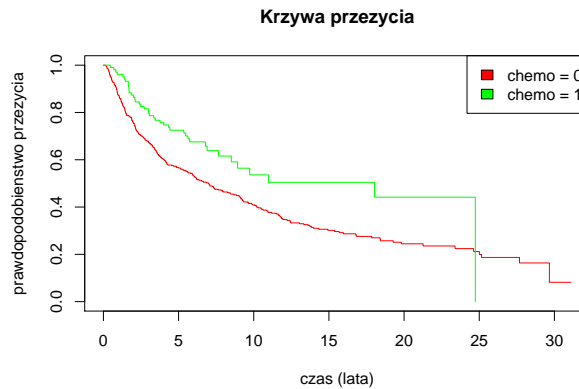


Krzywa kończy się po 31.10198 latach, kiedy to ostatni pacjent zrezygnował z badań (żył bez nawrotu). Następnie obliczyliśmy medianę czasu do wystąpienia zdarzenia, która wynosi 7.835729 (nanieśliśmy na wykres prostą oznaczającą prawdopodobieństwo przeżycia równe 50%). Mediana jest stosunkowo mała w porównaniu do maksymalnego czasu na wykresie oraz wykres poniżej prostej zmniejsza nachylenie. Podejrzewamy więc, że pacjenci, dla których leczenie przyniosło rezultaty przeżywają dość duży okres czasu.

Przetestujemy każdy z podanych czynników, sprawdzając czy ma wpływ na czas przeżycia. Wszystkie testy oprzemy na teście logrank, przyjmując poziom istotności równy 0.05.

Zacznijmy od stopnia zaawansowania choroby. Otrzymaliśmy p-value równe 0.2, więc nie możemy stwierdzić, że stopień zaawansowania choroby ma znaczenie.

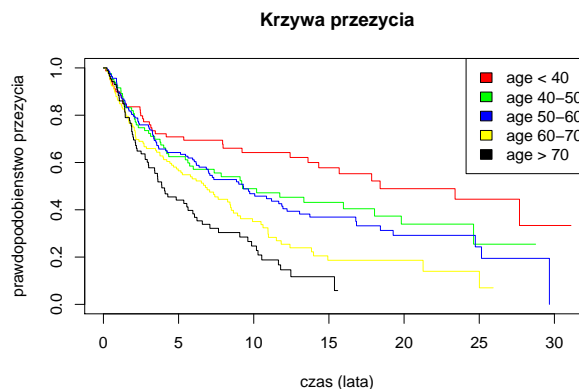
Dla chemioterapii otrzymaliśmy p-value równe 0.002, więc możemy przyjąć, że chemioterapia wpływa na czas przeżycia. Możemy to zauważyć na poniższym wykresie.



Widzimy, że osoby z chemioterapią przeżywają dłużej.

Ponieważ poziom hemoglobiny może przyjmować wiele różnych wartości, to podzieliśmy pacjentów na grupy posługując się kwartylami. Następnie sprawdziliśmy czy w każdej grupie jest porównywalna liczba zdarzeń. Otrzymaliśmy p-value równe 0.9, więc poziom hemoglobiny nie ma wpływu na czas przeżycia.

Ponieważ wiek może przyjmować wiele różnych wartości, to podzieliśmy pacjentów na typowe grupy wiekowe. Tak jak w przypadku hemoglobiny ocenialiśmy liczbę zdarzeń w każdej grupie. Otrzymaliśmy p-value równe  $1e-07$ , więc możemy przyjąć, że wiek wpływa na czas przeżycia.



Analizując wykres możemy wysnuć przypuszczenie, że istnieje trend krzywych przeżycia ze względu na wiek. Sprawdźmy to testem dla trendu. Otrzymaliśmy p-value równe 0.022954, co potwierdza nasze przypuszczenia, że krzywa przeżycia maleje ze względu na wiek.

Przetestujmy teraz dokładniej czynniki, które mają wpływ na krzywą przeżycia, stosując testy warstwowe. Dla każdego czynnika wykonamy trzy oddzielne testy, każdy z inną zmienną towarzyszącą. Zauważmy, że mamy tu do czynienia z wielokrotnym testowaniem. Uwzględniając fakt, że pracujemy na danych medycznych, przy których błędne odrzucenie poprawnej hipotezy może być kosztowne, zastosowaliśmy konserwatywną poprawkę Bonferroniego. Dzięki temu testy warstwowe dla chemioterapii i dla wieku mają poziom istotności równy  $0.05/3 = 0.017$ .

Zacznijmy od chemioterapii. Jako pierwszy rozważmy test chemioterapii z podziałem na warstwy stopnia zaawansowania choroby. Otrzymaliśmy p-value równe  $8e-04$ , więc potwierdza to, że chemioterapia ma wpływ na długość przeżycia. Dodatkowo zauważmy, że otrzymane p-value jest mniejsze od uzyskanego w zwykłym teście logrank dla chemioterapii (0.002), więc analiza warstwowa mocniej utwierdza nas w przekonaniu, że chemioterapia ma znaczenie.

W przypadku hemoglobiny wyciągnęliśmy takie same wnioski (uzyskaliśmy takie samo p-value równe 0.002).

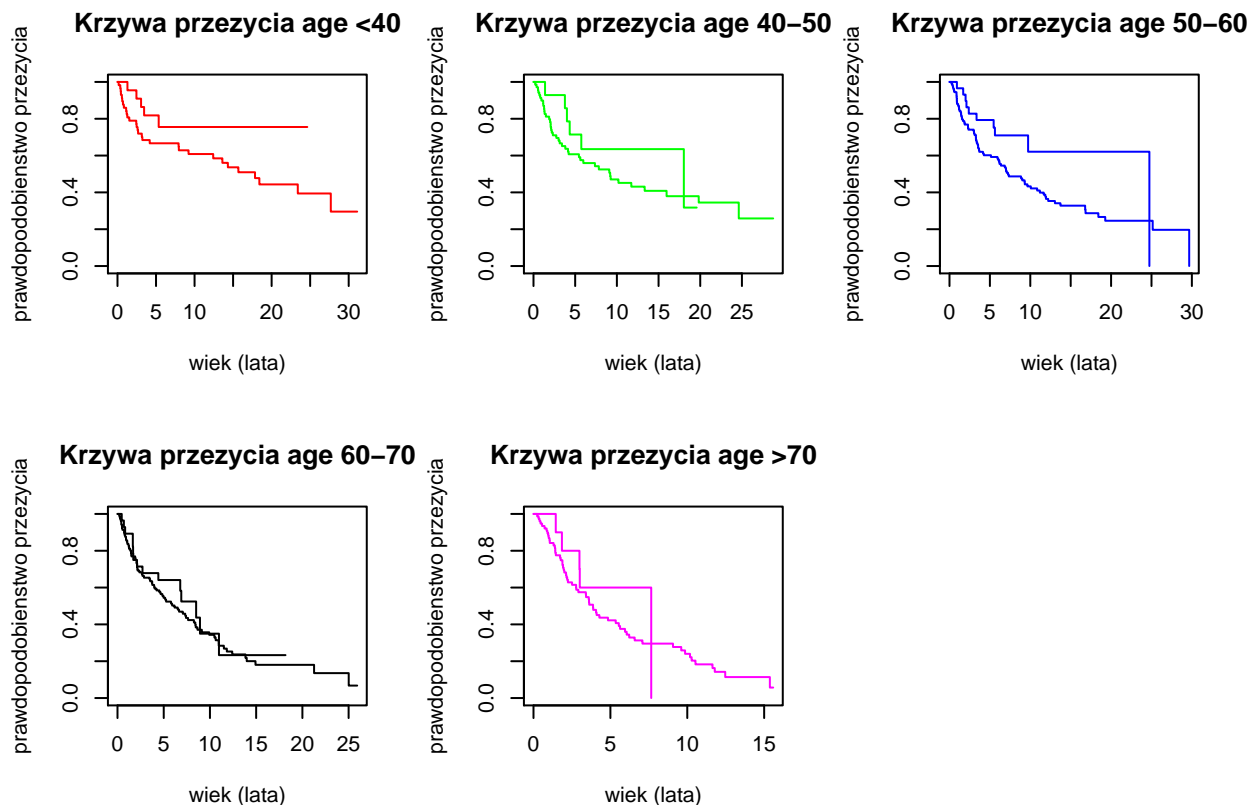
Zanim zrobimy test warstwowy dla wieku sprawdzimy czy istnieje zależność wieku od stosowania chemioterapii, gdyż oba czynniki mają wpływ na krzywą przeżycia. W tym celu porównujemy liczbę pacjentów z chemioterapią w stosunku do liczby wszystkich pacjentów w grupie wiekowej.

procent chemioterapii w grupie wiekowej  $< 40$  wynosi 0.278481  
 procent chemioterapii w grupie wiekowej (41,50) wynosi 0.168674  
 procent chemioterapii w grupie wiekowej (51,60) wynosi 0.2116788  
 procent chemioterapii w grupie wiekowej (61,70) wynosi 0.2121212  
 procent chemioterapii w grupie wiekowej  $> 70$  wynosi 0.1162791

Na podstawie uzyskanych wyników nie jesteśmy w stanie dostrzec żadnej zależności.

W przeprowadzonym teście warstwowym dla wieku p-value wynosiło 0.009 - wpływ chemioterapii jest wciąż istotny statystycznie, ale mamy mniejszy poziom krytyczny testu. Zatem analiza warstwowa jest mniej zadowalająca.

Sprawdźmy jak wyglądają wykresy przeżycia dla każdej z warstw.



Zauważamy, że potrzeba dalszych badań w tym zakresie, gdyż nie da się stwierdzić czy istnieje wpływ wieku na skuteczność chemioterapii.

Po zrobieniu testów dla wieku z podziałem na warstwy stopnia zaawansowania choroby, hemoglobiny i chemioterapii otrzymaliśmy p-value odpowiednio równe:  $8e-09$ ,  $8e-08$ ,  $7e-07$ , (p-value testu logrank dla wieku wynosiło  $1e-7$ ), więc możemy wyciągnąć te same wnioski dla każdej warstwy, co w przypadku analizy warstwowej chemioterapii.

Podsumowując, chemioterapia i wiek mają wpływ na czas przeżycia bez nawrotu choroby. W przypadku chemioterapii jest to wpływ pozytywny, a w przypadku wieku negatywny.

```

df <- read.csv("C://Users//PC/Desktop//Studia//BIO//PDI//follic_short_cr.csv")
library(survMisc);library(survival);library(foreign);library(rms);library(dplyr)
#----- ogólna krzywa przeżycia i mediana
par(mfrow = c(1,1))
krzywa_przezycia <- survfit(Surv(dftime, dstat) ~ 1,data = df,conf.type = "log-log")
plot(krzywa_przezycia, col = c("red"), lty = c(2),xlab = "czas (lata)",ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia")
abline(0.5, 0)

mediana <- krzywa_przezycia$time[which(krzywa_przezycia$surv<0.5)[1]]
max(df$dftime) # ostatnia osoba wycofała się z badaniach po 31.10198 latach
#----- st. zaawansowania choroby
krzywa_przezycia_clinstg1 <- survfit(Surv(dftime, dstat) ~ 1,data = df,subset = clinstg == 1,se.fit = FALSE)
krzywa_przezycia_clinstg2 <- survfit(Surv(dftime, dstat) ~ 1,data = df,subset = clinstg == 2,se.fit = FALSE)
plot(krzywa_przezycia_clinstg2,col="red", xlab = "czas (lata)", ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia")
lines(krzywa_przezycia_clinstg1,col="green")
legend("topright", legend = c("clinstg = 2", "clinstg = 1"), fill = c("red", "green"))

# Test, czy te dwie krzywe różnią się istotnie, test logrank dymprobkowy dla st. zaawansowania choroby
test_clinstg <- survdiff(Surv(dftime, dstat) ~ clinstg, data = df, rho = 0)
print(test_clinstg) # p = 0.2 => krzywe nie różnią się istotnie
#----- chemioterapia
krzywa_przezycia_chemio0 <- survfit(Surv(dftime, dstat) ~ 1,data = df,subset = chemo == 0,se.fit = FALSE)
krzywa_przezycia_chemio1 <- survfit(Surv(dftime, dstat) ~ 1,data = df,subset = chemo == 1,se.fit = FALSE)
plot(krzywa_przezycia_chemio0,col="red",xlab = "czas (lata)", ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia")
lines(krzywa_przezycia_chemio1,col="green")
legend("topright", legend = c("chemo = 0", "chemo = 1"), fill = c("red", "green"))

# test, czy te krzywe różnią się istotnie
test_chemio <- survdiff(Surv(dftime, dstat) ~ chemo, data = df, rho = 0)
print(test_chemio) # p = 0.002 => krzywe różnią się, ludzie z chemioterapią przeżywają dłużej
#----- hemoglobina
# tworzymy grupy dla hemoglobiny, podział ze względu na kwantyle:
df$hb2<- cut(df$hb, breaks = c(0, 130, 140, 150, 100000),labels = c('<=130', '131-140', '141-150', '>=151'), include.lowest = TRUE)
# sprawdzamy, czy podział jest ok, liczba jedynek
df %>% group_by(hb2) %>% summarize(liczba_przypadkow = sum(dstat, na.rm = TRUE))
# podział taki, żeby w każdej grupie była ta sama (mniej więcej) liczba śmierci
krzywa_przezycia_hgb130 <- survfit(Surv(dftime, dstat) ~ 1, data=df, subset = hb2 == '<=130', se.fit = FALSE)
krzywa_przezycia_hgb140 <- survfit(Surv(dftime, dstat) ~ 1, data=df, subset = hb2 == '131-140',se.fit = FALSE)
krzywa_przezycia_hgb150 <- survfit(Surv(dftime, dstat) ~ 1, data=df, subset = hb2 == '141-150', se.fit = FALSE)
krzywa_przezycia_hgb151 <- survfit(Surv(dftime, dstat) ~ 1, data=df, subset = hb2 == '>=151', se.fit = FALSE)
plot(krzywa_przezycia_hgb130, col="red", xlab = "czas (lata)", ylab = "prawdopodobieństwo przeżycia", main = "Krzywa przeżycia")
lines(krzywa_przezycia_hgb140,col="green")
lines(krzywa_przezycia_hgb150,col="blue")
lines(krzywa_przezycia_hgb151,col="yellow")
legend("topright", legend = c("hgb <= 130", "hgb 131-140", "hgb 141-150", "hgb >= 151"),fill = c("red", "green", "blue", "yellow"))

test_hgb <- survdiff(Surv(dftime, dstat) ~ hb2, data = df, rho = 0)
print(test_hgb) # p = 0.9, nie odrzucamy H0, nie ma istotnej różnicy między krzywymi
#----- wiek, grupujemy wiek:
df$age_groups2<- cut(df$age, breaks = c(0, 40, 50, 60, 70, 90),labels = c(1, 2, 3, 4, 5),include.lowest = TRUE)
df$age_groups2 <- as.numeric(df$age_groups2)
# sprawdzamy, czy podział jest ok
df %>% group_by(age_groups2) %>% summarize(liczba_przypadkow = sum(dstat, na.rm = TRUE))
krzywa_przezycia_age40 <- survfit(Surv(dftime, dstat) ~ 1, data=df,subset = age_groups2 == 1, se.fit = FALSE)
krzywa_przezycia_age50 <- survfit(Surv(dftime, dstat) ~ 1, data=df,subset = age_groups2 == 2,se.fit = FALSE)
krzywa_przezycia_age60 <- survfit(Surv(dftime, dstat) ~ 1, data=df, subset = age_groups2 == 3,se.fit = FALSE)
krzywa_przezycia_age70 <- survfit(Surv(dftime, dstat) ~ 1, data=df, subset = age_groups2 == 4,se.fit = FALSE)
krzywa_przezycia_age71 <- survfit(Surv(dftime, dstat) ~ 1, data=df,subset = age_groups2 == 5,se.fit = FALSE)
plot(krzywa_przezycia_age40,col="red", xlab = "czas (lata)", ylab = "prawdopodobieństwo przeżycia", main = "Krzywa przeżycia")
lines(krzywa_przezycia_age50,col="green")
lines(krzywa_przezycia_age60,col="blue")
lines(krzywa_przezycia_age70,col="yellow")
lines(krzywa_przezycia_age71)
legend("topright",legend=c("age < 40", "age 40-50", "age 50-60", "age 60-70", "age > 70"),fill=c("red", "green", "blue", "yellow", "black"))

test_age <- survdiff(Surv(dftime, dstat) ~ age_groups2, data = df, rho = 0)
print(test_age) # p = 1e-07 < 0.05 => odrzucamy H0 => krzywe różnią się istotnie
# test trendu dla wieku
trend_age <- ten(Surv(dftime, dstat) ~ age_groups2, data = df)
print(trend_age)
comp(trend_age) # p-value = 0.022954 < 0.05 => odrzucamy H0 => jest trend
#----- testy z warstwami
# Robimy warstwy dla tych zmiennych, co mają znaczenie.
#----- chemioterapia
# chemioterapia i st. zaawansowania choroby
test_chemio_clinstg <- survdiff(Surv(dftime, dstat) ~ chemo + strata(clinstg), data = df)
print(test_chemio_clinstg) # p = 8e-04, p-value się zmniejszyło
#chemioterapia i hemoglobina
test_chemio_hgb <- survdiff(Surv(dftime, dstat) ~ chemo + strata(hgb2), data = df)
print(test_chemio_hgb) # p = 0.002, p-value się zwiększyło
# chemioterapia i wiek
test_chemio_age <- survdiff(Surv(dftime, dstat) ~ chemo + strata(age_groups2), data = df)
print(test_chemio_age) # p = 0.009
krzywa_przezycia_chemio_age40 <- survfit(Surv(dftime, dstat)~chemo,data=df,subset=age_groups2==1,se.fit=FALSE)
krzywa_przezycia_chemio_age50 <- survfit(Surv(dftime, dstat) ~ chemo,data=df,subset=age_groups2==2,se.fit=FALSE)
krzywa_przezycia_chemio_age60 <- survfit(Surv(dftime, dstat) ~ chemo,data=df,subset=age_groups2 == 3,se.fit = FALSE)
krzywa_przezycia_chemio_age70 <- survfit(Surv(dftime, dstat) ~ chemo,data=df,subset=age_groups2 == 4,se.fit = FALSE)
krzywa_przezycia_chemio_age90 <- survfit(Surv(dftime, dstat) ~ chemo,data=df,subset=age_groups2 == 5,se.fit = FALSE)
par(mfrow = c(2,3))
plot(krzywa_przezycia_chemio_age40,col="red", xlab = "wiek (lata)", ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia age <40")
plot(krzywa_przezycia_chemio_age50,col="green", xlab = "wiek (lata)", ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia age 40-50")
plot(krzywa_przezycia_chemio_age60,col="blue", xlab = "wiek (lata)", ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia age 50-60")
plot(krzywa_przezycia_chemio_age70,col="black", xlab = "wiek (lata)", ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia age 60-70")
plot(krzywa_przezycia_chemio_age90,col="magenta", xlab = "wiek (lata)", ylab = "prawdopodobieństwo przeżycia",main = "Krzywa przeżycia age >70")
#----- testy z warstwami dla wieku
test_age_clinstg <- survdiff(Surv(dftime, dstat) ~ age_groups2 + strata(clinstg), data = df)
print(test_age_clinstg) # p = 8e-09
test_age_hgb <- survdiff(Surv(dftime, dstat) ~ age_groups2 + strata(hgb2), data = df)
print(test_age_hgb) # p = 8e-08
test_age_chemio <- survdiff(Surv(dftime, dstat) ~ age_groups2 + strata(chemo), data = df)
print(test_age_chemio) # p = 7e-07
# sprawdzamy czy wiek i chemioterapia są jakos powiazane?
mean(df[df$age_groups2=="<40",]$chemo) # 0.278481
mean(df[df$age_groups2=="40-50",]$chemo) # 0.1686747 => 16% ma chemioterapię
mean(df[df$age_groups2=="50-60",]$chemo) # 0.2116788
mean(df[df$age_groups2=="60-70",]$chemo) # 0.2121212
mean(df[df$age_groups2==">70",]$chemo) # 0.1162791

```