Directives 4 and 7

Department of Biostatistics @ University of Washington

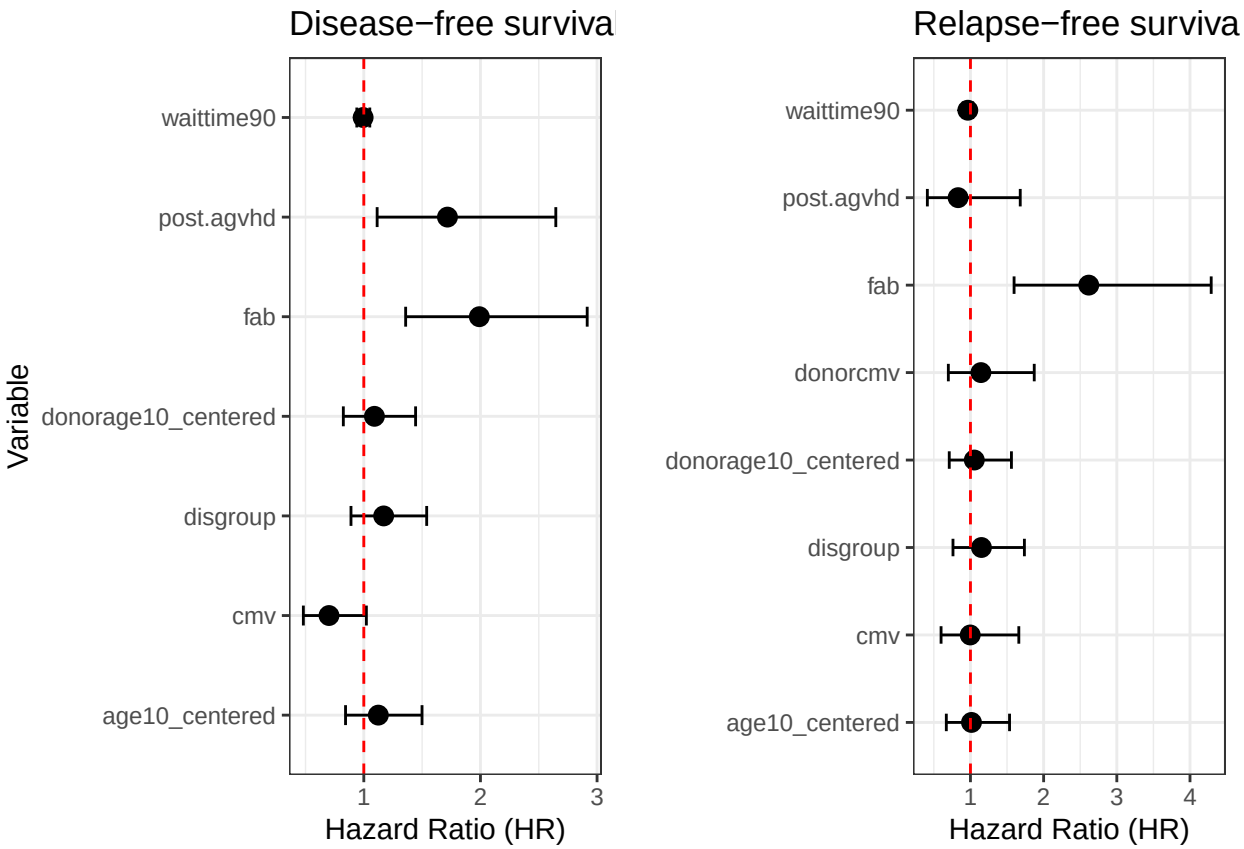
Alejandro Hernandez

Winter Quarter 2025

4. It is generally thought that aGVHD has an anti-leukemic effect. Based on the available data, is occurrence of aGVHD after transplantation associated with improved disease-free survival? Is it associated with a decreased risk of relapse? In view of this, do you consider aGVHD as an important prognostic event?

We believe the association of aGHVD with disease-free survival may be confounded by the baseline state of a patient's cancer, as well as determinants of their health and the health of their transplanted bone marrow. Then, when investigating the statistical effect of aGVHD, we elect to adjust for the wait time between diagnosis and transplant, the leukemia disease group, and the FAB leukemia subtype. Additionally, we elect to adjust for a patient's age and CMV status, as well as the age of their donor. Finally, we elect to stratify our proportional hazards model by hospital as we believe it may an additional confounder on our investigated relationship, and weaken the assumption of proportional hazards across hospitals.

We believe the association of aGHVD with relapse-free survival may be confounded by the baseline state of a patient's leukemia, as well as determinants of their health and the health of their donor. Then, when investigating the statistical effect of aGVHD, we elect to adjust for the wait time between diagnosis and transplant, the leukemia disease group, and the FAB leukemia subtype. Additionally, we elect to adjust for the age and CMV status of a patient and their donor. Finally, we elect to stratify our proportional hazards model by hospital as we believe it may an additional confounder on our investigated relationship, and weaken the assumption of proportional hazards across hospitals.

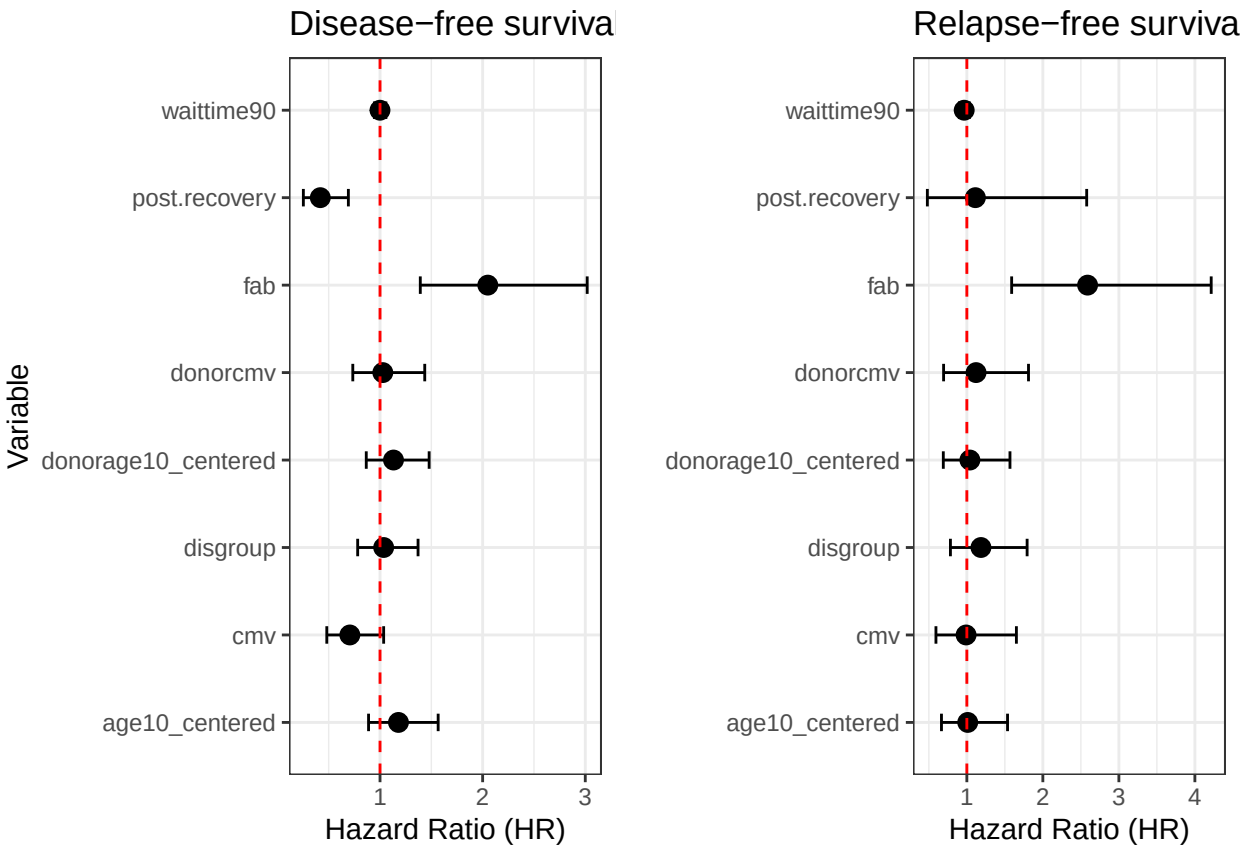


We fit two Cox proportional hazards models, one for disease-free survival and another for relapse-free survival. In both models, our target is the hazard ratio associated with the development of aGHVD, which is modeled as a time-varying covariate. In the second PH model, death is a competing event with relapse, because cancer relapse cannot occur after death, and modeling death as a censoring event may misrepresent relapse-free survival.

- For disease-free survival, the hazard ratio associated with the occurrence of aGVHD is 1.72 (90% CI: 1.11-2.65), adjusting for covariates. In other words, comparing two patients of the same covariates, one who did develop aGHVD during the study period and another who did not, the instantaneous risk of death or relapse for the first patient is 1.72 greater than times the second, throughout the study period.
- For relapse-free survival, the hazard ratio associated with the occurrence of aGVHD is 0.83 (90% CI: 0.41-1.68), controlling for covariates related to patient/donor health and state of initial cancer.

Both confidence interval for the hazard ratio ranges above and below 1, therefore we cannot claim with certainty that the sample has sufficient evidence that aGVHD is associated with better or worse outcomes.

7. Based on the available data, is recovery of normal platelet levels associated with improved disease-free survival? Is it associated with a decreased risk of relapse?



We fit two Cox proportional hazards models, one for disease-free survival and another for relapse-free survival. In both models, our target is the hazard ratio associated with platelet recovery, which is modeled as a time-varying covariate. In the second PH model, death is a competing event with relapse, because cancer relapse cannot occur after death, and modeling death as a censoring event may misrepresent relapse-free survival.

- For disease-free survival, the ratio associated with platelet recovery is 0.42 (90% CI: 0.25-0.69), adjusting for covariates.
- For relapse-free survival, the ratio associated with platelet recovery is 1.11 (90% CI: 0.48-2.58), adjusting for covariates. In other words, comparing two patients of the same covariates, one whose platelet count returned to a normal level during the study period and another who did not, the instantaneous risk of relapse for the first patient is 1.11 times the second, throughout the study period.

We can claim with certainty that the sample has sufficient evidence that platelet recovery is positively associated with disease-free survival. The latter confidence interval for the hazard ratio ranges above and below 1, therefore we cannot claim with certainty that the sample has sufficient evidence that platelet is associated with better or worse relapse-free survival.

End of report. Code appendix begins on the next page.

Code Appendix

```
# Clear environment
rm(list=ls())

# Setup options
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, results='hide')
options(knitr.kable.NA = '-', digits = 2)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "allcode")]
# Load relevant packages
library(survival)  # survival models
library(dplyr)     # data manipulation
library(broom)     # combine and reshape model output
library(ggplot2)   # data visualization
library(gridExtra) # plot configuration

theme_set(theme_bw()) # set ggplot theme

# Load data
getwd()
bmt <- read.csv("../data/bmt_clean.csv")
dim(bmt) # 137 rows, 24 columns
names(bmt)

# Handle missing data (there is none)
anyNA(bmt)

## Helper functions

# Extract model results
get_robustci <- function (model, var=NULL, alpha=0.1) {
  # Organize model results
  tidy_model <- broom::tidy(model)
  # If desired, filter the results to a specific variable
  if (!is.null(var)) tidy_model = tidy_model %>% dplyr::filter(term==var)
  coef = tidy_model$estimate
  std.error = tidy_model$std.error
  robust.se = tidy_model$robust.se
  crit.val = qnorm(alpha/2, lower.tail=F)

  res <- list(
    term = tidy_model$term,
    # Exponentiate coefficient
    estimate = exp(coef),
    std.error = std.error,
    robust.se = robust.se,
    # Robust confidence interval
    conf.low = exp(coef - crit.val*robust.se),
    conf.high = exp(coef + crit.val*robust.se)
    )

  return(res)
```

```r
}

# Create a forest plot of HR estimates
create_HRforest <- function (model) {
  results <- tidy(model, conf.int=TRUE, conf.level=0.90, exponentiate=TRUE)
  gg <- ggplot(results, aes(x=estimate, y=term)) +
    geom_point(size=3) + # Plot hazard ratio points
    geom_errorbarh(aes(xmin=conf.low, xmax=conf.high), height=0.2) +  # CI bars
    geom_vline(xintercept=1, linetype="dashed", color="red") +  # Reference line
    labs(title = "Forest Plot from Cox PH Model",
         y="Variable", x="Hazard Ratio (HR)") +
    theme(legend.position="bottom")
  return(gg)
}

## Restructing dataframe to including indicators of competing risks and
## time-varing covariates
tbmt <- survival::tmerge(
  data1 = bmt, data2 = bmt, id = id,
  # Death
  death = event(ts, deltas),
  # Relapse
  relapse = event(tdfs, deltar),
  # Death or relapse
  death.relapse = event(tdfs, deltadfs),
  # Occurrence of aGVHD
  post.agvhd = tdc(ta),
  # Occurrence of platelet recovery
  post.recovery = tdc(tp)
  )

# Calculate survival times for distinct end points
death.relapse_surv <- with(tbmt, survival::Surv(tstart, tstop, death.relapse))
relapse_surv <- with(tbmt, Surv(tstart, tstop, relapse))

######################
#### DIRECTIVE 4 ####
######################

## Investigating acute graft-versus-host disease (aGVHD)
# HR of disease-free survival associated with the occurrence of aGVHD
formula1 <-
  # Association of interest
  death.relapse_surv ~ post.agvhd +
    # Patient and donor health
    age10_centered + donorage10_centered + cmv +
    # Initial state of cancer
    waittime90 + disgroup + fab +
    # Stratify by hospital and account for data structure
    + strata(hospital) + cluster(id)
coxfit1 <- survival::coxph(formula1, data=tbmt)
tidy(coxfit1, conf.int=TRUE, conf.level=0.90, exponentiate=TRUE)[-c(5,6)]
get_robustci(coxfit1, "post.agvhd")[-1] %>% sapply(round, 4)
```

```r
# HR of relapse-free survival associated with the occurrence of aGVHD
formula2 <-
  # Association of interest
  relapse_surv ~ post.agvhd +
    # Patient and donor health
    age10_centered + donorage10_centered + cmv + donorcmv +
    # Initial state of cancer
    waittime90 + disgroup + fab +
    # Stratify by hospital and account for data structure
    + strata(hospital) + cluster(id)
coxfit2 <- coxph(formula2, data=tbmt)
tidy(coxfit2, conf.int=TRUE, conf.level=0.90, exponentiate=TRUE)[-c(5,6)]
get_robustci(coxfit2, "post.agvhd")[-1] %>% sapply(round, 4)

# Forest plots of HR estimates
gridExtra::grid.arrange(
  create_HRforest(coxfit1) + labs(title="Disease-free survival"),
  create_HRforest(coxfit2) + labs(title="Relapse-free survival", y=""),
  nrow=1)

## Diagnostics

# Martingale residuals (form of wait time)
coxfit <- coxph(formula1, data=tbmt)
coxfit2 <- update(coxfit, . ~ . - waittime) # Remove wait time from predictors
mgresid1 <- residuals(coxfit, type="martingale")
mgresid2 <- residuals(coxfit2, type="martingale")
values <- tbmt$waittime

par(mfrow=c(1,2)) # set to print two plots side-by-side
plot(values, mgresid2,
     xlab="waittime", ylab="martingale residuals", # ylim=c(-1,1),
     main="waittime variable not included")
mgresid1.loess = loess(mgresid1 ~ values, degree=1)
lines(sort(values), predict(mgresid1.loess, sort(values)), col=2, lwd=2)
abline(h=0, lty=3)

plot(values, mgresid1,
     xlab="waittime", ylab="", main="waittime variable included")
mgresid1.loess = loess(mgresid1 ~ values, degree=1)
lines(sort(values), predict(mgresid1.loess, sort(values)), col=2, lwd=2)
abline(h=0, lty=3)
par(mfrow=c(1,1)) # reset to print a single plot

# Deviance residuals (identifying outliers)
coxfit <- coxph(formula1, data=tbmt)
devresid = residuals(coxfit, type="deviance")
plot(1:coxfit$n, devresid, xlab="observation", ylab="deviance residuals")
abline(h=0, lty=3)

# Schoenfeld residuals (assessing proportional hazards)
coxfit <- coxph(formula1, data=tbmt)
schoenresid = residuals(coxfit, type="scaledsch")
```

```r
times = as.numeric(rownames(schoenresid))

plot(times, schoenresid[,1], xlab="time", ylab="scaled Schoenfeld residuals")
schoenresid.loess = loess(schoenresid[,1] ~ times, degree=1)
lines(unique(times), predict(schoenresid.loess, unique(times)), col=2, lwd=2)
abline(h=0, lty=3)

# Assessing proportional hazards of a time-varying covariate
coxfit.tv = coxph(death.relapse_surv ~ tt(post.agvhd) + post.agvhd,
                  tt = function(x,t,...) x*t,
                  data=tbmt)
summary(coxfit.tv)$coef

######################
#### DIRECTIVE 7 ####
######################

## Investigating return of platelet counts to a normal level
# HR of disease-free survival associated with platelet recovery
formula3 <-
  # Association of interest
  death.relapse_surv ~ post.recovery +
    # Patient and donor health
    age10_centered + donorage10_centered + cmv + donorcmv +
    # Initial state of cancer
    waittime90 + disgroup + fab +
    # Stratify by hospital and account for data structure
    + strata(hospital) + cluster(id)
coxfit3 <- coxph(formula3, data=tbmt)
tidy(coxfit3, conf.int=TRUE, conf.level=0.90, exponentiate=TRUE)[-c(5,6)]
get_robustci(coxfit3, "post.recovery")[-1] %>% sapply(round, 4)

# HR of relapse-free survival associated with platelet recovery
formula4 <-
  # Association of interest
  relapse_surv ~ post.recovery  +
    # Patient and donor health
    age10_centered + donorage10_centered + cmv + donorcmv +
    # Initial state of cancer
    waittime90 + disgroup + fab +
    # Stratify by hospital and account for data structure
    + strata(hospital) + cluster(id)
coxfit4 <- coxph(formula4, data=tbmt)
tidy(coxfit4, conf.int=TRUE, conf.level=0.90, exponentiate=TRUE)[-c(5,6)]
get_robustci(coxfit4, "post.recovery")[-1] %>% sapply(round, 4)

# Forest plots of HR estimates
grid.arrange(
  create_HRforest(coxfit3) + labs(title="Disease-free survival"),
  create_HRforest(coxfit4) + labs(title="Relapse-free survival", y=""),
  nrow=1)

## Diagnostics
```

End of document.