MAE0514 - Prova 1 - Parte 2

Rubens Santos Andrade Filho[1]

Junho de 2021

Sumário

[1]Número USP: 10370336

Questão 2

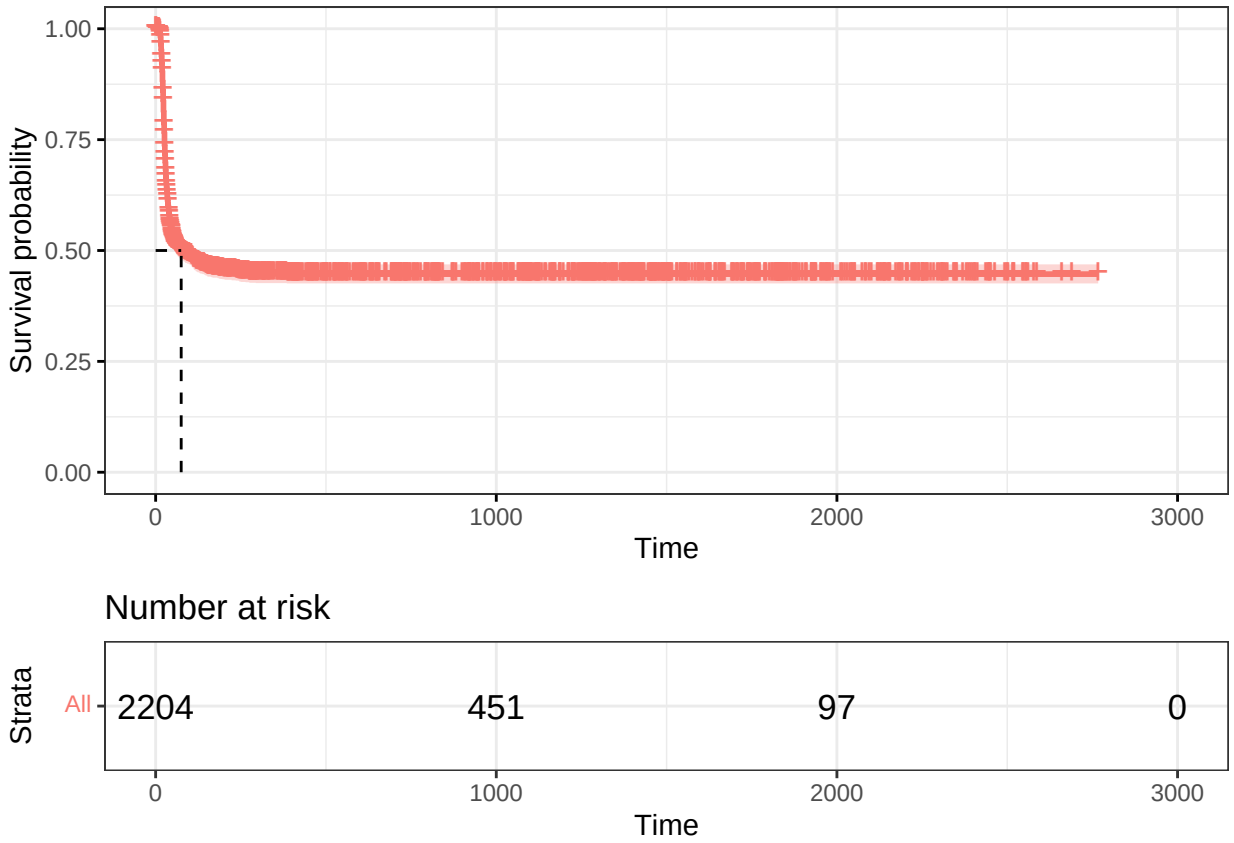
Foram considerados os dados provenientes do EBMT (*European Resgistry for Blood and Marrow Transplatation*), discutido em Putter, Fiocco Geskus (2007). Os dados são referentes a 2204 pacientes que reccberam transplante de medula óssea entre 1995 e 1998 reportados ao EBMT. Os dados estão disponíveis no arquivo **bmt3.csv** e a descrição em **bmt3.des**. As variáveis nos dados são:

- Tempo, em dias, após o transplante até recuperaço das plaquetas ou perda de acompanhamento
- Indicador de recuperação das plaquetas: 1, se recuperado; 0, se perda de acompanhamento
- Tempo livre de doença: tempo, em dias, após o transplante até óbito ou reincidência do câncer;
- Indicador de evento: 1, se óbito ou reincidência; 0 se censura
- Classificaçao da doença: leucemia linfoblástica aguda (ALL), leucemia mielóide aguda (AML) c leucemia mielóide crônica (CML)
- Idade, em categorias
- Correspondência de gênero doador-receptor
- Esgotamento de células T

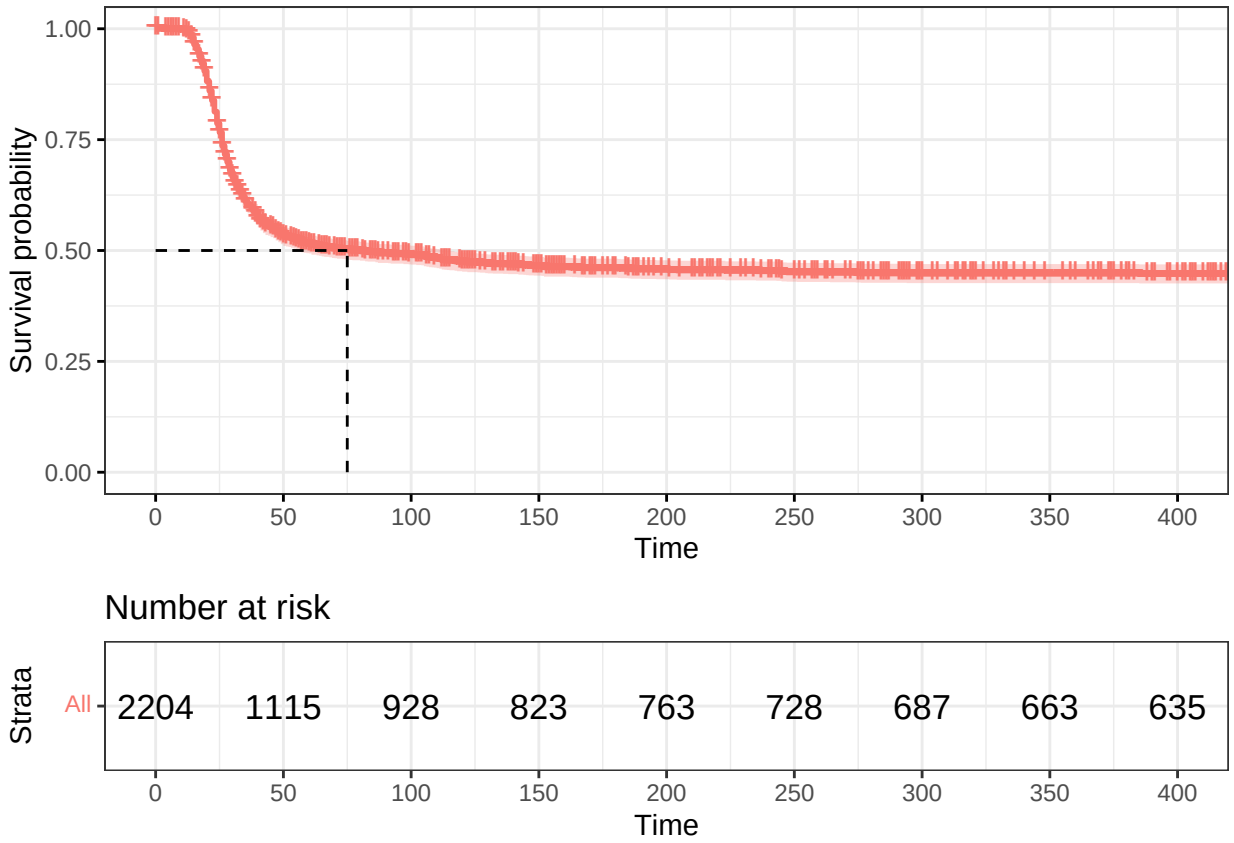
Os itens a seguir foram respondidos utilizando esses dados.

a) Análise descritiva, tempo de recuperação das plaquetas como desfecho

Nesse item, fazemos uma análise descritiva dos dados considerando como desfecho ou variável resposta o tempo até recuperação das plaquetas, `prtime`. Começamos fazendo o gráfico com a estimativa de Kaplan-Meier para a curva de sobrevivência do tempo até recuperação das plaquetas.



Vemos no gráfico que a estivativa da curva de sobrevivência decai rapídamente nos primeiros 300 dias para pouco menos de 50%. De fato, observando os dados, vemos que o maior tempo observado de recuperação das plaquetas ocorre no dia 385, com isso estimaviva da curva de sobrevivência não decai mais a partir desse dia. Com isso, para observar melhor o que acontece no início da curva, fazemos novamente o mesmo gráfico, porém, apenas até pouco após o dia 385.



Agora, podemos ver claramente, seguindo a linha pontilhada, que o tempo de sobrevivência mediano é de 75 dias. Isto é, 75 dias é a estimativa do tempo decorrido após o transplante em que a probabilidade de recuperação é 50%.

```
## Call: survfit(formula = s_plaq ~ 1, data = dados_raw)
##
##       n  events  median 0.95LCL 0.95UCL
##    2204    1169      75      56     112
```

Além disso, o número de observações é igual a 2204, número de recuperações observadas é 1169, e uma intervalar de 95% para o tempo mediano é $[56, 112]$ dias.

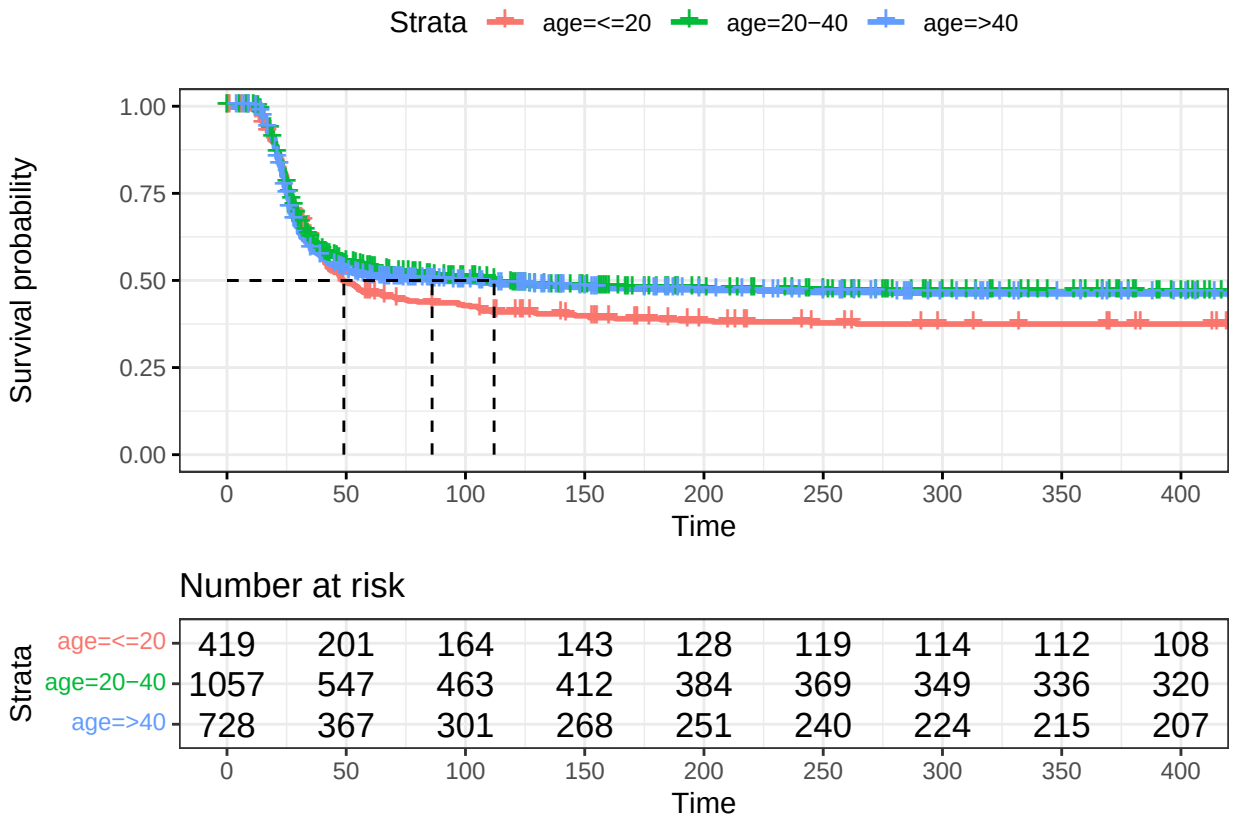
A seguir, iremos análisar comparar o tempo de recuperação das plaquetas entre os níveis de cada fator `age`, `dissub`, `drmatch` e `tcd`. É interessante também análisar o tempo mediano de recuperação das plaquetas para cada nível dos fatores. A tabela a seguir mostra as estimativas pontual e intervalar dos tempos medianos geral e em cada nível de cada fator.

Variável	**Tempo mediano**
Geral	75 (56, 112)
Idade	
<=20	49 (42, 71)
20-40	112 (64, 274)
>40	86 (48, -)
Subclassificação da doença	

Variável	Tempo mediano
AML	48 (40, 61)
ALL	59 (42, 105)
CML	- (144, -)
Correspondência de gênero	
No gender mismatch	86 (55, 142)
Gender mismatch	64 (47, 118)
Depleção de células T	
No TCD	112 (76, 223)
TCD	36 (33, 43)

Começando pela idade, `age`, construimos o gráfico com as estimativas de Kaplan-Meier para a curva de sobrevivência para cada categoria ("<=20", "20-40" e ">40"). Novamente, para melhor visualizar as curvas, iremos mostrar até o dia 400, uma vez que as estimaivas das curvas são constantes depois disso.

```
## Call: survfit(formula = s_plaq ~ age, data = dados_raw)
##
##               n events median 0.95LCL 0.95UCL
## age=<=20    419    253     49      42      71
## age=20-40  1057    540    112      64     274
## age=>40     728    376     86      48      NA
```

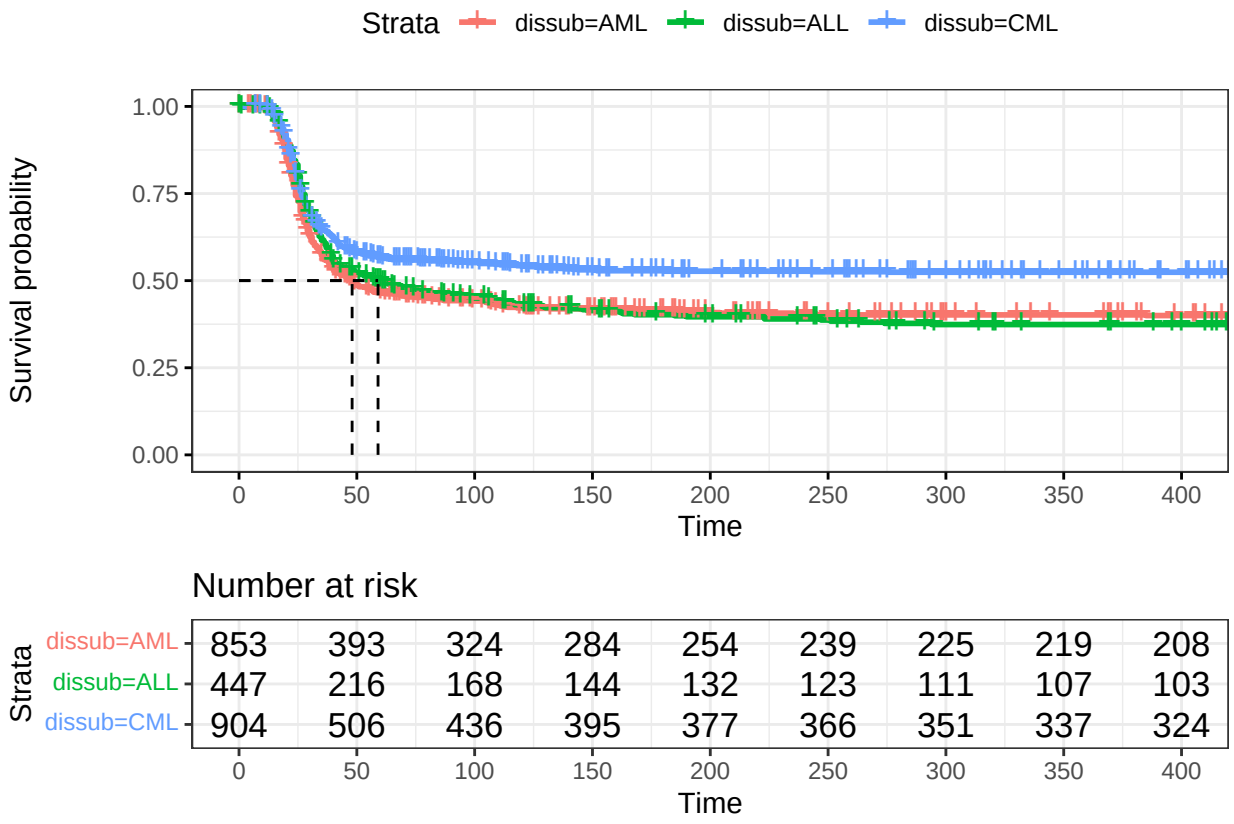


O gráfico mostra que os pacientes na faixa etária menor ou igual a 20 anos tem uma aparente menor estimativa da curva de sobrevivencia em relação às outras faixas etárias e em especial após o seu

tempo mediano. Antes do tempo mediano, é dificil ver indícios de diferenças entre as curvas. Além disso, as estimativas intervalares (com confiança de 95%) para o tempo mediano se sobrepõem, com isso, não podemos dizer que existe diferença significativa entre os tempos medianos a um nível de 5%. entretanto, mais a frente, testaremos formalmente isso.

Agora, fazemos a mesma análise com a covariável de Classificaçao da doença, `dissub`.

```
## Call: survfit(formula = s_plaq ~ dissub, data = dados)
##
##               n events median 0.95LCL 0.95UCL
## dissub=AML 853    490     48      40      61
## dissub=ALL 447    260     59      42     105
## dissub=CML 904    419     NA     144      NA
```

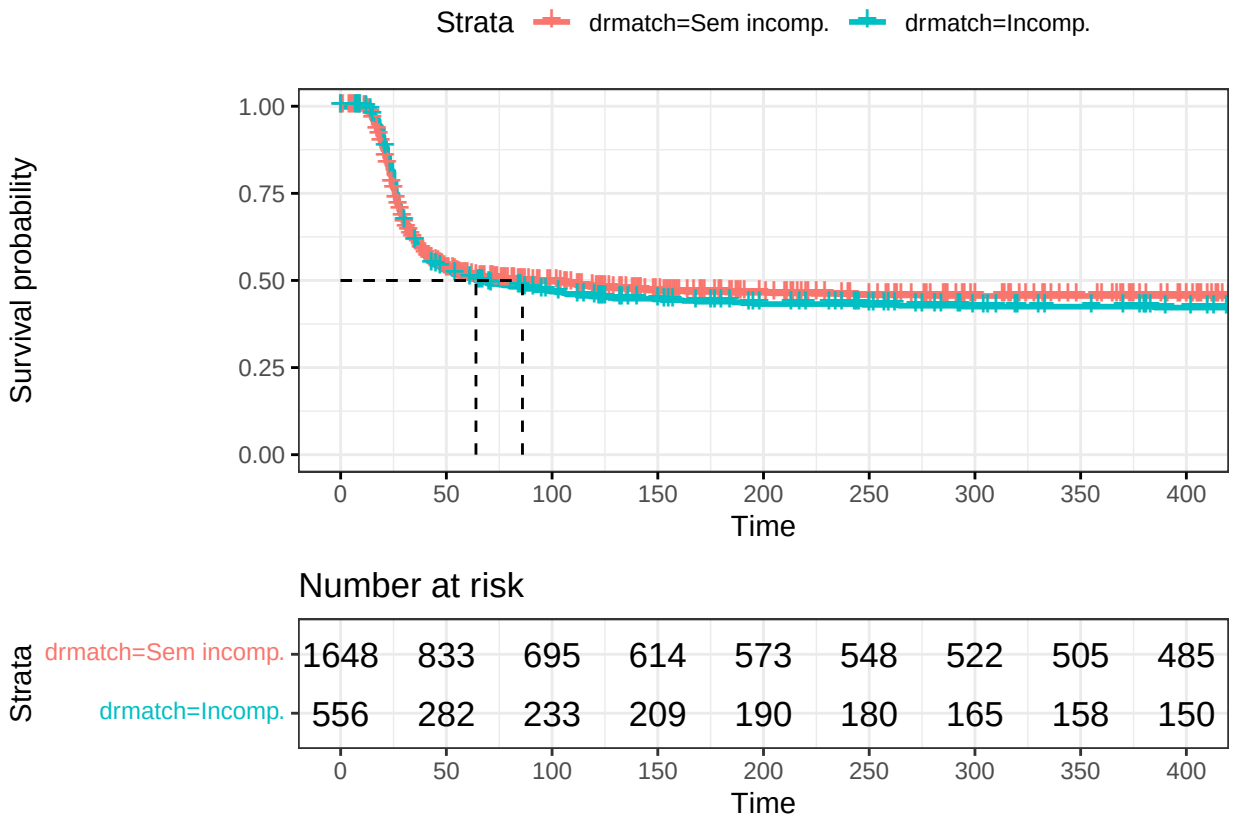


Não parece haver indícios de diferença entre as curvas de sobrevivência para as classificações de leucemia linfoblástica aguda (ALL) e leucemia mielóide aguda (AML). Entretanto, a curva para os pacientes com leucemia mielóide crônica (CML) apresentam uma aparente maior probabilidade de sobrevivência ao longo do tempo. Inclusive nem foi possível estimar o tempo mediano de sobrevivência para esse grupo. Já as estimativas do tempo mediano de sobrevivência, isto é, o tempo no qual metade dos pacientes já recuperaram as plaquetas, são bem próximas para os pacientes com leucemia linfoblástica aguda (ALL) e leucemia mielóide aguda (AML), 48 e 59 respecitamente. O resumo acima também mostra as estimativas intervalares com 955 de confiança.

Já para a covariável de Correspondência de gênero, `drmatch`, não parece haver diferenças entre as curvas de sobrevivência estimadas por Kaplan-Meier. As estimativas pontuais do tempo de sobrevivência

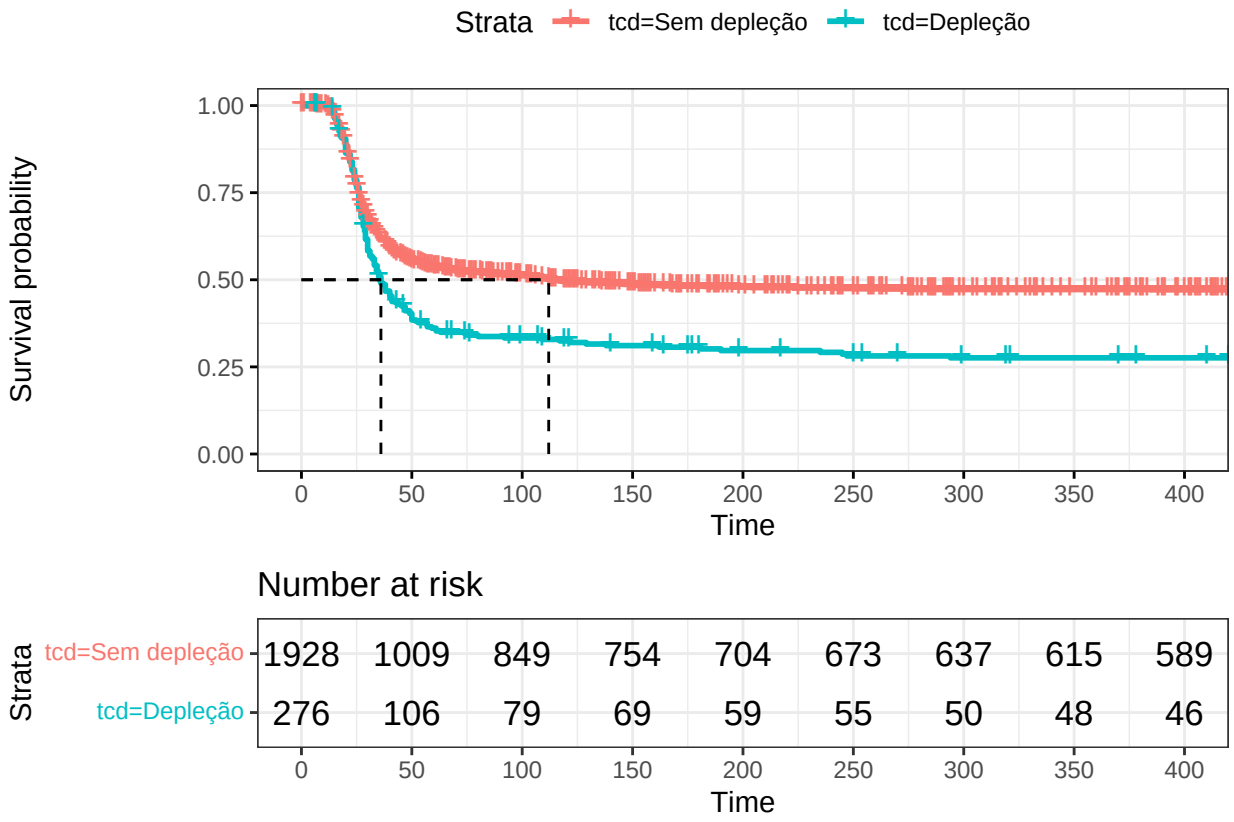
mediano também ficaram bem próximas, 86 e 64 dias para pacientes sem e com incompatibilidade de gênero, respectivamente. E com as estimativas intervalares se sobrepondo.

```
## Call: survfit(formula = s_plaq ~ drmatch, data = dados)
##
##                         n events median 0.95LCL 0.95UCL
## drmatch=Sem incomp. 1648    860     86      55     142
## drmatch=Incomp.      556    309     64      47     118
```



Por fim, quanto a Depleção de células T, `tcd`, observamos que existem indícios de diferença entre as curvas de sobrevivência estimadas. O pacietes com depleção de células T apresentam, ao longo do tempo, uma menor sobrevivência, iso é, um menor tempo de recuperação das plaquetas. Isso também é evidenciado pelo tempo mediano de sobrevivência: a estimativa do tempo até metade dos paciente sem depleção de células T recuperarem as plaquetas é de 112 dias, enquanto que nos pacientes com depleção, essa estimativa é de 36 dias.

```
## Call: survfit(formula = s_plaq ~ tcd, data = dados)
##
##                       n events median 0.95LCL 0.95UCL
## tcd=Sem depleção 1928    978    112      76     223
## tcd=Depleção      276    191     36      33      43
```



b) Testes de log-rank, tempo de recuperação das plaquetas como desfecho

Fizemos os testes de log-rank comparando as categorias das covariáveis no estudo. No teste de log rank, a hipótese nula é que as funções de sobrevivência em cada nível de um fator são iguais em todos os tempos. A hipótese alternativa é de que pelo menos uma função é diferente para algum tempo dentro de um intervalo de zero a um tempo razoável estabelecido.

```
## Call:
## survdiff(formula = s_plaq ~ age, data = dados_raw)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## age=<=20    419      253      221    4.5047    5.6627
## age=20-40  1057      540      569    1.4323    2.8400
## age=>40     728      376      379    0.0245    0.0369
##
##  Chisq= 6.1  on 2 degrees of freedom, p= 0.05


## valor p exato
##    0.04797336


## Call:
## survdiff(formula = s_plaq ~ dissub, data = dados_raw)
##
```

```
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## dissub=AML 853      490      428      9.10     14.62
## dissub=ALL 447      260      235      2.64      3.37
## dissub=CML 904      419      506     15.05     27.08
##
##  Chisq= 27.3  on 2 degrees of freedom, p= 1e-06
```

```
## Call:
## survdiff(formula = s_plaq ~ drmatch, data = dados_raw)
##
##                               N Observed Expected (O-E)^2/E (O-E)^2/V
## drmatch=No gender mismatch 1648      860      870     0.120      0.48
## drmatch=Gender mismatch     556      309      299     0.351      0.48
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

```
## Call:
## survdiff(formula = s_plaq ~ tcd, data = dados_raw)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## tcd=No TCD 1928      978     1037      3.32      29.9
## tcd=TCD     276      191      132     25.97      29.9
##
##  Chisq= 29.9  on 1 degrees of freedom, p= 5e-08
```

A um nível de significância de 5%, rejeitamos a hipótese de nula que as curvas de sobrevivência são iguais em todos os tempos, isto é, existe diferença entre os tempos até a recuperação das plaquetas entre os níveis das covariáveis Idade (valor p = 0.048), Classificação da doença (valor p = 1e-06) e Depleção de células T (valor p = 5e-08). Observamos que para a idade essa diferença é discutível e ainda mais se notarmos que as curvas para as diferentes faixas etárias aparentam se cruzarem. Além disso, não rejeitamos a hipótese nula para os níveis da covariável de Correspondência de gênero, com os valores observados bem próximos dos esperados.

c) Análise descritiva, tempo livre da doença como desfecho

De maneira similar ao item a), fazemos os gráficos com as estimativas de Kaplan-Meier e estimativas pontuais e intervalares do tempo de sobrevivencia mediano geral e para os diferentes níveis de cada covariável.

d) Testes de log-rank, tempo livre da doença como desfecho

(d) Faça testes de log-rank comparando as categorias das covariáveis no estudo. Considere também a inclusão de ponderaçōes nesse caso. Discuta os resultados.

e) Modelo Weibull, tempo livre da doença como desfecho

(e) Considerando o tempo livre da doença, ajuste um modelo Weibull aos dados. Apresente os resultados do modelo completo, com todas as covariáveis incluídas. Faça um processo de seleça de variáveis utilizando o teste da razão de verossimilhanças e apresente o resultado do modelo final obtido. Você

precisa descrever claramente o processo de seleção das variáveis adotado, mas deve apresentar apenas as estimativas e resultados de dois modelos: modelo completo e modelo final. Você pode apresentar os resultados do modelo na parametrização de locação-escala.

f) Interpretação do modelo final em (e).

(f) Interprete os parâmetros do modelo final obtido em (e).

g) Modelo log-logístico, tempo livre da doença como desfecho

(g) De forma semelhante ao item (e), ajuste um modelo log-logístico aos dados. Faça da mesma forma (porém utilizando a distribuição log-logística) e apresente os resultados do modelo completo e do modelo final.

h) Interpretação do modelo final em (f).

(h) Interprete os parâmetros do modelo final obtido em (g).

Importante: Em todos os itens, os resultados apresentados devem ser interpretados. A redação será também avaliada. Não se esqueça de apresentar códigos dos programas utilizados.

Código Completo

```
library(knitr)
library(tidyverse)
library(dplyr)
library(readr)
library(ggplot2)
library(survival)
library(survminer)
library(gtsummary)

knitr::opts_chunk$set(warning=FALSE,
                      # fig.dim = c(5,5),
                      # out.height = '40%',
                      # fig.align = 'center',
                      message=FALSE,
                      echo=FALSE
                      )
# QUESTAO 2 ----

dados_raw <- readr::read_csv(
  'ebmt3.csv',
  col_types = readr::cols_only(
    id = col_integer(),
```

```
    prtime = col_double(),
    prstat = col_integer(),
    rfstime = col_double(),
    rfsstat = col_integer(),
    dissub = col_factor(c("AML", "ALL", "CML")),
    age = col_factor(levels = c("<=20", "20-40", ">40"), ordered = T),
    drmatch = col_factor(c("No gender mismatch", "Gender mismatch")),
    tcd = col_factor(c("No TCD", "TCD"))
  )
)

labels <- list(
  id="Identificação do paciente",
  prtime="Tempo de recuperação das plaquetas",
  prstat="Indicador de recuperação das plaquetas",
  rfstime="Tempo livre de doença",
  rfsstat="Indicador de evento",
  dissub="Subclassificação da doença",
  age="Idade",
  drmatch="Correspondência de gênero",
  tcd="Depleção de células T"
)

labelled::var_label(dados_raw) <- labels

dados_raw %>% str

# traduz fatores
dados <- dados_raw %>%
  mutate(
    #dissub = factor(dissub, c("AML", "ALL", "CML")),
    # age = forcats::fct_recode(
    #   age, "Até 20"="<=20", "Entre 20 e 40"="20-40", "Mais que 40"=">40"),
    drmatch = forcats::fct_recode(
      drmatch, "Sem incomp."="No gender mismatch",
      "Incomp."="Gender mismatch"),
    tcd = forcats::fct_recode(tcd, "Sem depleção"="No TCD", "Depleção"="TCD")
  )
dados

# QUESTÃO 2.a) geral ----

# ajuste
s_plaq <- with(dados_raw,Surv(prtime, prstat))
fit <- survfit(s_plaq~1, dados_raw)

# grafico das estimativas
ggsurvplot(
  fit,
  surv.median.line = "hv",
  ggtheme = theme_bw(),
  risk.table = T, tables.height = 0.32,
  legend='none'
```

```
)
ggsurvplot(
  fit,
  surv.median.line = "hv",
  ggtheme = theme_bw(),
  xlim=c(0,400),
  risk.table = T, tables.height = 0.32,
  break.time.by = 50,
  legend='none'
)

print(fit)

fits <- list(
  survfit(s_plaq ~ 1, dados_raw),
  survfit(s_plaq ~ age, dados_raw),
  survfit(s_plaq ~ dissub, dados_raw),
  survfit(s_plaq ~ drmatch, dados_raw),
  survfit(s_plaq ~ tcd, dados_raw)
)


fits %>%
  tbl_survfit(
  probs = 0.5,
  label_header = "**Tempo mediano**",
  missing ="-",
  label = list(
    1 ~ "Geral"
  )
) %>%
  #bold_labels() %>%
  italicize_levels() %>%
  modify_header(
    update = list(
      label ~ "**Variável**"
    )
  )

# QUESTÃO 2.a) age ----
# ajuste
fit <- survfit(s_plaq~age, dados_raw)

# tabela com medidas resumo
print(fit)

# grafico das estimativas
ggsurvplot(
  fit,
  surv.median.line = "hv",
  ggtheme = theme_bw(),
  xlim=c(0,400),
  risk.table = T, tables.height = 0.32,
```

```
  break.time.by = 50,
  legend='top'
)

# QUESTÃO 2.a) dissub ----
# ajuste
fit <- survfit(s_plaq~dissub, dados)

# tabela com medidas resumo
print(fit)

# grafico das estimativas
ggsurvplot(
  fit,
  surv.median.line = "hv",
  ggtheme = theme_bw(),
  xlim=c(0,400),
  risk.table = T, tables.height = 0.32,
  break.time.by = 50,
  legend='top'
)
# QUESTÃO 2.a) drmatch ----
# ajuste
fit <- survfit(s_plaq~drmatch, dados)

# tabela com medidas resumo
print(fit)

# grafico das estimativas
ggsurvplot(
  fit,
  surv.median.line = "hv",
  ggtheme = theme_bw(),
  xlim=c(0,400),
  risk.table = T, tables.height = 0.32,
  break.time.by = 50,
  legend='top'
)

# QUESTÃO 2.a) tcd ----
# ajuste
fit <- survfit(s_plaq~tcd, dados)

# tabela com medidas resumo
print(fit)

# grafico das estimativas
ggsurvplot(
  fit,
  surv.median.line = "hv",
  ggtheme = theme_bw(),
  xlim=c(0,400),
  risk.table = T, tables.height = 0.32,
```

```
  break.time.by = 50,
  legend='top'
)

(sdiff <- survdiff(s_plaq ~ age, dados_raw))
c("valor p exato" = pchisq(sdiff$chisq, df = 2, lower.tail = F))
survdiff(s_plaq ~ dissub, dados_raw)
survdiff(s_plaq ~ drmatch, dados_raw)
survdiff(s_plaq ~ tcd, dados_raw)
```