SA_group_assignment

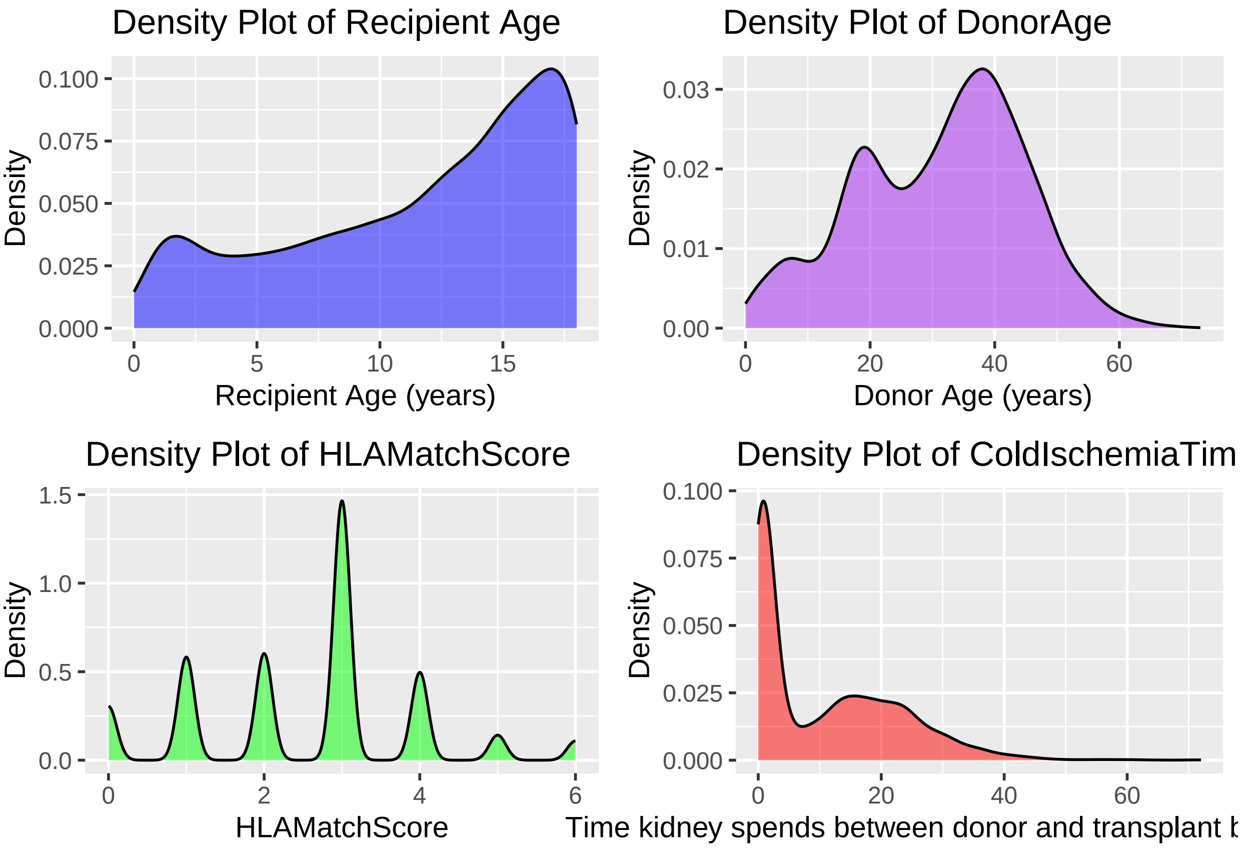
Niels van der Drift, Mikdad Kanbar, Severin Holtmann

2023-04-30

Exercise 1: Data Inspection

We will investigate a table with summary statistics and make some plots:

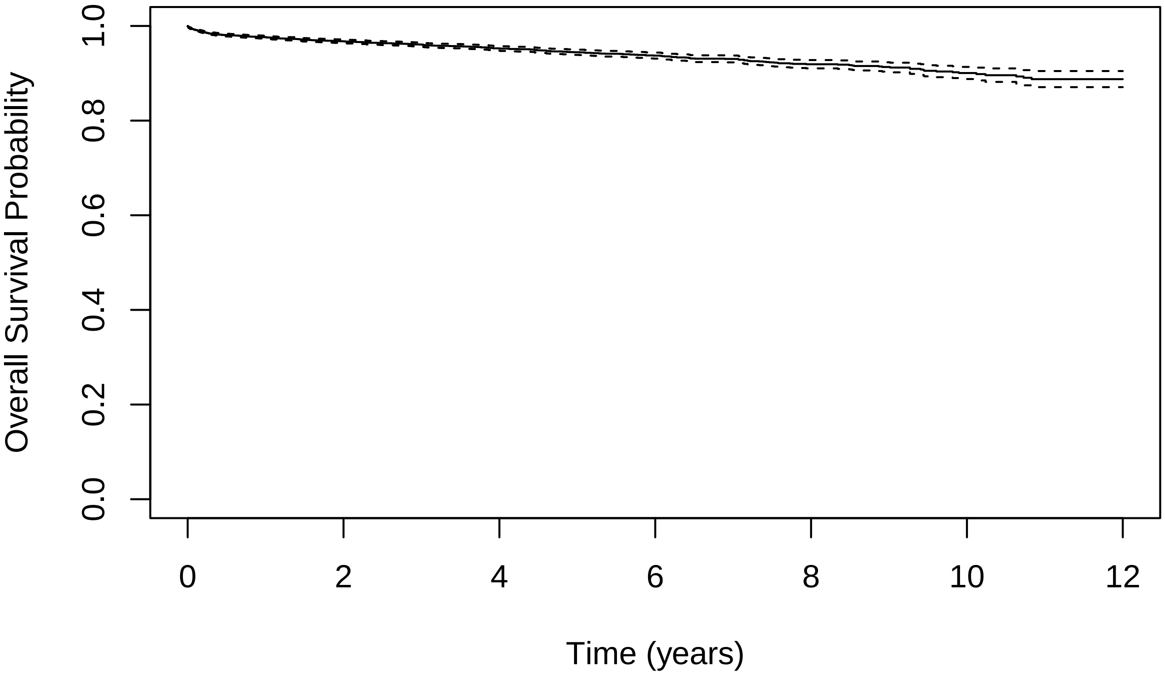
hlamat	age	age.1	cold_isc	death	year	sex	txtype	fu
Min. :0.000	Min. : 0.0	Min. : 0.00	Min. : 0.00	Min. :0.00000	Min. :1990	Min. :0.0000	Min. :0.0000	Min. : 0.000
1st Qu.:2.000	1st Qu.:21.0	1st Qu.: 8.00	1st Qu.: 1.00	1st Qu.:0.00000	1st Qu.:1993	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.: 1.096
Median :3.000	Median :33.0	Median :13.00	Median : 7.00	Median :0.00000	Median :1996	Median :1.0000	Median :0.0000	Median : 3.115
Mean :2.574	Mean :31.3	Mean :11.65	Mean :10.86	Mean :0.04757	Mean :1996	Mean :0.5894	Mean :0.4734	Mean : 3.888
3rd Qu.:3.000	3rd Qu.:41.0	3rd Qu.:16.00	3rd Qu.:19.00	3rd Qu.:0.00000	3rd Qu.:1999	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.: 5.978
Max. :6.000	Max. :73.0	Max. :18.00	Max. :72.00	Max. :1.00000	Max. :2002	Max. :1.0000	Max. :1.0000	Max. :12.532
NA's :234	NA's :113	NA's :9	NA's :2250	NA	NA	NA	NA	NA



We observe that the recipient age is most frequent between ages 10 and 18, while the donor age is mostly populated between 18 and 45 years old. HLAMatch Score has mode 3 and is right skewed, most of the patients have a HLAMatch of 3 or lower. ColdIschemiaTime is right skewed, most of its mass lies between 0 and 5 hours, this is expected as it makes sense minimize the time a kidney spends outside of the donor and recipient body.

Exercise 2: Plot Overall Kaplan-Meier

Next, we will plot a Kaplan-Meier curve for overall survival during the first 12 years after transplantation.



We observe good survival rates for the kidney transplants. Roughly 90% is still alive after 12 years.

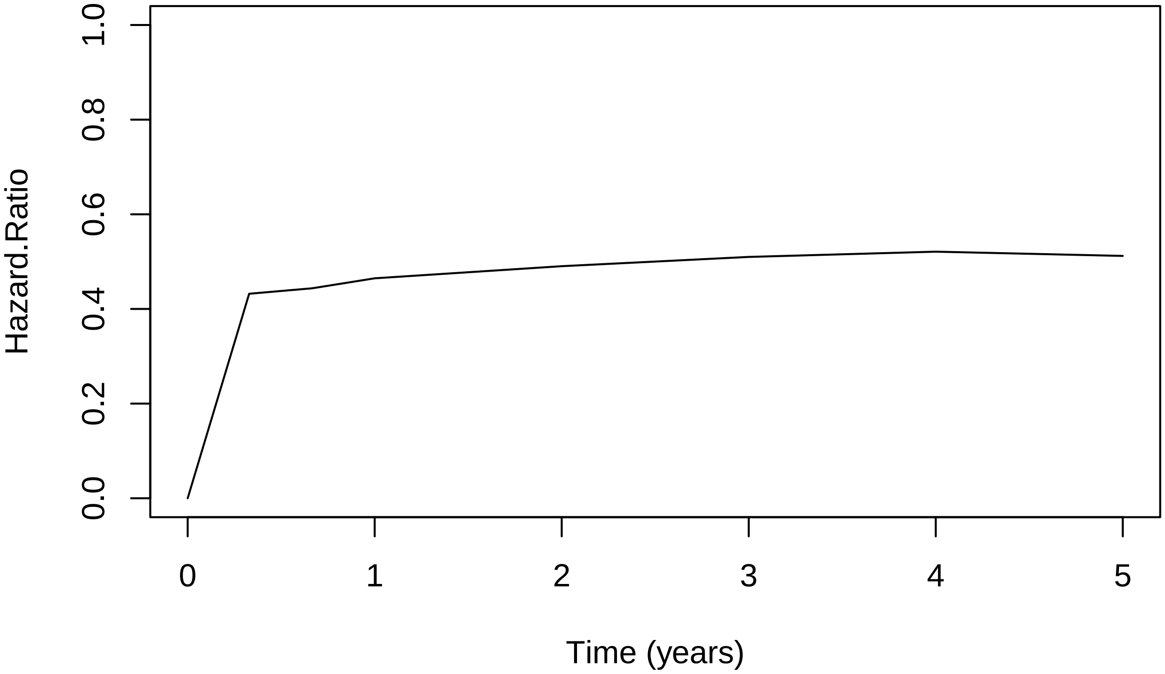
Exercise 3: Compare Mortality Rates

Next, we are going to compare mortality rates (hazard functions) between children whose transplanted kidney was provided by a living donor and those whose source was recently deceased, as is specified by the dummy variable *txtype*. The information will be provided in the following Kaplan-Meier tables.

	Time	Occurrences	People at Risk	Hazard	1 - Hazard	Survival	Mortality	cumHaz
1	0.0000000	0	5148	0	1	1.0000000	0.0000000	0.0000000
99	0.3287671	0	4637	0	1	0.9898905	0.0101095	0.0101592
190	0.6657534	0	4374	0	1	0.9870652	0.0129348	0.0130170
286	1.0000000	0	4056	0	1	0.9845543	0.0154457	0.0155638
515	2.0000000	0	3428	0	1	0.9782992	0.0217008	0.0219362
745	3.0000000	0	2829	0	1	0.9730770	0.0269230	0.0272876
973	4.0000000	0	2297	0	1	0.9669280	0.0330720	0.0336255
1206	5.0000000	0	1789	0	1	0.9616331	0.0383669	0.0391150

	Time	Occurrences	People at Risk	Hazard	1 - Hazard	Survival	Mortality	cumHaz
1	0.0000000	4	4627	0.0008645	0.9991355	0.9991355	0.0008645	0.0008645
106	0.3287671	0	4158	0.0000000	1.0000000	0.9766033	0.0233967	0.0236692
203	0.6657534	0	3897	0.0000000	1.0000000	0.9708459	0.0291541	0.0295813
305	1.0000000	2	3630	0.0005510	0.9994490	0.9667654	0.0332346	0.0337926
562	2.0000000	0	3023	0.0000000	1.0000000	0.9557460	0.0442540	0.0452542
829	3.0000000	0	2419	0.0000000	1.0000000	0.9472001	0.0527999	0.0542342
1094	4.0000000	0	1853	0.0000000	1.0000000	0.9365245	0.0634755	0.0655662
1343	5.0000000	0	1408	0.0000000	1.0000000	0.9250764	0.0749236	0.0778614

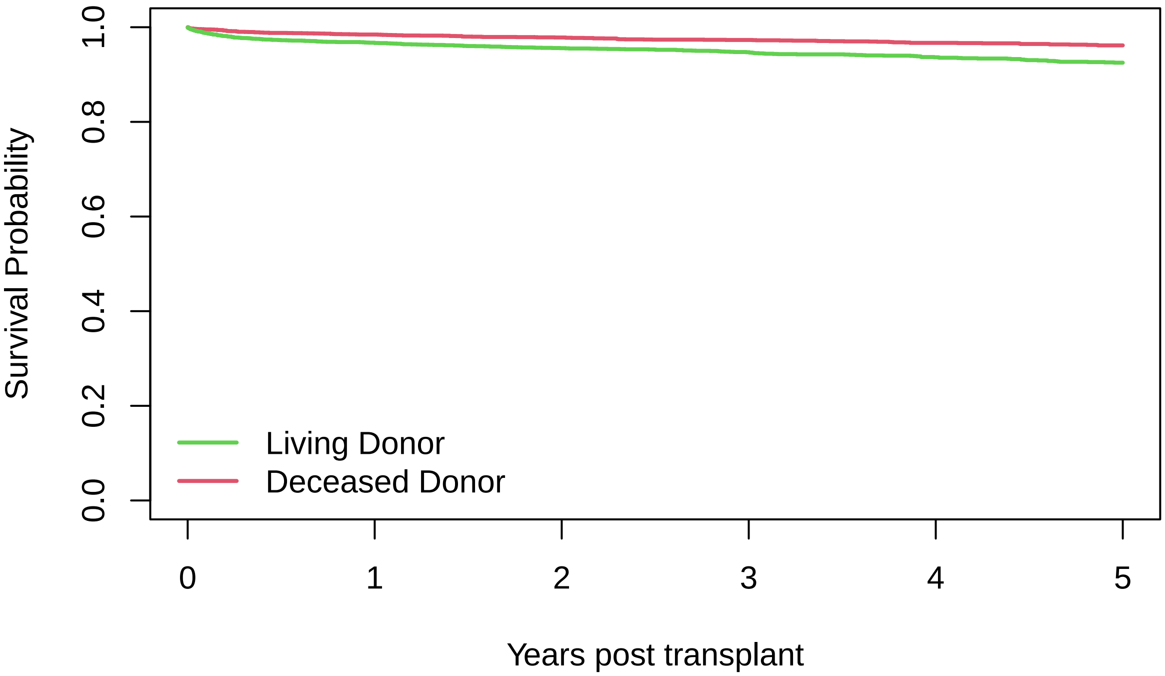
We can observe that the survival rate of patients with a living donor is constantly lower, than that of patients with a dead donor. We can also see that the disparity becomes slightly larger over time, which is also visualised in the following plot of the ratio between the two mortality rates.



As we can see, the ratio is constantly at around 45% - 50%. This suggests that the risk of death is about twice as high if the donor is still alive. This tendency will become more prevalent during the following tests. It does seem a bit counter-intuitive to us, as one would think that a kidney from a live donor is better.

Exercise 4: Plot Kaplan-Meier curves for both donor types

Now, we will present a plot with Kaplan-Meier survival curves for the two donor types.



From this, we may gather similar information as from the table and the hazard ratio plot; Living donor patients tend to survive less.

Exercise 5: Fit Cox model with donor type as predictor

To formally investigate the difference between donor types, we will make a Cox proportional hazards regression, dependent on donor type.

```
## Call:
## coxph(formula = Surv(fu, death) ~ txtype, data = unos_data)
##
##   n= 9775, number of events= 465
##
##            coef exp(coef) se(coef)     z Pr(>|z|)
## txtype 0.64469   1.90539  0.09558 6.745 1.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## txtype     1.905     0.5248      1.58     2.298
##
## Concordance= 0.586  (se = 0.012 )
## Likelihood ratio test= 47.1  on 1 df,   p=7e-12
## Wald test            = 45.5  on 1 df,   p=2e-11
## Score (logrank) test = 47.09  on 1 df,   p=7e-12
```

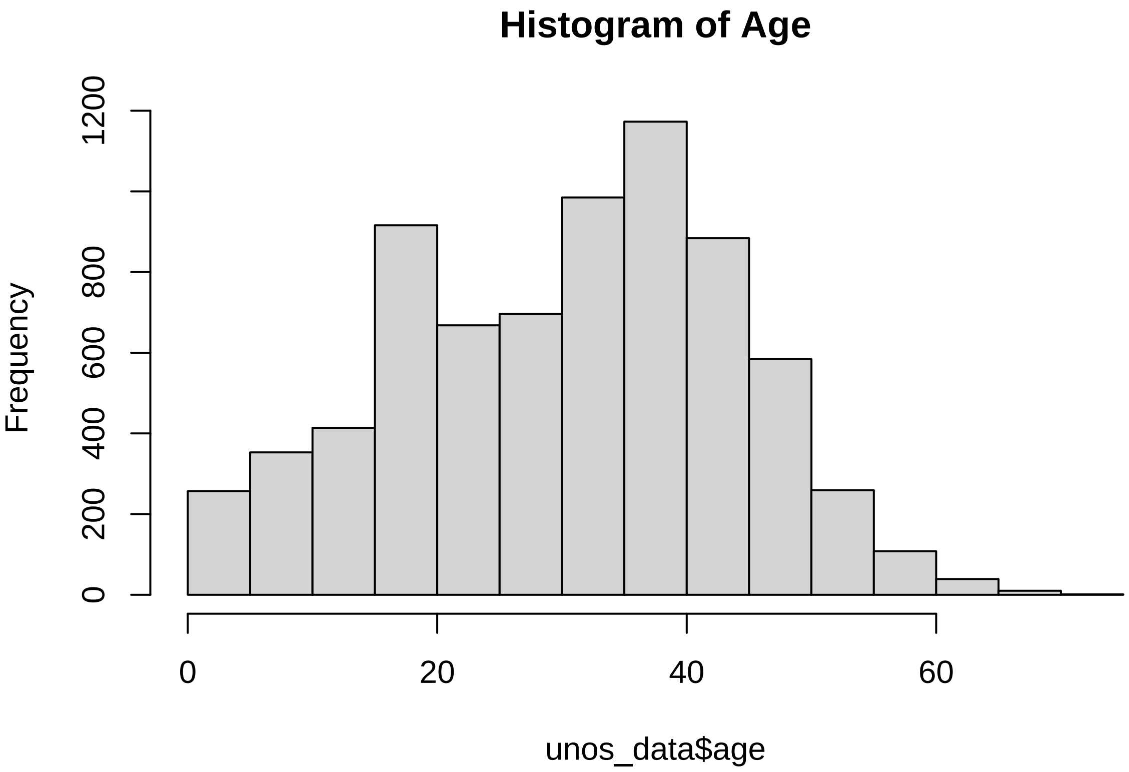
The exp(coef) value of 1.90539 represents the hazard ratio for deceased donor kidney recipients compared to living donor kidney recipients. This means that the hazard of mortality is 1.91 times higher for living donor kidney recipients compared to deceased donor kidney recipients.

The confidence interval for the hazard ratio is between 1.58 and 2.298, indicating that we can be 95% confident that the true hazard ratio lies within this range.

Exercise 6: Fit a Cox model with age as predictor and estimate the hazard ratio and its confidence interval

Research shows that an important determinant of mortality after kidney transplant is the age of the recipient. Fit a Cox model with age as predictor and estimate the hazard ratio and its confidence interval. First, we will consider age as a continuous variable, and thereafter divide it into categories.

Numerical Age



```
## Call:
## coxph(formula = Surv(fu, death) ~ age.1, data = unos_data)
##
##   n= 7347, number of events= 351
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## age.1 -0.01822   0.98194  0.01002 -1.819   0.0689 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## age.1    0.9819      1.018    0.9628     1.001
##
## Concordance= 0.543  (se = 0.018 )
## Likelihood ratio test= 3.26  on 1 df,   p=0.07
## Wald test            = 3.31  on 1 df,   p=0.07
## Score (logrank) test = 3.32  on 1 df,   p=0.07
```

Categorical Age

```
## Call:
## coxph(formula = Surv(fu, death) ~ age_cat, data = unos_data)
##
##   n= 7303, number of events= 348
##    (44 observations deleted due to missingness)
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## age_cat(10,15]  -0.1336    0.8750   0.1294 -1.032    0.302
## age_cat(15,Inf] -0.1093    0.8964   0.1303 -0.839    0.401
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## age_cat(10,15]     0.8750      1.143    0.6789     1.128
## age_cat(15,Inf]    0.8964      1.116    0.6944     1.157
##
## Concordance= 0.53  (se = 0.016 )
## Likelihood ratio test= 1.25  on 2 df,   p=0.5
## Wald test            = 1.26  on 2 df,   p=0.5
## Score (logrank) test = 1.26  on 2 df,   p=0.5
```

- Hazard Ratio for continuous age: 0.9785
- CI: (0.9719, 0.9852)
- Hazard Ratio's for age categories:
- 10-15: 0.8750 CI(0.411, 0.979)
- 12-18: 0.3525 CI(0.22)
- 18-30: 0.3385
- 30+: 0.3309

Covariate	Hazard Ratio	95% lower bound	95% upper bound
Continuous Age	0.9785	0.9719	0.9852
Age 10-15	0.8750	0.6789	1.1280
Age 15+	0.8964	0.6944	1.1570

We can observe that the bounds for continuous age are very narrow, making the estimate very precise. For categorical age on the other hand, this is not at all the case. The bounds are quite similar and have a lot of spread. 1 is also within the bounds, implying that the coefficients are insignificant. It is thus better to stick with age as a continuous variable.

Exercise 7: Fit a multivariate Cox model by using other predictors and describe your results.

We will start off by making a full model of all relevant variables, including an interaction term between donor and patient age.

```
## Call:
## coxph(formula = Surv(fu, death) ~ cold_isc + sex + txtype + hlamat +
##     age.1 * age, data = unos_data, method = "breslow")
##
##   n= 7347, number of events= 351
##    (2428 observations deleted due to missingness)
```

```
##
##                 coef  exp(coef)   se(coef)      z Pr(>|z|)
## cold_isc   0.0047135  1.0047247  0.0067681  0.696  0.48616
## sex       -0.0806924  0.9224774  0.1086716 -0.743  0.45776
## txtype     0.3471755  1.4150650  0.1891947  1.835  0.06650 .
## hlamat    -0.0612583  0.9405802  0.0443540 -1.381  0.16724
## age.1     -0.0666207  0.9355500  0.0222801 -2.990  0.00279 **
## age       -0.0224923  0.9777588  0.0096357 -2.334  0.01958 *
## age.1:age  0.0016306  1.0016319  0.0007316  2.229  0.02583 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## cold_isc     1.0047     0.9953    0.9915    1.0181
## sex          0.9225     1.0840    0.7455    1.1415
## txtype       1.4151     0.7067    0.9766    2.0503
## hlamat       0.9406     1.0632    0.8623    1.0260
## age.1        0.9355     1.0689    0.8956    0.9773
## age          0.9778     1.0227    0.9595    0.9964
## age.1:age    1.0016     0.9984    1.0002    1.0031
##
## Concordance= 0.604  (se = 0.017 )
## Likelihood ratio test= 39.47  on 7 df,   p=2e-06
## Wald test            = 41.27  on 7 df,   p=7e-07
## Score (logrank) test = 42.35  on 7 df,   p=4e-07
```

We can see that all coefficients besides age are insignificant. We will start reducing the model by removing the most insignicant coefficient - cold_isc.

```
## Call:
## coxph(formula = Surv(fu, death) ~ sex + txtype + hlamat + age.1 *
##     age, data = unos_data, method = "breslow")
##
##   n= 9433, number of events= 449
##    (342 observations deleted due to missingness)
##
##                 coef  exp(coef)   se(coef)      z Pr(>|z|)
## sex       -0.0999397  0.9048920  0.0959073 -1.042  0.29739
## txtype     0.3773892  1.4584718  0.1222194  3.088  0.00202 **
## hlamat    -0.0915721  0.9124955  0.0402870 -2.273  0.02303 *
## age.1     -0.1313814  0.8768832  0.0178443 -7.363 1.80e-13 ***
## age       -0.0453338  0.9556784  0.0077646 -5.838 5.27e-09 ***
## age.1:age  0.0033324  1.0033380  0.0005994  5.560 2.70e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## sex          0.9049     1.1051    0.7498    1.0920
## txtype       1.4585     0.6856    1.1478    1.8532
## hlamat       0.9125     1.0959    0.8432    0.9875
## age.1        0.8769     1.1404    0.8467    0.9081
## age          0.9557     1.0464    0.9412    0.9703
## age.1:age    1.0033     0.9967    1.0022    1.0045
##
## Concordance= 0.625  (se = 0.016 )
## Likelihood ratio test= 116.3  on 6 df,   p=<2e-16
## Wald test            = 140.7  on 6 df,   p=<2e-16
## Score (logrank) test = 147.8  on 6 df,   p=<2e-16
```

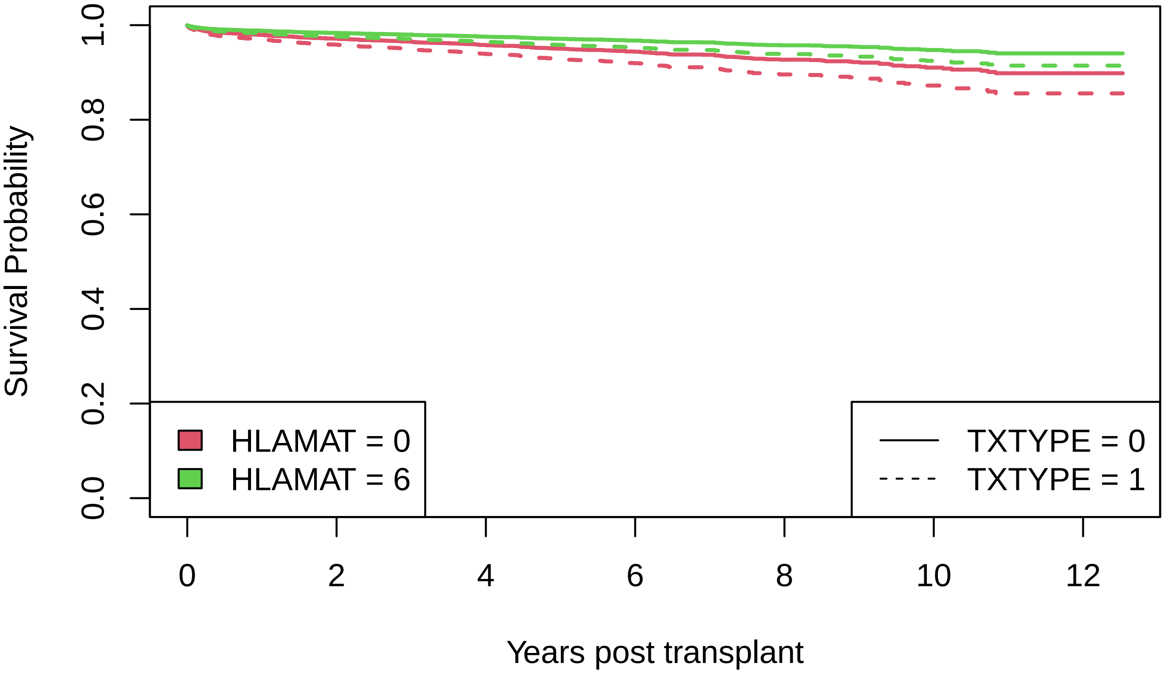
Having dropped cold_isc, we can see that more covariates have become significant. Nevertheles, sex is still insignificant so we will remove this too.

```
## Call:
## coxph(formula = Surv(fu, death) ~ txtype + hlamat + age.1 * age,
##     data = unos_data, method = "breslow")
##
##   n= 9433, number of events= 449
##    (342 observations deleted due to missingness)
##
##                  coef  exp(coef)   se(coef)      z Pr(>|z|)
## txtype      0.3735962  1.4529503  0.1221736  3.058  0.00223 **
## hlamat     -0.0924870  0.9116611  0.0402858 -2.296  0.02169 *
## age.1      -0.1303368  0.8777997  0.0178149 -7.316 2.55e-13 ***
## age        -0.0453928  0.9556221  0.0077665 -5.845 5.08e-09 ***
## age.1:age   0.0033232  1.0033287  0.0005993  5.545 2.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## txtype        1.4530     0.6883    1.1436    1.8461
## hlamat        0.9117     1.0969    0.8424    0.9866
## age.1         0.8778     1.1392    0.8477    0.9090
## age           0.9556     1.0464    0.9412    0.9703
## age.1:age     1.0033     0.9967    1.0022    1.0045
##
## Concordance= 0.628  (se = 0.016 )
## Likelihood ratio test= 115.2  on 5 df,   p=<2e-16
## Wald test            = 139.5  on 5 df,   p=<2e-16
## Score (logrank) test = 146.6  on 5 df,   p=<2e-16
```

We now have a nice condensed model with only significant coefficients. As was already previously conveyed, having a living donor raises the hazard rate by ~50% (we are still unsure if this might be due to a variable coding mistake). The negative coefficient of hlamat implies that having a better donor-patient match reduces the hazard rate. This makes intuitive sense. Both age and donor age have an exponential coefficient below one, stating that being older and having an older donor is beneficial. This should however be considered in conjunction with the interaction, which does have a hazard increasing coefficient. The interaction aims to uncover a non-linear effect that becomes particularly prevalent when either, or especially both, donor and patient are very old. In most extreme cases, this could also highlight that a big disparity between patient and donor age becomes a risk as well. We also attempted to model this property directly by accounting for the absolute difference between patient and donor age, but this did not yield significant results. It should be noted that the coefficient is much small than the normal age coefficients as the multiplicative property of the interaction can easily lead to values in the hundreds.

Exercise 8: Estimate the survival function for specific covariate patterns

Based on the previous results, we have chosen the final model to use *txtype, hlamat, age.1, age, age.1:age*. Using this model, we will now evaluate the survival rate of some exemplary cases. Specifically, we will investigate the extreme cases of donor matching [0, 6] for both living donors and dead donors. Preemptively, we would expect that people with a better match [6] have a higher survival rate. As previously discovered, a living donor increases the risk of death. We expect to observe the same pattern here. To simulate these patients, the median ages for patient age [13], and for donor age [33] were used.



Indeed, we can see that a better match leads to a higher survival rate. Even if a patient has a living donor, having a perfect match still leads to a higher rate of survival than having a dead donor but a terrible match.

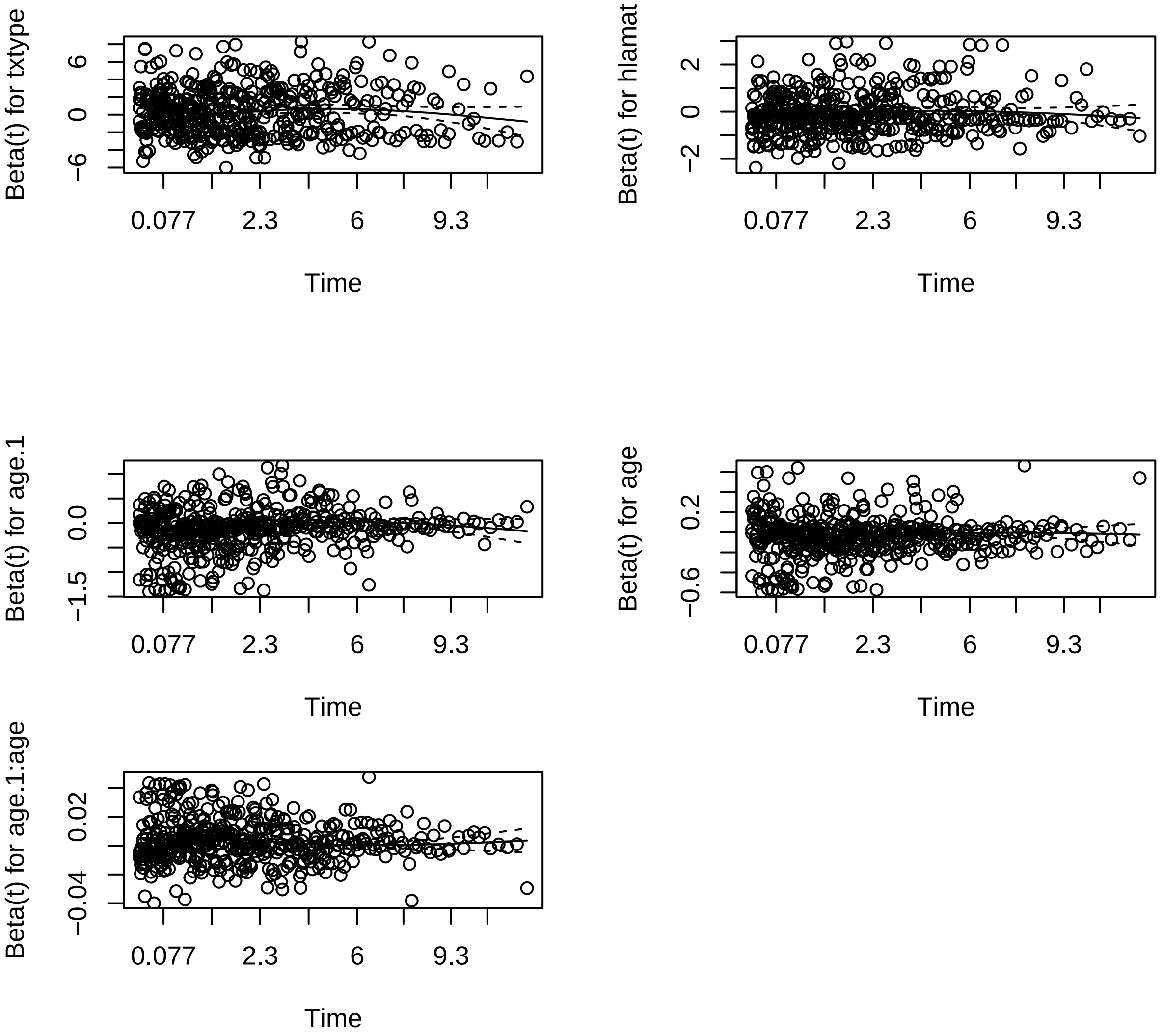
Exercise 9: Check the proportional hazards assumption.

Using our model, we will now investigate the proportional hazard assumption.

```
##            chisq df       p
## txtype      2.95  1 0.08577
## hlamat      8.36  1 0.00384
## age.1      34.19  1 5.0e-09
## age        11.43  1 0.00072
## age.1:age  22.53  1 2.1e-06
## GLOBAL     42.68  5 4.3e-08
```

As we can see, all tests except for txtype are significant. The global test is also significant. This means that we cannot assume that the hazards of different subgroups behave in a parallel manner. Some solutions for this may include adding time-dependent covariates (we unfortunately have none), stratifying (this might be a bit challenging for our continuous age variables, but could be interesting for hlamat), or using a different type of model all together, such as an accelerated failure time model.

Exercise 10: Plot the Schoenfeld residuals and comment.



It should be noted that time is presented on a logistic scale. All plots show a similar non-linear trend. This is particularly apparent during the first year. One conclusion we may draw from this, is that there is a systematic component which is latent to our model. We can see that in earlier years, there are quite a lot of occurrences, which makes the confidence bands rather tight. This makes the absence of proportionality especially significant and gives more power to our conclusions.

In summary, our model is very significant overall which suggests that the selected covariates explain variability in the survival data well. Nevertheless, we can notice a non-linear trend in the residuals which may suggest that some components are missing. Seeing as our data mostly covers rather superficial information about the patients, it could likely be that more specific medical information could be of use. Especially as the trends we observe are non-proportional over time, adding time-dependent information could greatly alleviate our issues and thereby improve both the ability of our model to predict, as well as the causal conclusions we may draw.