

Assignment - 2 (Survival Analysis)

ID	Name	Course
G238135	Htun Teza	RACE 626 - Advanced Analysis in Clinical Research

1. Data Preprocessing

- Using quantiles, Age is divided into three groups as per instruction.

Age Group	Lower Bound	Upper Bound
1	10	36
2	37	47
3	48	70

- Dates are entered in both Gregorian format and Buddhist format. They will all be converted into Gregorian format.

No	Enrollment Data Set	Variable	No	Follow up Data Set	Variable
1	Date of birth	datebirth	1	Date of Visit	datevisit
2	Date of Pre-transplantation	datepret	2		year
3		year	3	Date of Acute Rejection	datear
4	Date of Operation	dateop	4	Date of Graft Failure	gategf
5	Date of Best CR	datebcr	5	Date of Death	datedie
6	Date of Discharge	datedis			
7	Date of Acute Rejection	datear			
8	Date of UTI	dateuti			

- While merging the enrollment data and follow up data, 10 subjects are enrolled and not followed up. 10 subjects will be removed.

No	Subj ID
151	1804688
165	2223194
121	2593548

No	Subj ID
25	2689017
153	2796553
40	2872803
358	3227209
90	3307933
47	3727186
60	4008529

Final Dataset has 500 subjects and 4915 observations (visits).

2. Survival Analysis

The dataset is declared to be for survival analysis. The time is scaled to months.

```

          id: subjid
failure event: gf == 2
obs. time interval: (datevisit1[_n-1], datevisit1]
exit on or before: failure
t for analysis: (time-origin)/30.5
origin: time dateop1

-----
4,915 total observations
43 observations begin on or after (first) failure

-----
4,872 observations remaining, representing
500 subjects
59 failures in single-failure-per-subject data
25,977.639 total analysis time at risk and under observation
at risk from t = 0
earliest observed entry t = 0
last observed exit t = 120.1311

```

A

10 records were excluded beforehand.

B

10 subjects are included in enrollment data set, but not in follow up data set. It means those subjects are not followed up, or lost to follow up. 10 enrollment records are removed prior to merging.

C

There were 43 records that showed “observations begin on or after (first) failure”! For survival analysis, the subjects are to be removed from the study after first failure. However, in the dataset, those patients still followed up after the first failure.

For instance, for subjid 3125636, after first failure, Survival Analysis will stop counting the time to failure. So, the last 2 records are censored.

	subjid	dateop1	datevis~1	_st	_t	gf
3558.	3125636	02dec1998	22dec1998	1	.6557377	1. function
3559.	3125636	02dec1998	18feb1999	1	2.557377	1. function
3560.	3125636	02dec1998	25mar1999	1	3.704918	1. function
3561.	3125636	02dec1998	29apr1999	1	4.852459	1. function
3562.	3125636	02dec1998	10jun1999	1	6.2295082	1. function
3563.	3125636	02dec1998	08jul1999	1	7.147541	1. function
3564.	3125636	02dec1998	23sep1999	1	9.6721311	1. function
3565.	3125636	02dec1998	14oct1999	1	10.360656	2. graft failure
3566.	3125636	02dec1998	16dec1999	0	.	2. graft failure
3567.	3125636	02dec1998	13jan2000	0	.	2. graft failure

It is to be noted that since our failure scenario is graft failure ($gf==2$), loss to follow up ($gf==3$) is not considered failure, and the analysis will count such record as function. There are 27 of such records.

	subjid	dateop1	datevis~1	_st	_t	gf
4579.	3422842	09may2001	14jun2001	1	1.1803279	1. function
4580.	3422842	09may2001	25jun2001	1	1.5409836	3. loss followup

D

```
failure _d: gf == 2
analysis time _t: (datevisit1-origin)/30.5
origin: time dateop1
id: subjid
```

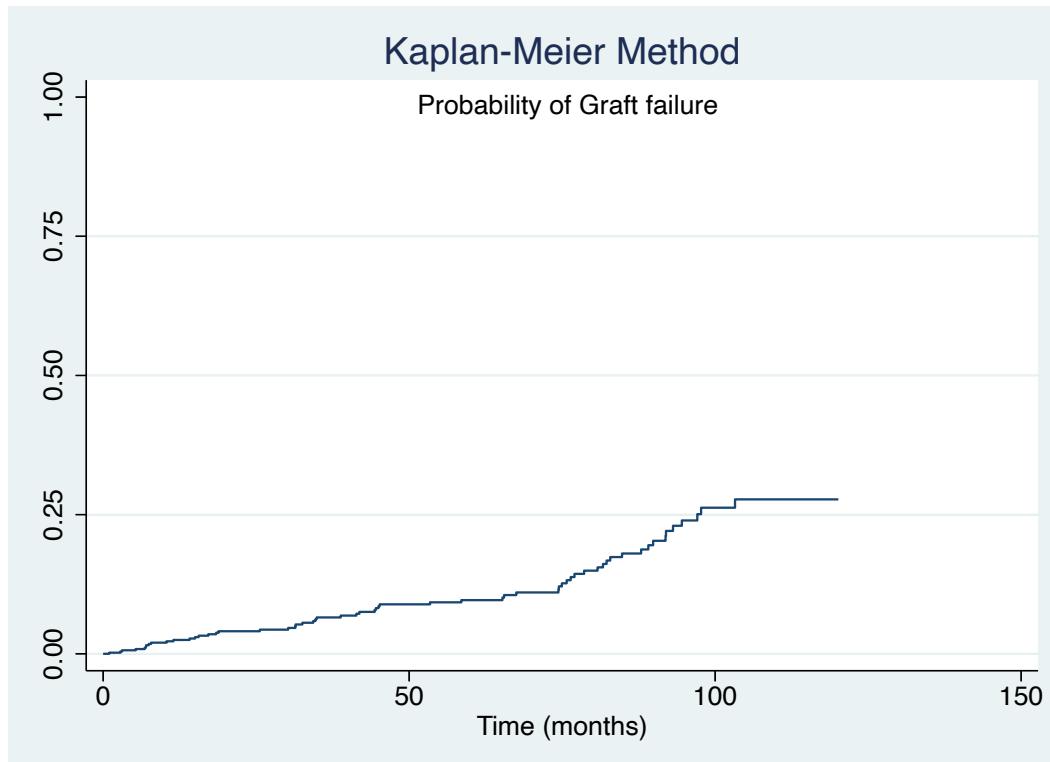
Time	Beg. Total	Fail	Failure Function	Std. Error	[95% Conf. Int.]
------	---------------	------	---------------------	---------------	------------------

24	342	17	0.0405	0.0097	0.0253	0.0645
60	226	17	0.0964	0.0160	0.0694	0.1331

Note: Failure function is calculated over full data and evaluated at indicated times; it is not calculated from aggregates shown at left.

Failure probability at 2 years (24 months) is 4.05% (95% CI = 2.53%, 6.45%).

Failure probability at 5 years (60 months) is 9.64% (95% CI = 6.94%, 13.31%).



```

failure _d: gf == 2
analysis time _t: (datevisit1-origin)/30.5
origin: time dateop1
id: subjid

|           incidence          no. of      |----- Survival time -----|
| time at risk       rate        subjects   25%    50%    75%   |
+-----+
total | 25977.63934 .0022712      500     97.08197 . . .

```

```

failure _d: gf == 2
analysis time _t: (datevisit1-origin)/30.5
origin: time dateop1
id: subjid

```

Estimated rates and lower/upper bounds of 95% confidence intervals

(4872 records included in the analysis)

D	Y	Rate	Lower	Upper
59	2.6e+04	0.0022712	0.0017597	0.0029314

Among 500 patients in 25978 person-months, the estimated incidence was 0.227 per 100 patients per month. (95% CI = 0.175 , 0.293)

25% of patients survive about 97 months or longer.

Median survival time is not applicable.

Cox

Univariate Regression

1. Age Group of the recipient

H_0 : Survival Function is the same for all age groups.

H_A : Survival Function is not the same for all age groups.

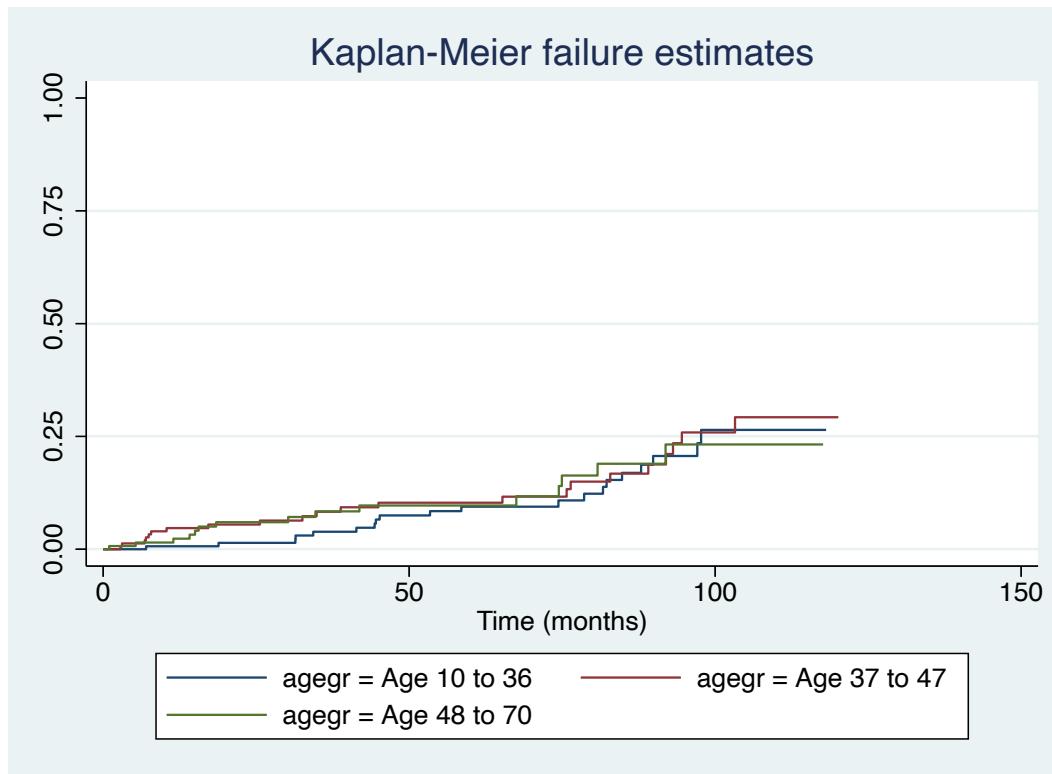
Log-rank test for equality of survivor functions

agegr		Events	Events
		observed	expected
Age 10 to 36		20	22.50
Age 37 to 47		22	20.16
Age 48 to 70		15	14.34
Total		57	57.00

chi2(2) = 0.48
Pr>chi2 = 0.7879

H_0 is failed to be rejected. ($p=0.7879$)

Survival Function is the same for all age groups.



2. Gender of the recipient

H_0 : Survival Function is the same for both gender.

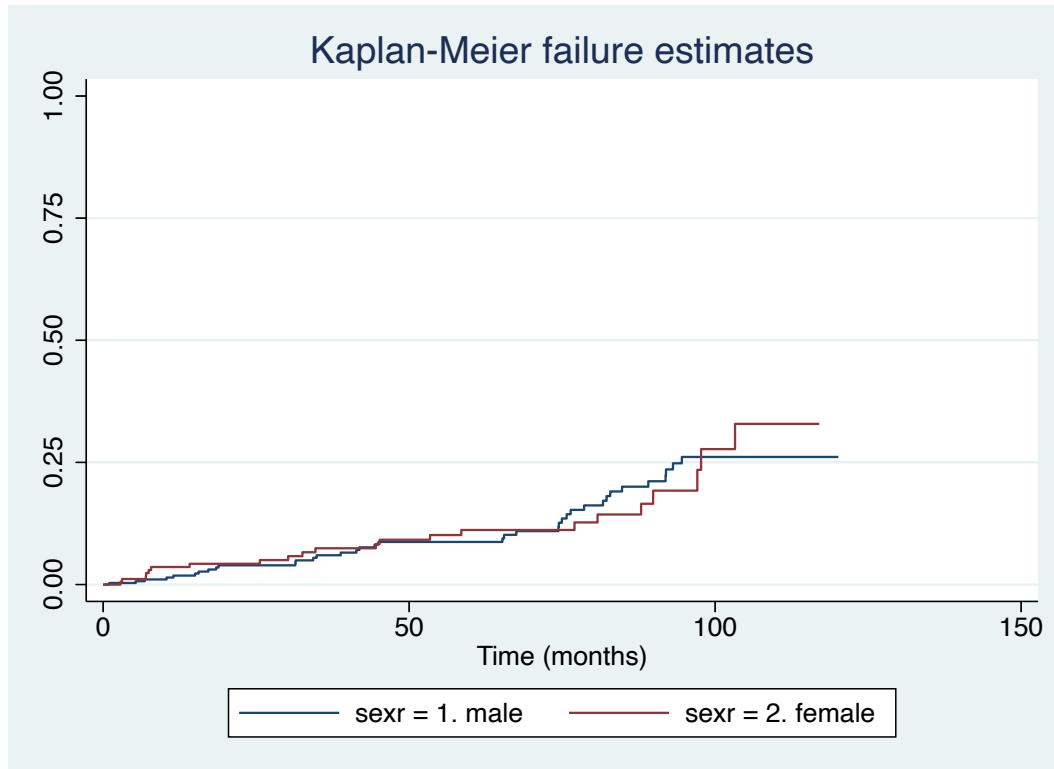
H_A : Survival Function is not the same for both gender.

Log-rank test for equality of survivor functions

sexr		Events	Events
		observed	expected
1. male		37	37.63
2. female		22	21.37
Total		59	59.00

chi2(1) = 0.03
 Pr>chi2 = 0.8651

H_0 is failed to be rejected. ($p=0.8651$)
 Survival Function is the same for both gender.



3. Blood group of the recipient

H_0 : Survival Function is the same for all blood groups.

H_A : Survival Function is not the same for all blood groups.

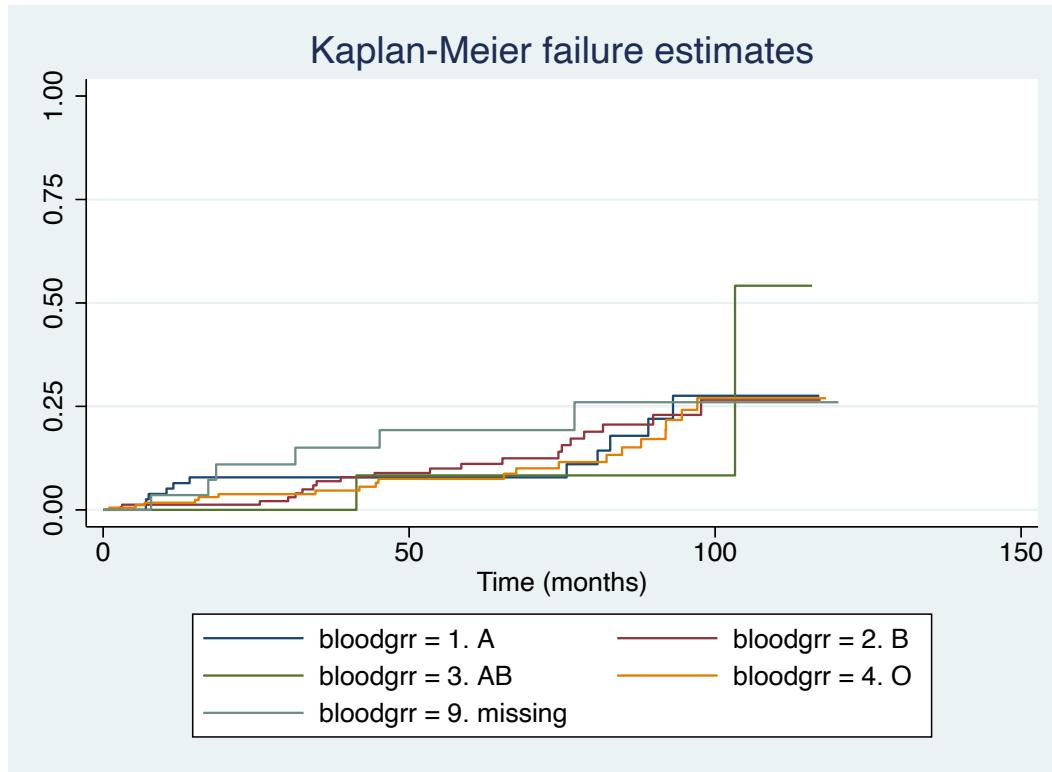
Log-rank test for equality of survivor functions

bloodgrr		Events	Events
		observed	expected
1. A		11	10.50
2. B		20	19.70
3. AB		2	2.41
4. O		20	22.07
9. missing		6	4.32
Total		59	59.00

chi2(4) = 0.94
 Pr>chi2 = 0.9182

H_0 is failed to be rejected. ($p=0.9182$)

Survival Function is the same for all blood groups.



4. Sex of the donor

H_0 : Survival Function is the same for both gender.

H_A : Survival Function is not the same for both gender.

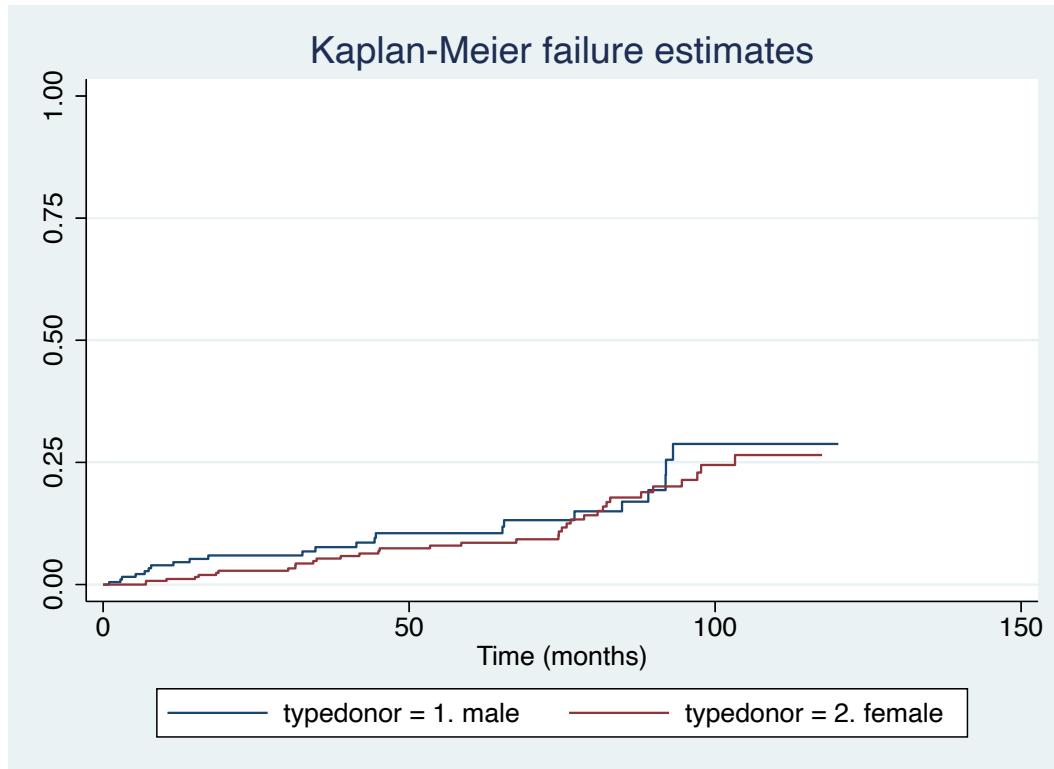
Log-rank test for equality of survivor functions

typedonor		Events	Events
		observed	expected
1. male		23	19.71
2. female		35	38.29
Total		58	58.00

chi2(1) = 0.84
Pr>chi2 = 0.3596

H_0 is failed to be rejected. ($p=0.3586$)

Survival Function is the same for both sex.



5. HLA Matching

H_0 : Survival Function is the same for both matching and mismatching.

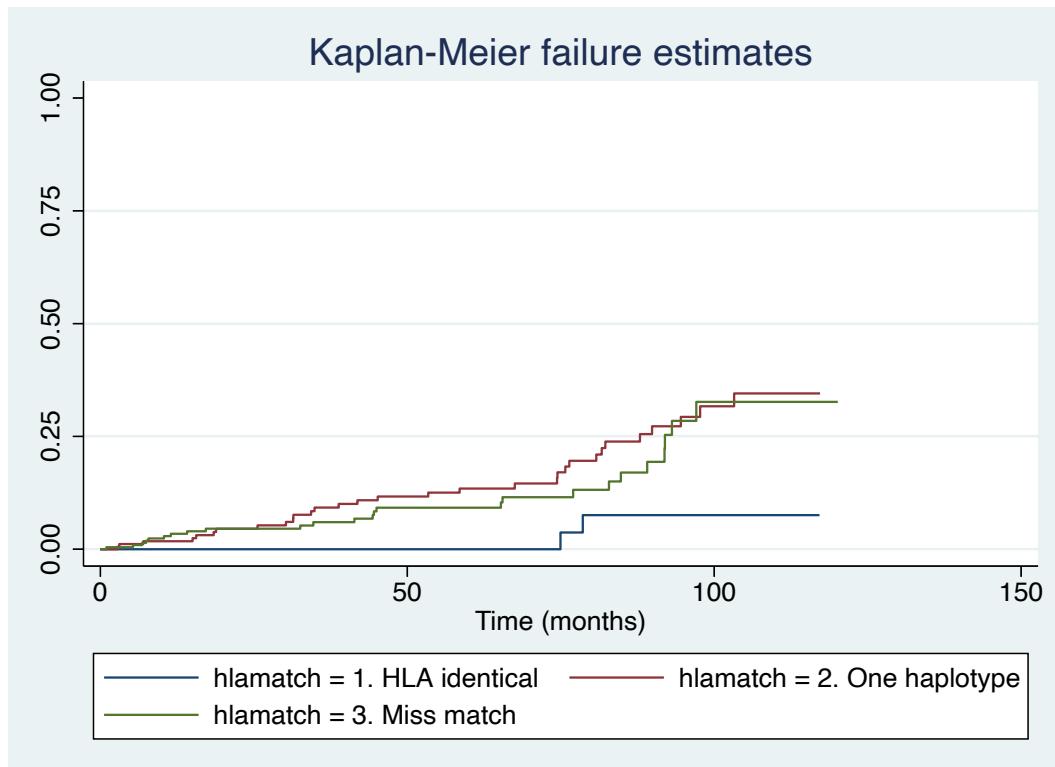
H_A : Survival Function is not the same for both matching and mismatching.

Log-rank test for equality of survivor functions

hla.match		Events	Events
		observed	expected
1. HLA identical		2	9.12
2. One haplotype		31	24.62
3. Miss match		25	24.26
Total		58	58.00
		chi2(2) =	7.25
		Pr>chi2 =	0.0266

H_0 is rejected. ($p=0.0266$)

Survival Function is not the same for both matching and mismatching.



HLA Match	No of Subjects	No of events	Time at risk	Median Survival Time	Incidence (per 100 patients per month)
HLA Identical	63	2	3549.64	NA	0.05634
One Haplotype	191	31	10242.69	NA (25% - 88 months)	0.30265
Mismatch	228	25	10871.67	NA (25% - 92 months)	0.22996

Cox regression -- no ties

No. of subjects =	482	Number of obs =	4,687
No. of failures =	58		
Time at risk =	24664	LR chi2(2) =	9.73
Log likelihood =	-304.01904	Prob > chi2 =	0.0077

Interval	_t Haz. Ratio	Std. Err.	z	P> z	[95% Conf.]
-----+-----	-----	-----	-----	-----	-----
-----	hlamatch				

2. One haplotype		5.754036	4.199213	2.40	0.016	1.376505
24.05289						
3. Miss match		4.726303	3.479613	2.11	0.035	1.116471
20.00763						

HLA Matching "One Haplotype" are 5.75 (95% CI = 1.38, 24.05) times higher risk of graft failure than "HLA Identical", and it is statistically significant. (p=0.016)

HLA Matching "Mismatch" are 4.73 (95% CI = 1.12, 20) times higher risk of graft failure than "HLA Identical", and it is statistically significant. (p=0.035)

6. Presence of HIV antigen in the recipient

There is no Positive samples in the dataset.

7. Presence of Hepatitis B antigen in the recipient

H_0 : Survival Function is the same for both positive and negative.

H_A : Survival Function is not the same for both postive and negative.

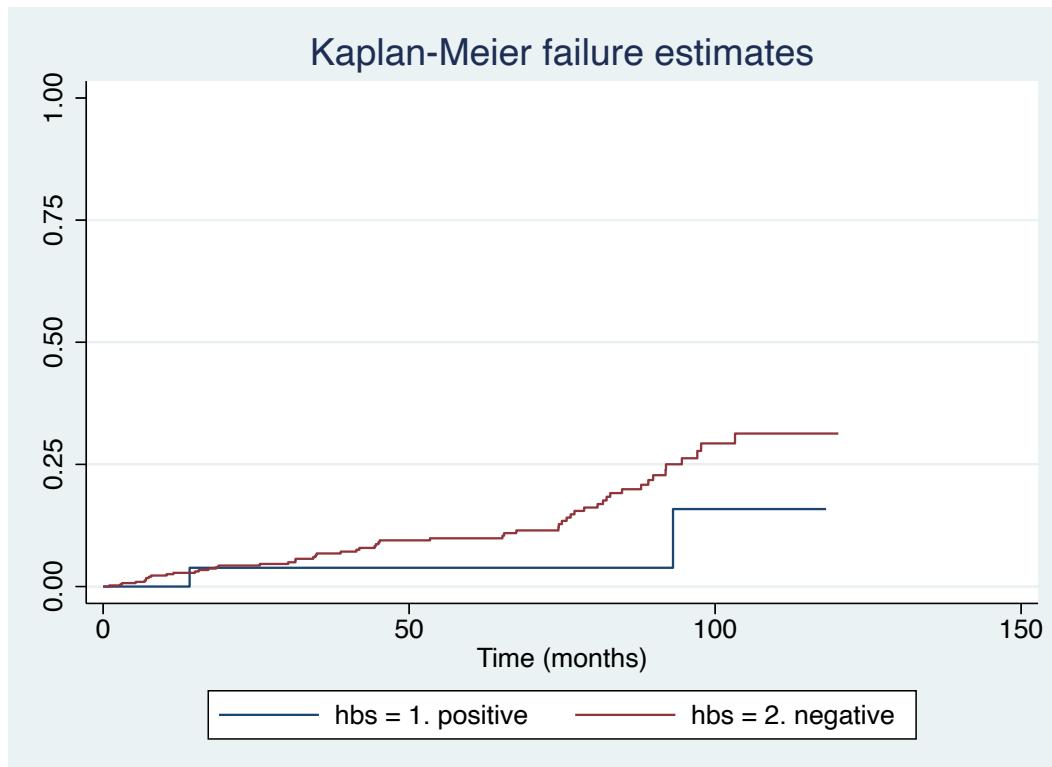
Log-rank test for equality of survivor functions

hbs		Events	
		observed	expected
1. positive		2	4.56
2. negative		55	52.44
Total		57	57.00

chi2(1) = 1.56
Pr>chi2 = 0.2114

H_0 is failed to be rejected. (p=0.2114)

Survival Function is the same for both positive and negative.



8. Presence of Hepatitis B antibody in the recipient

H_0 : Survival Function is the same for both positive and negative.

H_A : Survival Function is not the same for both positive and negative.

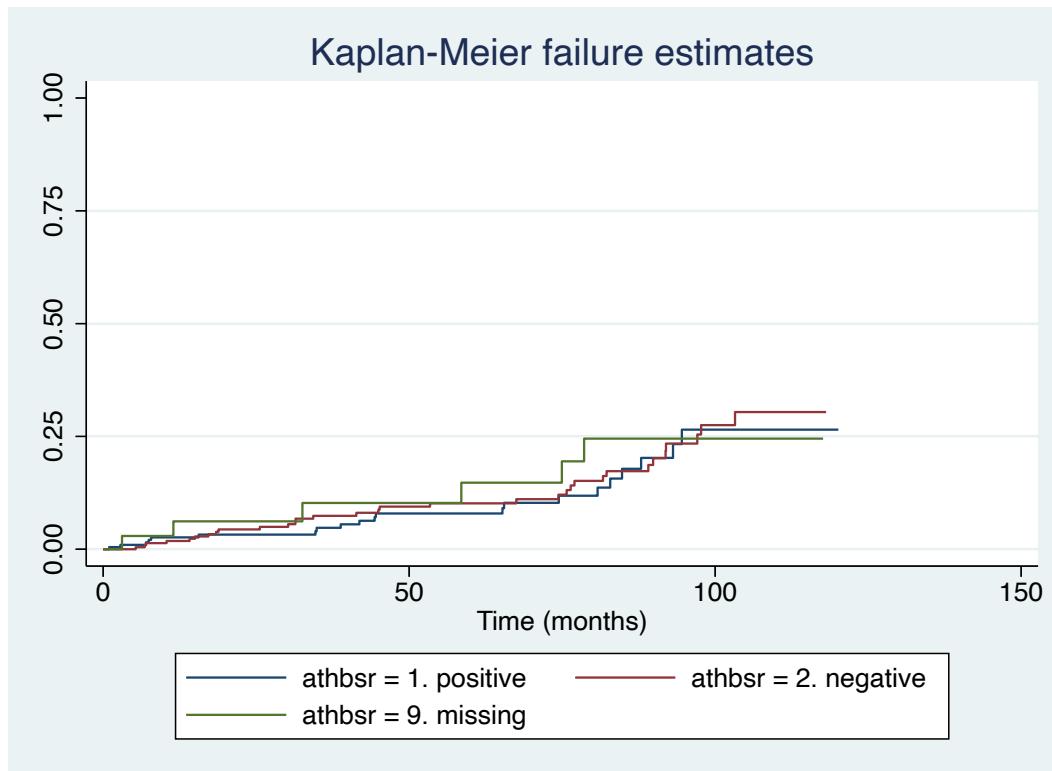
Log-rank test for equality of survivor functions

athbsr	Events observed	Events expected
1. positive	21	22.43
2. negative	32	30.80
9. missing	6	5.77
Total	59	59.00

chi2(2) = 0.15
 Pr>chi2 = 0.9286

H_0 is failed to be rejected. ($p=0.9286$)

Survival Function is the same for both positive and negative.



9. Presence of Hepatitis C antibody in the recipient

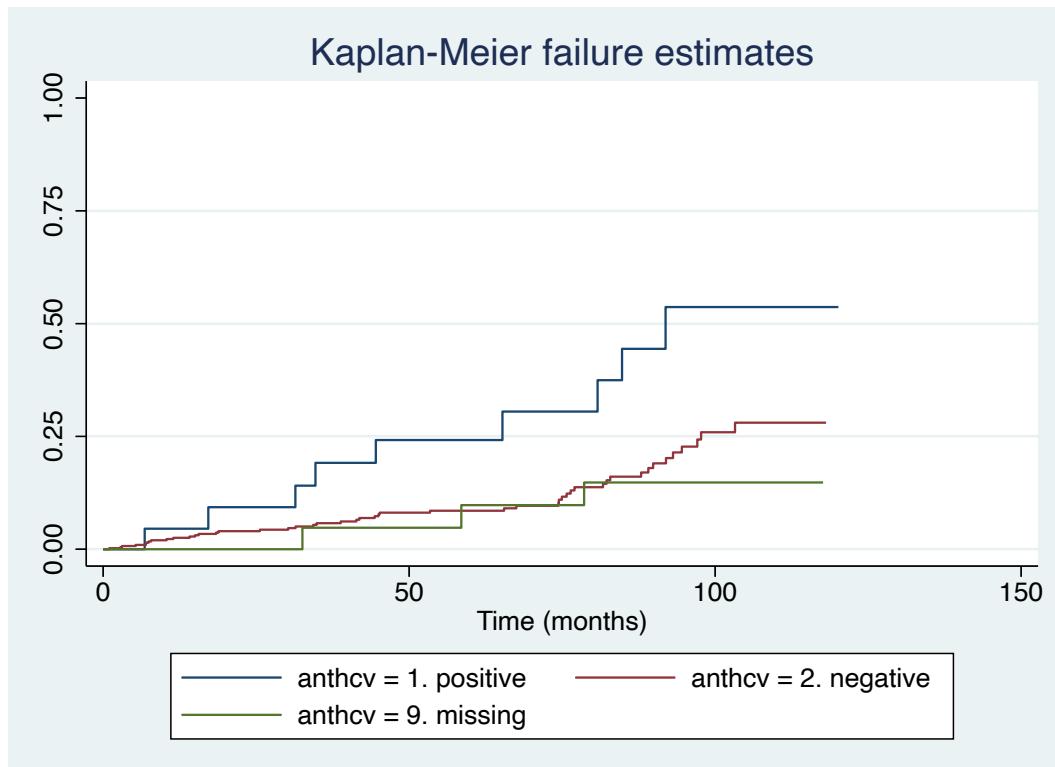
H_0 : Survival Function is the same for both positive and negative.

H_A : Survival Function is not the same for both positive and negative.

Log-rank test for equality of survivor functions			
anthcv	Events		Events expected
	observed		
1. positive	9		3.45
2. negative	47		49.86
9. missing	3		5.69
Total	59		59.00
	chi2(2) = 10.36		
	Pr>chi2 = 0.0056		

H_0 is rejected. ($p=0.0056$)

Survival Function is not the same for both positive and negative.



Hepatitis C	No of Subjects	No of events	Time at risk	Median Survival Time	Incidence (per 100 patients per month)
Positive	24	9	1479.57	91.9 months	0.60828
Negative	450	47	22412.26	NA (25% - 97.7 months)	0.60828
Missing	26	3	2086.8	NA	0.14383

Cox regression -- no ties

No. of subjects =	500	Number of obs =	4,872
No. of failures =	59		
Time at risk =	25977.63934	LR chi2(2) =	7.87
Log likelihood =	-314.40312	Prob > chi2 =	0.0196
<hr/>			
<u>_t</u>	Haz. Ratio	Std. Err.	z
+			P> z
			[95% Conf. Interval]
<hr/>			
anthcv			
1. positive	2.763832	1.0075	2.79
9. missing	.5568952	.3337271	-0.98
<hr/>			

Hepatitis Antibody positive subjects are 2.76 (95% CI = 1.35, 5.65) times higher risk of graft failure than Hepatitis Antibody negative subjects, and it is statistically significant. (p=0.005)

10. Mode of Dialysis

H_0 : Survival Function is the same for all modes of dialysis.

H_A : Survival Function is not the same for all modes of dialysis.

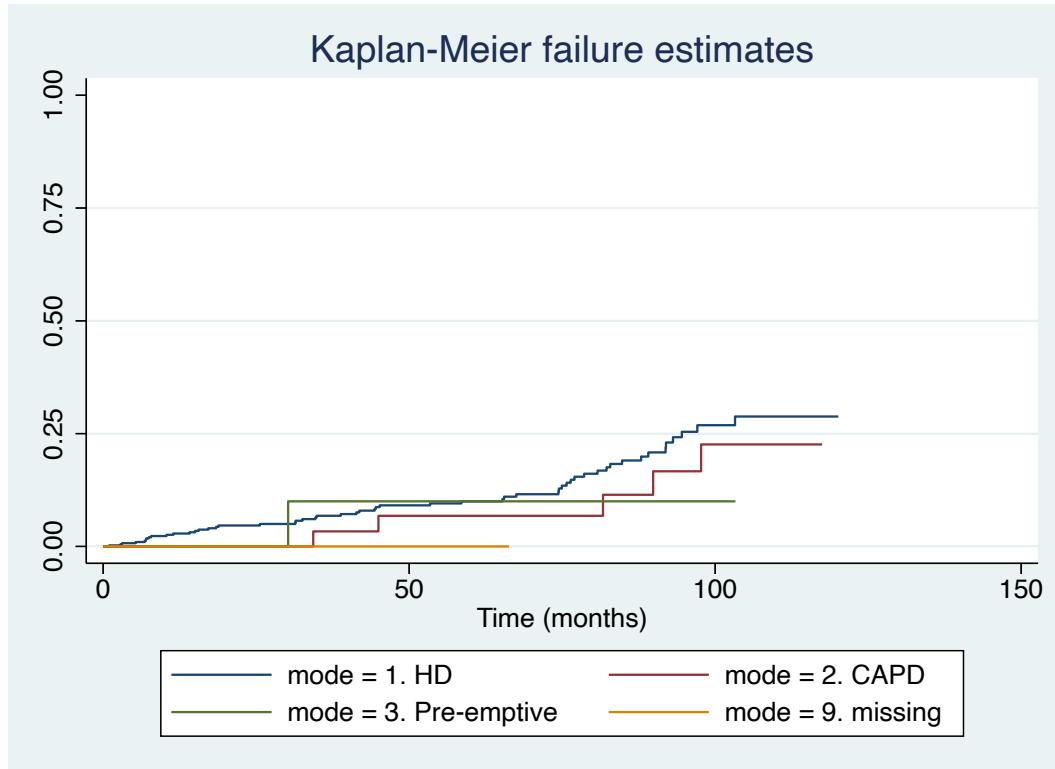
Log-rank test for equality of survivor functions

mode	Events observed	Events expected
1. HD	53	50.11
2. CAPD	5	7.27
3. Pre-emptive	1	1.50
9. missing	0	0.12
Total	59	59.00

chi2(3) = 1.17
Pr>chi2 = 0.7606

H_0 is failed to be rejected. ($p=0.7606$)

Survival Function is the same for all modes of dialysis.



11. Diabetes Mellitus

H_0 : Survival Function is the same for both positive and negative.

H_A : Survival Function is not the same for both positive and negative.

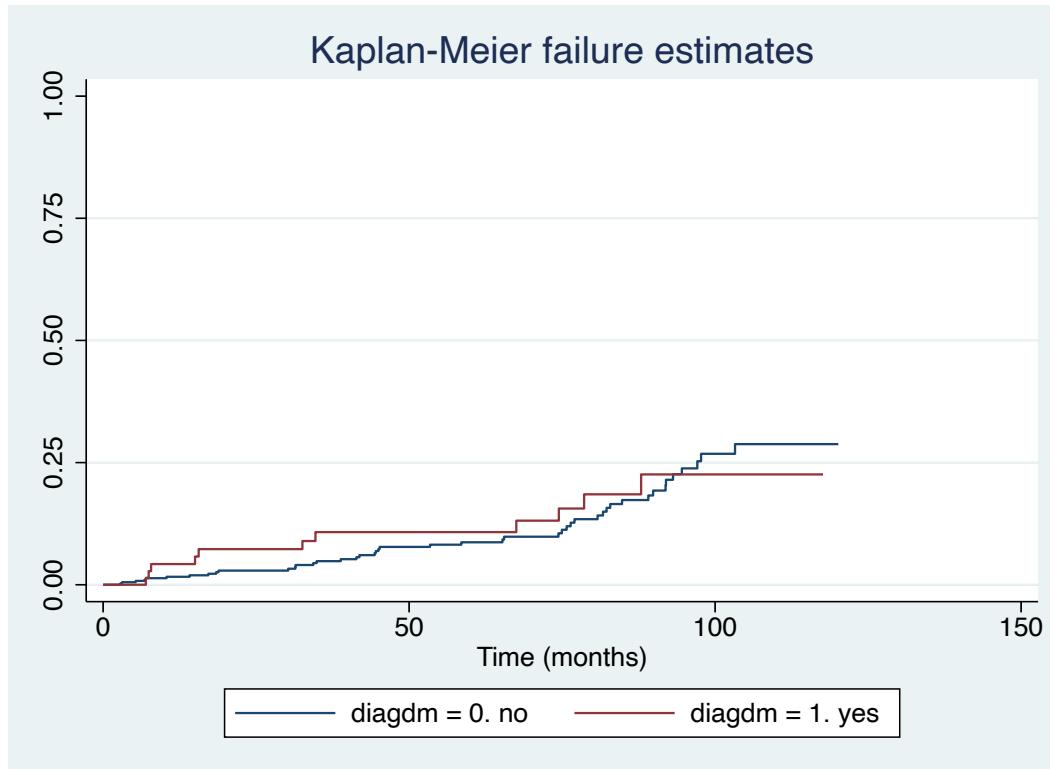
Log-rank test for equality of survivor functions

diagdm	Events observed	Events expected
0. no	45	45.80
1. yes	11	10.20
Total	56	56.00

chi2(1) = 0.08
Pr>chi2 = 0.7813

H_0 is failed to be rejected. ($p=0.7813$)

Survival Function is the same for both positive and negative.



11. Hypertension

H_0 : Survival Function is the same for both positive and negative.

H_A : Survival Function is not the same for both positive and negative.

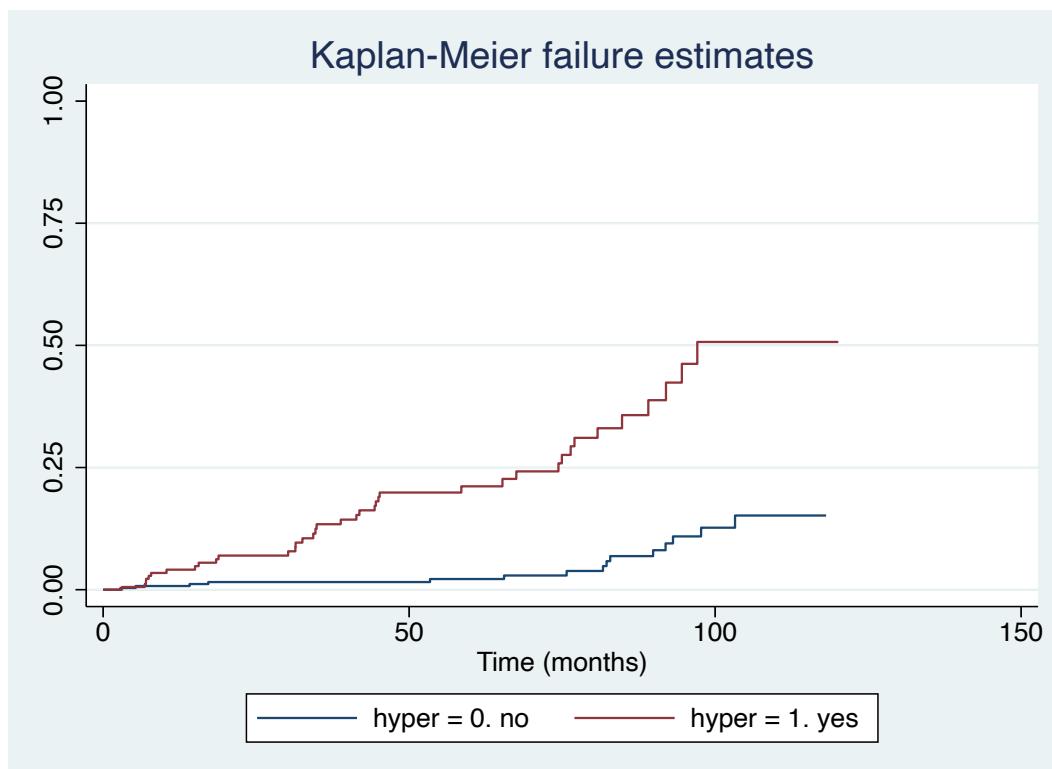
Log-rank test for equality of survivor functions

	Events observed	Events expected
hyper = 0. no	15	36.63
hyper = 1. yes	38	16.37
Total	53	53.00

chi2(1) = 42.27
 Pr>chi2 = 0.0000

H_0 is rejected. ($p < 0.001$)

Survival Function is not the same for both positive and negative.



Hypertension	No of Subjects	No of events	Time at risk	Median Survival Time	Incidence (per 100 patients per month)
Negative	454	15	16741.28	NA	0.0896
Positive	375	38	8247.48	97.1 months	0.46075

Cox regression -- no ties

No. of subjects =	496	Number of obs =	4,793
No. of failures =	53		
Time at risk =	24988.7541	LR chi2(1) =	38.20
Log likelihood =	-263.92571	Prob > chi2 =	0.0000

	<u>t</u>	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
+						
hyper						
1. yes		5.940506	1.83433	5.77	0.000	3.243295 10.88079

Hypertension positive subjects are 5.94 (95% CI = 3.24, 10.88) times higher risk of graft failure than Hypertension negative subjects, and it is statistically significant. ($p<0.001$)

12. High Cholestrol

H_0 : Survival Function is the same for all level of cholestrol.

H_A : Survival Function is not the same for all level of cholestrol.

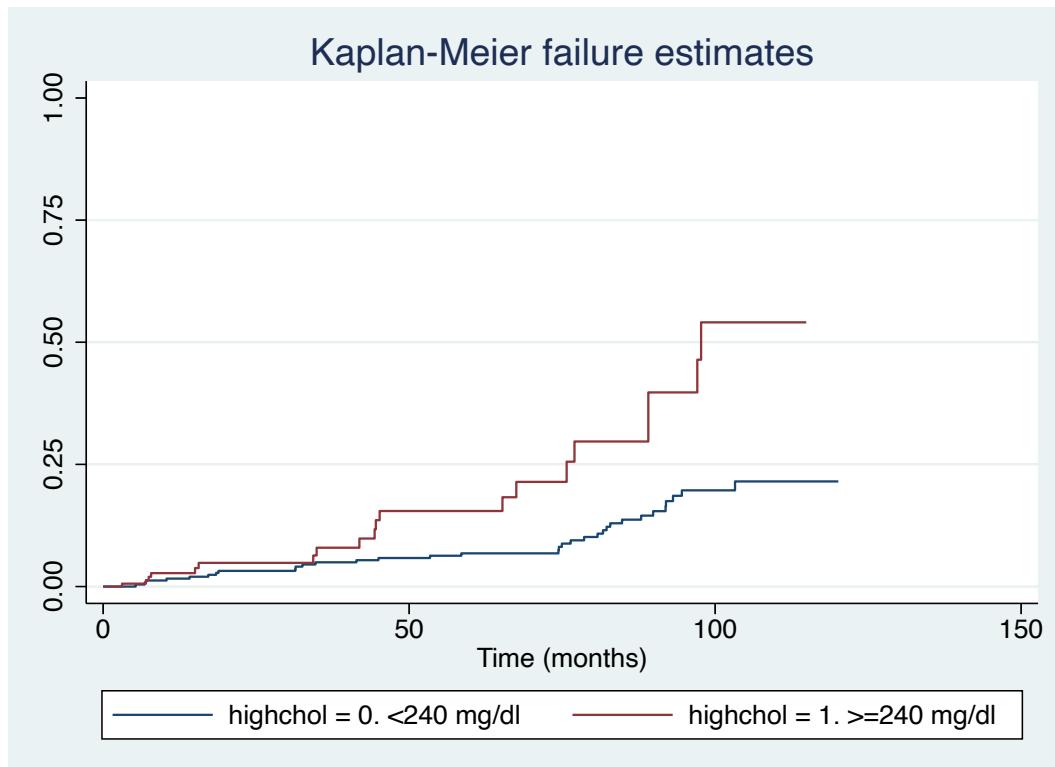
Log-rank test for equality of survivor functions

highchol		Events	
		observed	expected
0. <240 mg/dl		33	42.36
1. >=240 mg/dl		19	9.64
Total		52	52.00

chi2(1) =	11.91
Pr>chi2 =	0.0006

H_0 is rejected. ($p=0.0006$)

Survival Function is not the same for all level of cholestrol.



Cholesterol	No of Subjects	No of events	Time at risk	Median Survival Time	Incidence (per 100 patients per month)
Less than 240 mg/dl	458	33	18975.11	NA	0.17391
Equal or more than 240 mg/dl	324	19	5316.95	97.7	0.35735

Cox regression -- no ties

No. of subjects =	493	Number of obs =	4,287
No. of failures =	52		
Time at risk =	24292.06557	LR chi2(1) =	10.05
Log likelihood =	-269.99322	Prob > chi2 =	0.0015

Interval	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf.]
highchol						
1. >=240 mg/dl	2.739801	.8281296		3.33	0.001	1.51508
4.954529						

Subjects with cholesterol equal or more than 240 mg/dl are 2.74 (95% CI = 1.52, 4.95) times higher risk of graft failure than subjects with cholesterol lower than 240 mg/dl, and it is statistically significant. ($p=0.001$)

13. Receiving Immunosuppressive Drugs

a. Cyclosporin

H_0 : Survival Function is the same for whether or not taking cyclosporin.

H_A : Survival Function is not the same for whether or not taking cyclosporin.

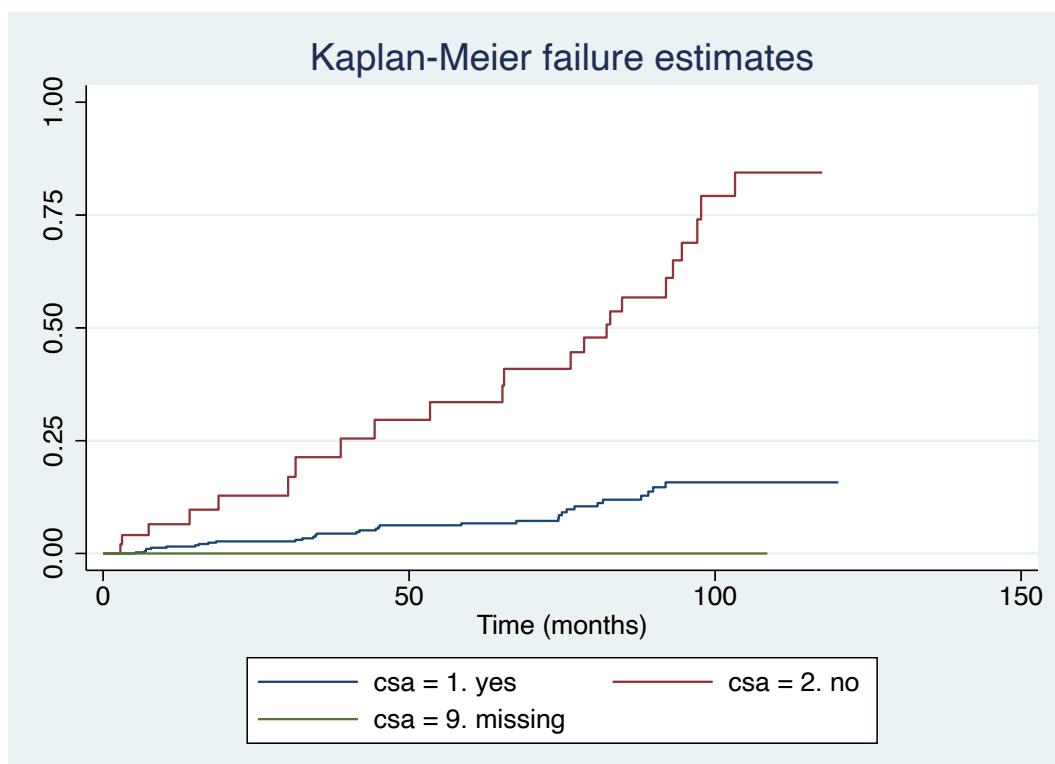
Log-rank test for equality of survivor functions

csa	Events observed	Events expected
1. yes	33	50.92
2. no	23	4.96
9. missing	0	0.12
Total	56	56.00

chi2(2) = 72.40
Pr>chi2 = 0.0000

H_0 is rejected. ($p=0.0006$)

Survival Function is not the same for whether or not taking cyclosporin.



Cyclosporin	No of Subjects	No of events	Time at risk	Median Survival Time	Incidence (per 100 patients per month)
Yes	444	33	23800.95	NA	0.13865
No	102	23	2084.39	82.26	1.10344
Missing	4	0	45.11	NA	0

Cox regression -- no ties

No. of subjects =	497	Number of obs	=	4,865
No. of failures =	56			
Time at risk =	25930.45902	LR chi2(2)	=	42.31
Log likelihood =	-279.05216	Prob > chi2	=	0.0000
<hr/>				
_t Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
<hr/>				
csa				
2. no 7.309908	2.001893	7.26	0.000	4.273678 12.50323
9. missing 6.87e-14	9.84e-07	-0.00	1.000	0 .
<hr/>				

Subjects not taking Cyclosporin are 7.31 (95% CI = 4.27, 12.5) times higher risk of graft failure than subjects taking cyclosporin, and it is statistically significant. ($p<0.001$)

b. Prednisolone

H_0 : Survival Function is the same for whether or not taking prednisolone.

H_A : Survival Function is not the same for whether or not taking prednisolone.

Log-rank test for equality of survivor functions

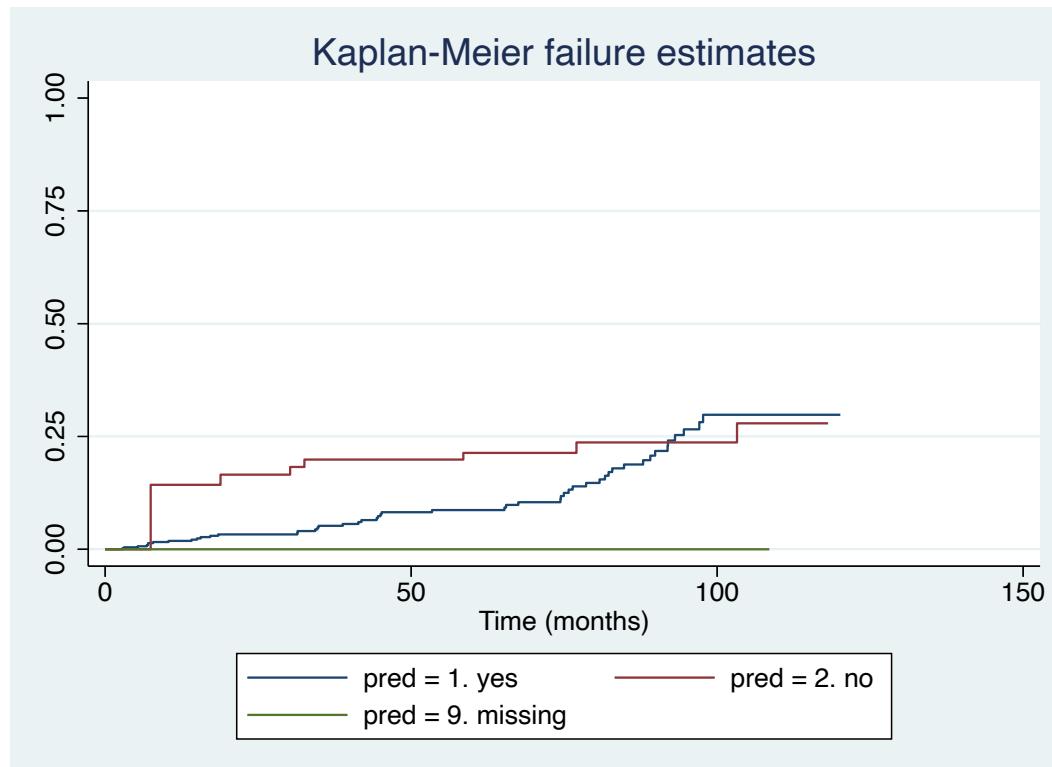
pred	Events	
	observed	expected
1. yes	49	46.14
2. no	7	9.74
9. missing	0	0.12
Total	56	56.00

$\chi^2(2) = 1.12$

Pr>chi2 = 0.5699

H_0 is failed to be rejected. ($p=0.5699$)

Survival Function is the same for whether or not taking prednisolone.



c. MMF

H_0 : Survival Function is the same for whether or not taking MMF.

H_A : Survival Function is not the same for whether or not taking MMF.

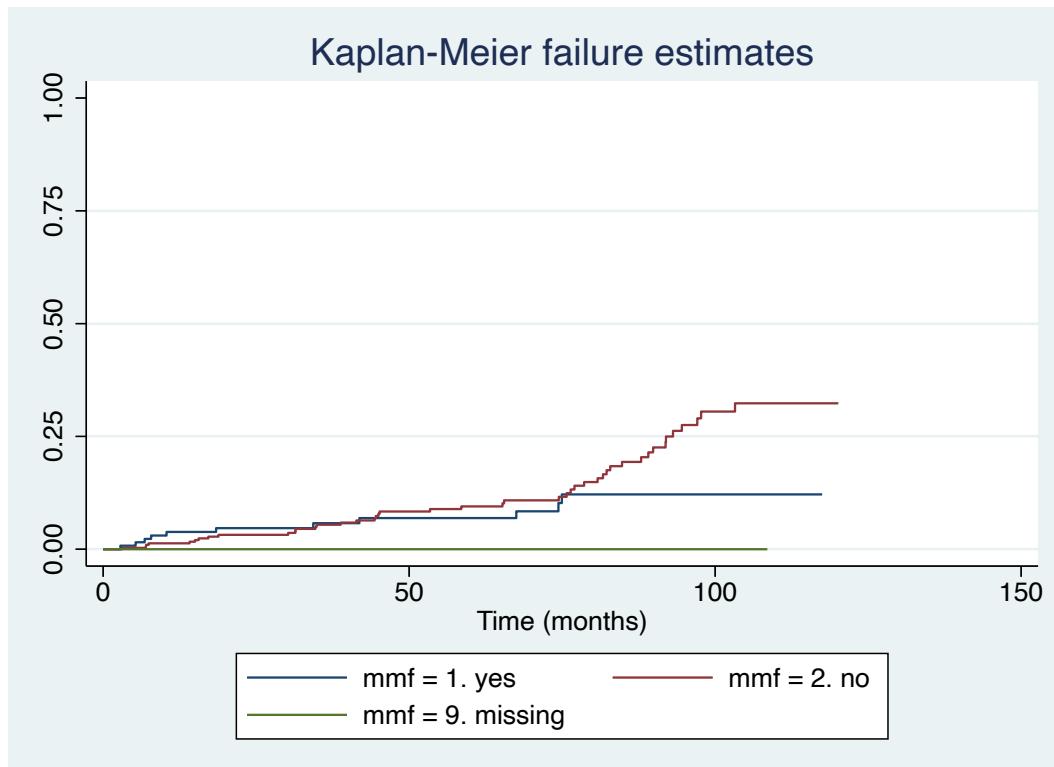
Log-rank test for equality of survivor functions

mmf	Events observed	Events expected
1. yes	11	16.35
2. no	44	38.53
9. missing	0	0.12
Total	55	55.00

chi2(2) = 2.65
Pr>chi2 = 0.2656

H_0 is failed to be rejected. ($p=0.2656$)

Survival Function is the same for whether or not taking MMF.



d. Cardil

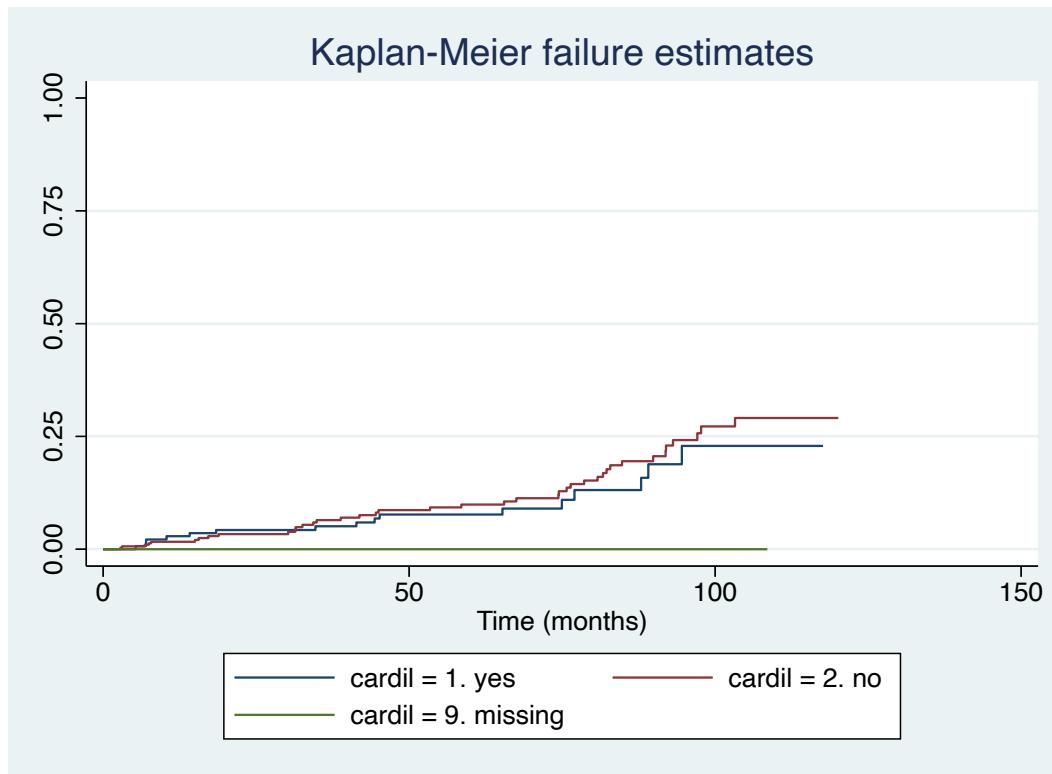
H_0 : Survival Function is the same for whether or not taking Cardil.

H_A : Survival Function is not the same for whether or not taking Cardil.

Log-rank test for equality of survivor functions			
cardil	Events		Events expected
	observed		
1. yes	16		18.24
2. no	40		37.55
9. missing	0		0.21
Total	56		56.00
	chi2(2) =	0.65	
	Pr>chi2 =	0.7236	

H_0 is failed to be rejected. ($p=0.7236$)

Survival Function is the same for whether or not taking Cardil.



Multivariate Regression

Forward Selection

Step 0 : Start with an empty model.

From Univariate Analysis with restraints applied

Feature	LR	P-value
HLA Matching	10.09	0.0064
Hepatitis C	11.21	0.0037
Hypertension	38.13	<0.0001
Cholestrol	12.43	0.0004
Cyclosporin	29.27	<0.0001

Step 1 : Hypertension is included in the model.

Feature	Est Store	LR	P-value
HLA Matching	B	8.43	0.0148
Hepatitis C	C	10.44	0.0054
Cholestrol	D	7.63	0.0057
Cyclosporin	E	29.01	<0.0001

Step 2 : Cyclosporin is included in the model.

Feature	Est Store	LR	P-value
HLA Matching	B1	8.25	0.0041
Hepatitis C	C1	4.61	0.1000
Cholesterol	D1	df(unrestricted) = df(restricted) = 3	

Model with or without high cholesterol has the same level of freedom. The model with high cholesterol will be chosen.

Step 3 : Cholesterol and HLA Matching are both included in the model.

Feature	Est Store	LR	P-value
Hepatitis C	C2	5.73	0.0571

Hepatitis C is not included in the final model.

Multivariate Model

```
Cox regression -- no ties

No. of subjects =          474                Number of obs      =   4,071
No. of failures =         48                 LR chi2(5)        =    84.71
Time at risk     = 22269.14754           Prob > chi2       = 0.0000
Log likelihood   = -206.96805           [95% Conf.

-----+-----+-----+-----+-----+
-----|     _t | Haz. Ratio  Std. Err.      z      P>|z|      [95% Conf.
Interval]
-----+-----+-----+-----+-----+
-----|     hyper |          |
10.64636 1. yes | 5.54144   1.846136   5.14   0.000    2.884324
-----|     csa |          |
12.64806 2. no | 6.811098  2.150924   6.08   0.000    3.667839
-----|  9. missing |  2.22e-18   .       .       .       .
-----|     highchol |
5.110907 1. >=240 mg/dl | 2.737481  .8720136   3.16   0.002    1.466237
```

hla match						
2. One haplotype	8.370519	8.556532	2.08	0.038	1.128847	
62.06827						
3. Miss match	5.625254	5.787593	1.68	0.093	.748817	
42.25796						

Hypertension positive subjects are 5.54 (95% CI = 2.88, 10.65) times higher risk of graft failure than Hypertension negative subjects, and it is statistically significant. ($p<0.001$)

Subjects not taking Cyclosporin are 6.08 (95% CI = 3.67, 12.65) times higher risk of graft failure than subjects taking cyclosporin, and it is statistically significant. ($p<0.001$)

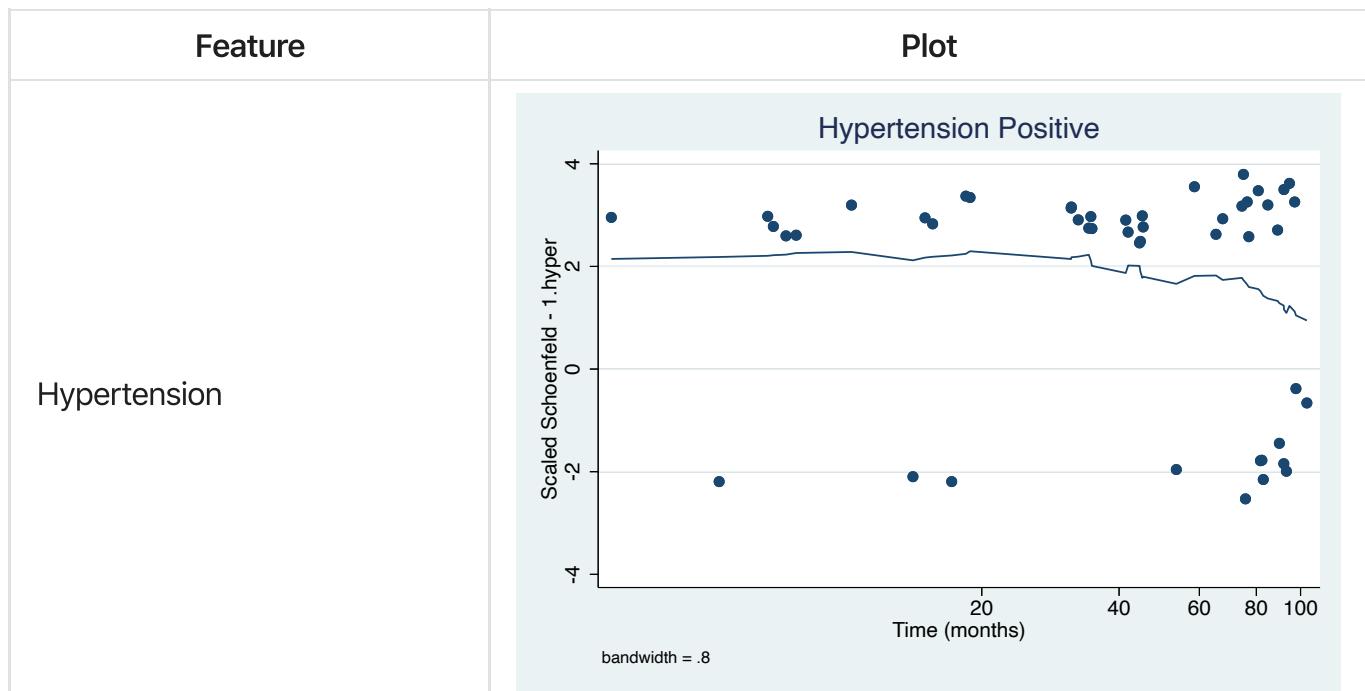
Subjects with cholesterol equal or more than 240 mg/dl are 2.74 (95% CI = 1.47, 5.11) times higher risk of graft failure than subjects with cholesterol lower than 240 mg/dl, and it is statistically significant. ($p=0.002$)

HLA Matching "One Haplotype" are 8.37 (95% CI = 1.12, 62.07) times higher risk of graft failure than "HLA Identical", and it is statistically significant. ($p=0.038$)

HLA Matching "Mismatch" are 5.63 (95% CI = 0.75, 42.26) times higher risk of graft failure than "HLA Identical", and it is not statistically significant. ($p=0.098$)

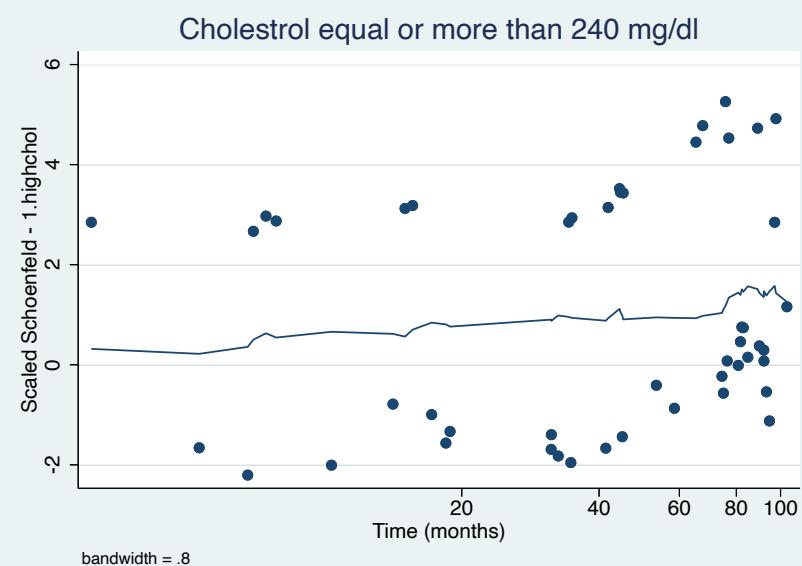
Assumption Checking

1. Plotting scaled schoenfeld residual versus function time

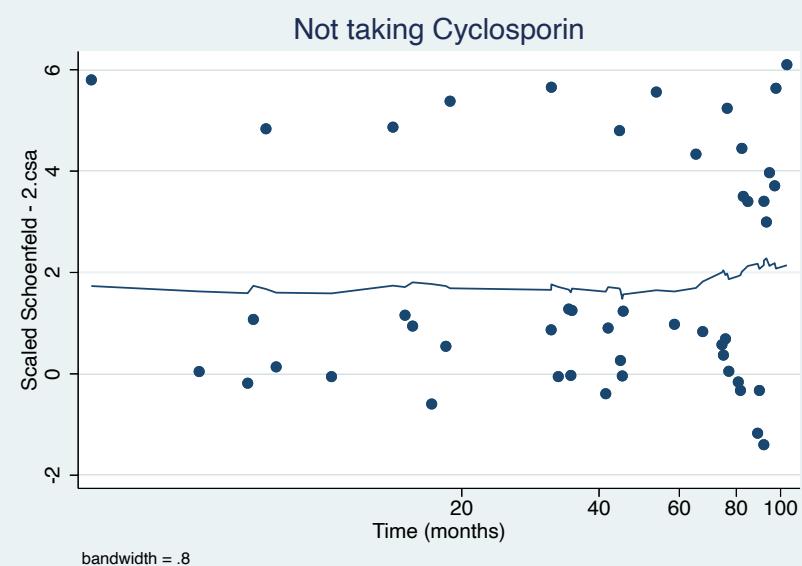


Feature**Plot**

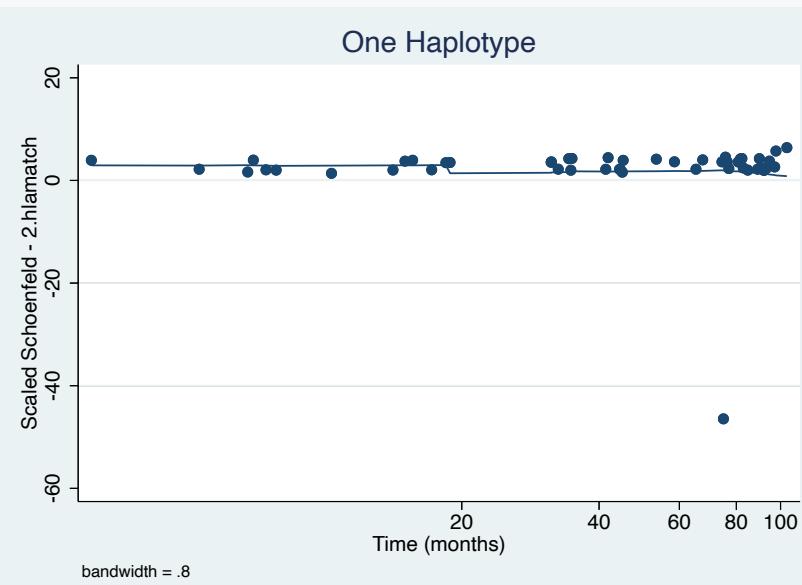
Cholesterol equal or more than 240 mg/dl

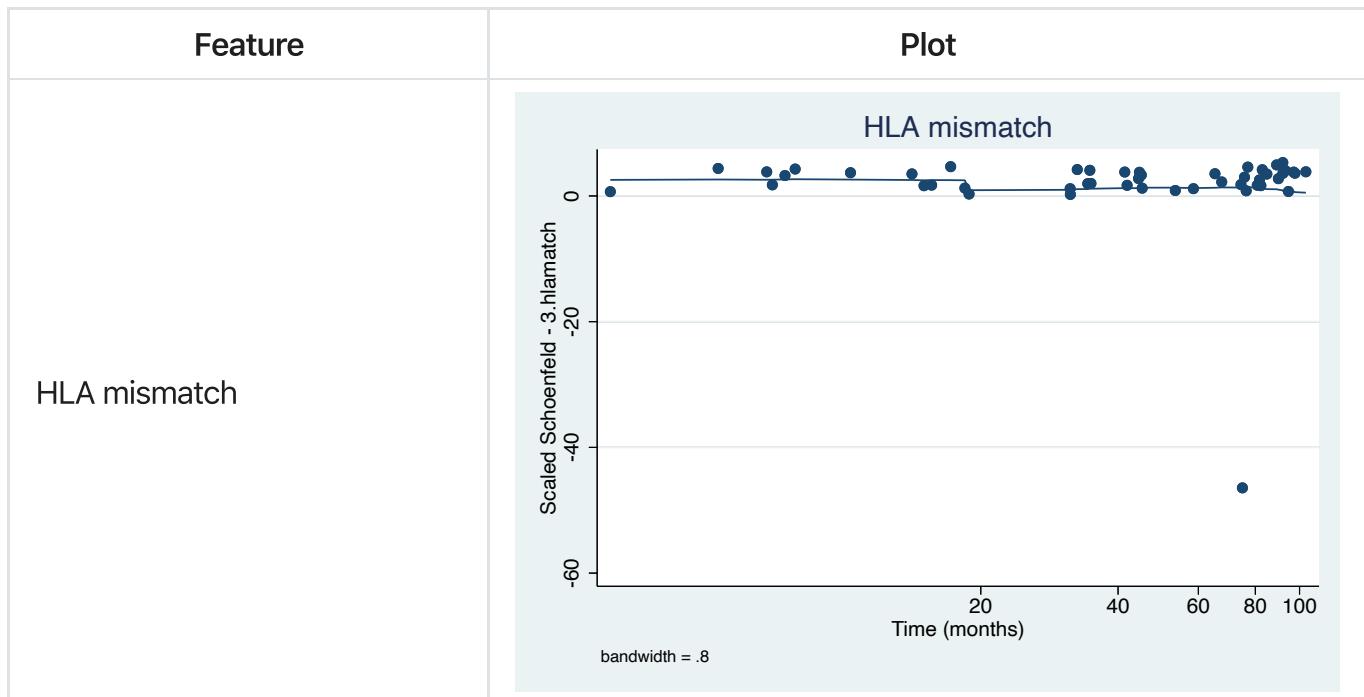


Not taking Cyclosporin

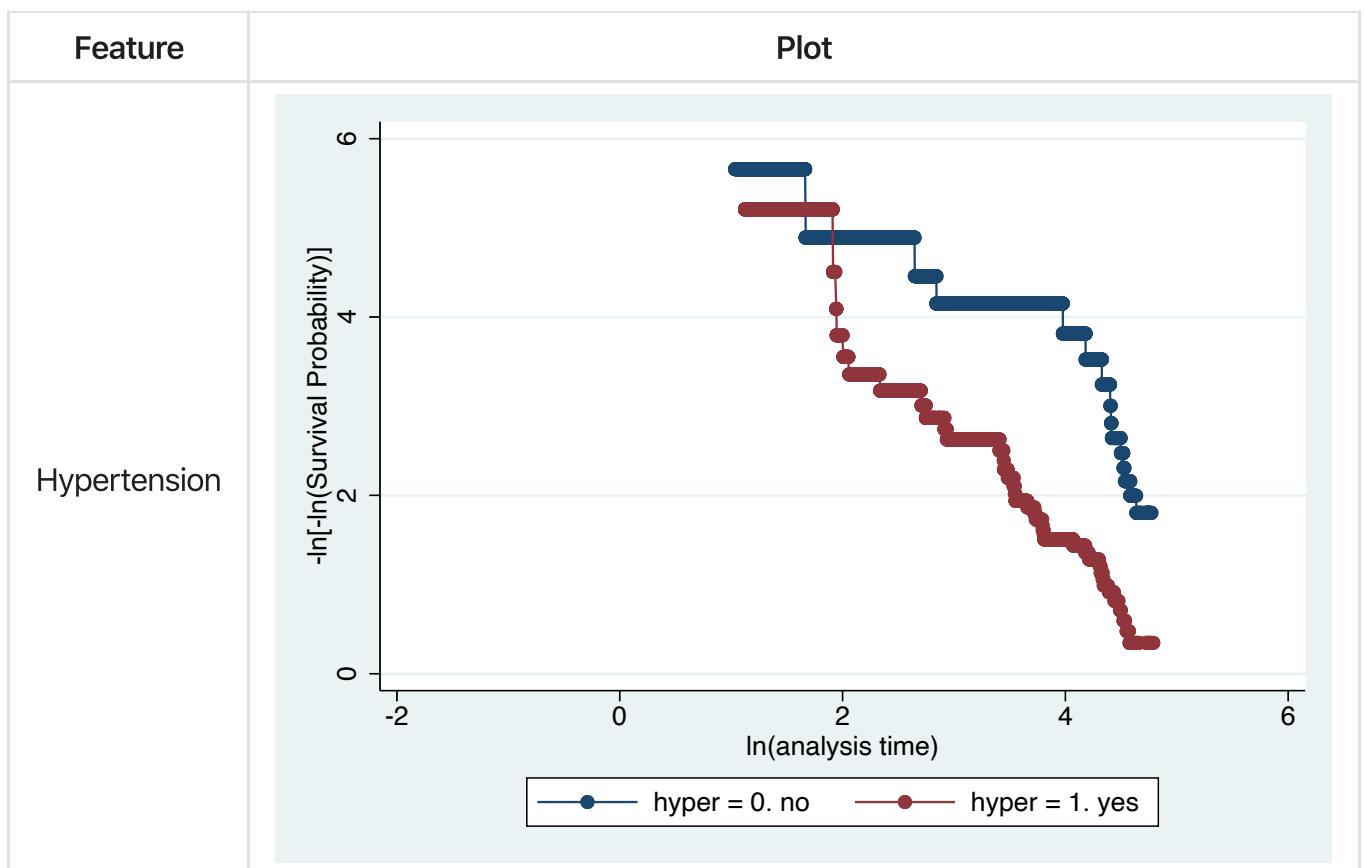


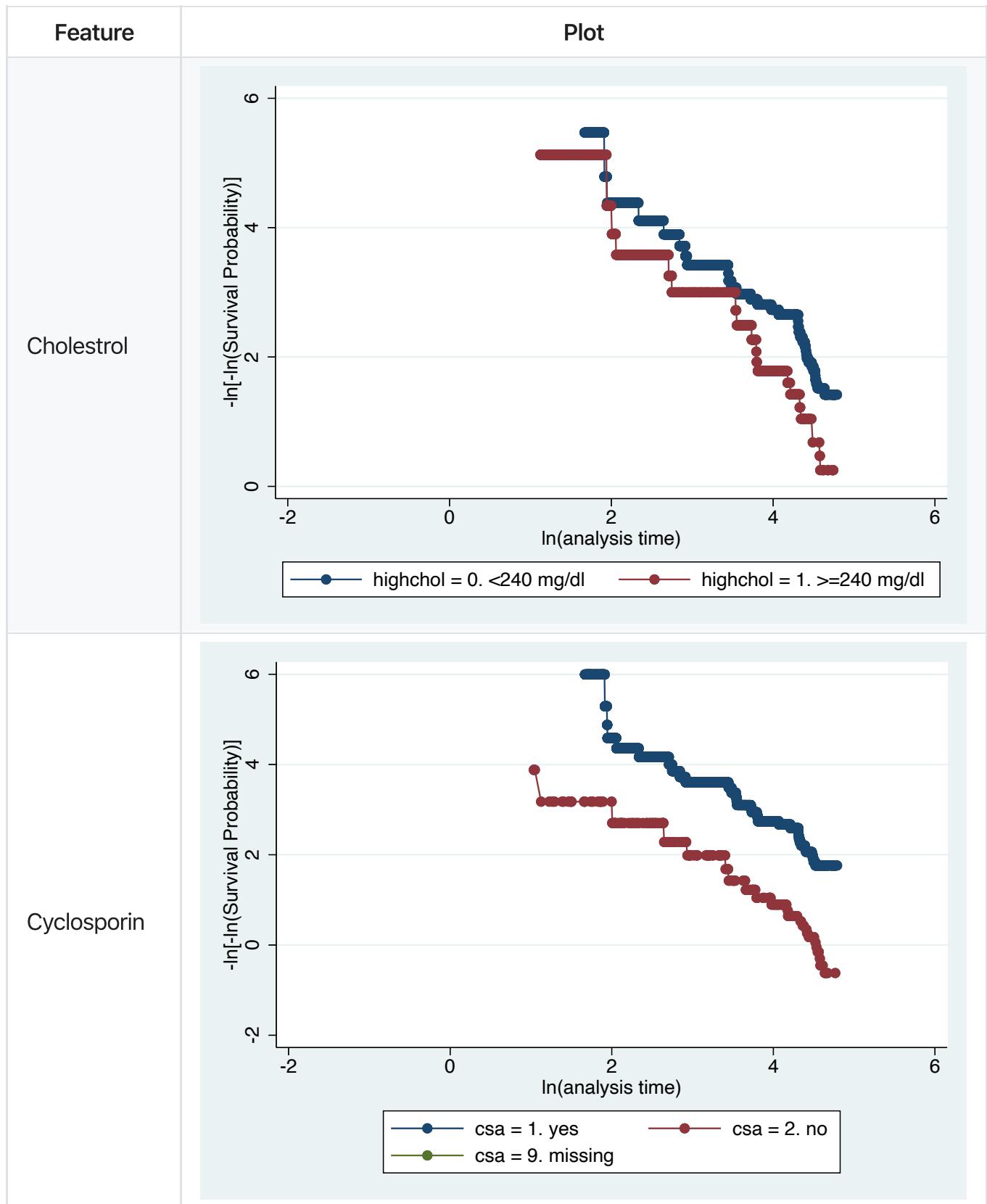
HLA match One haplotype

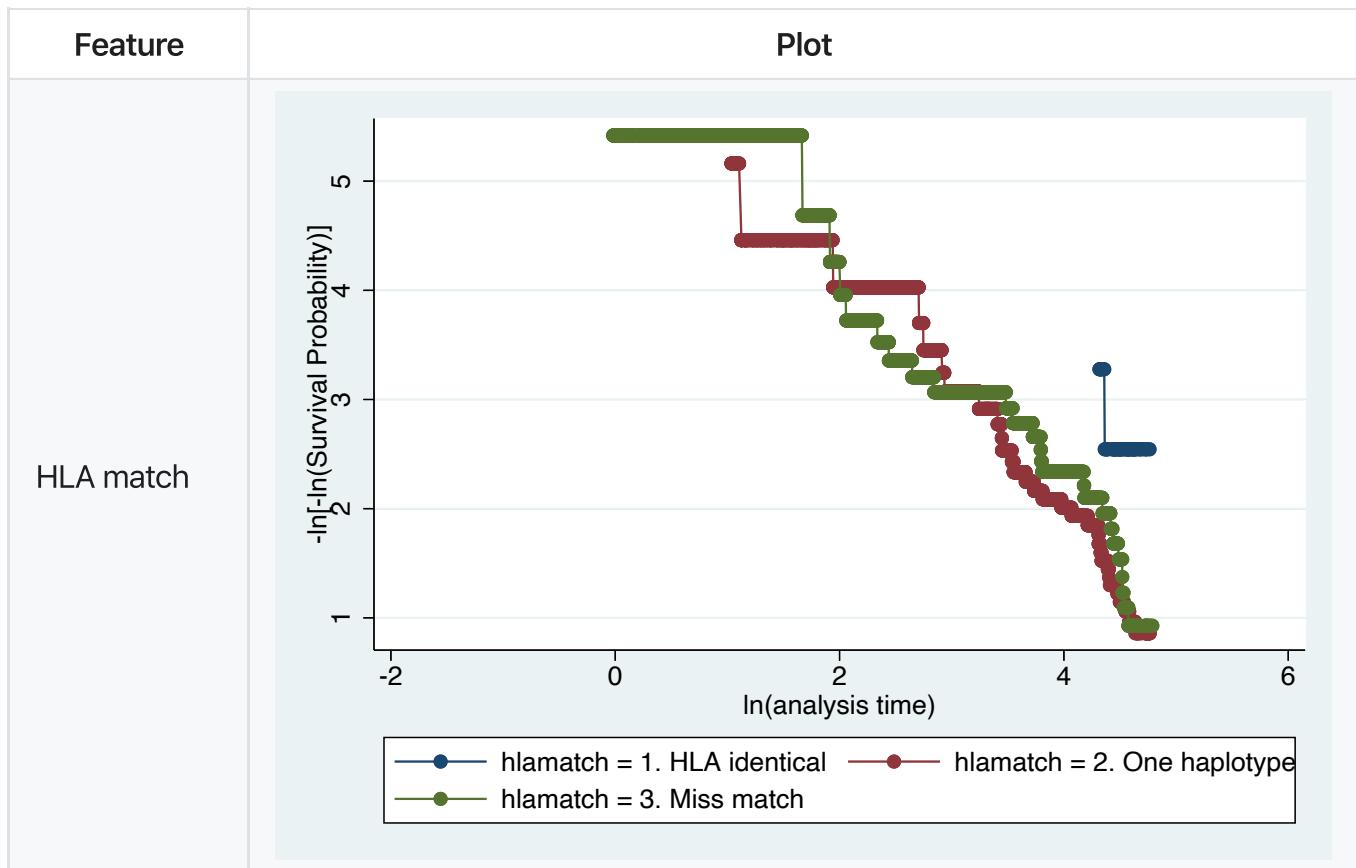




2. Log-log plot







3. Statistical test

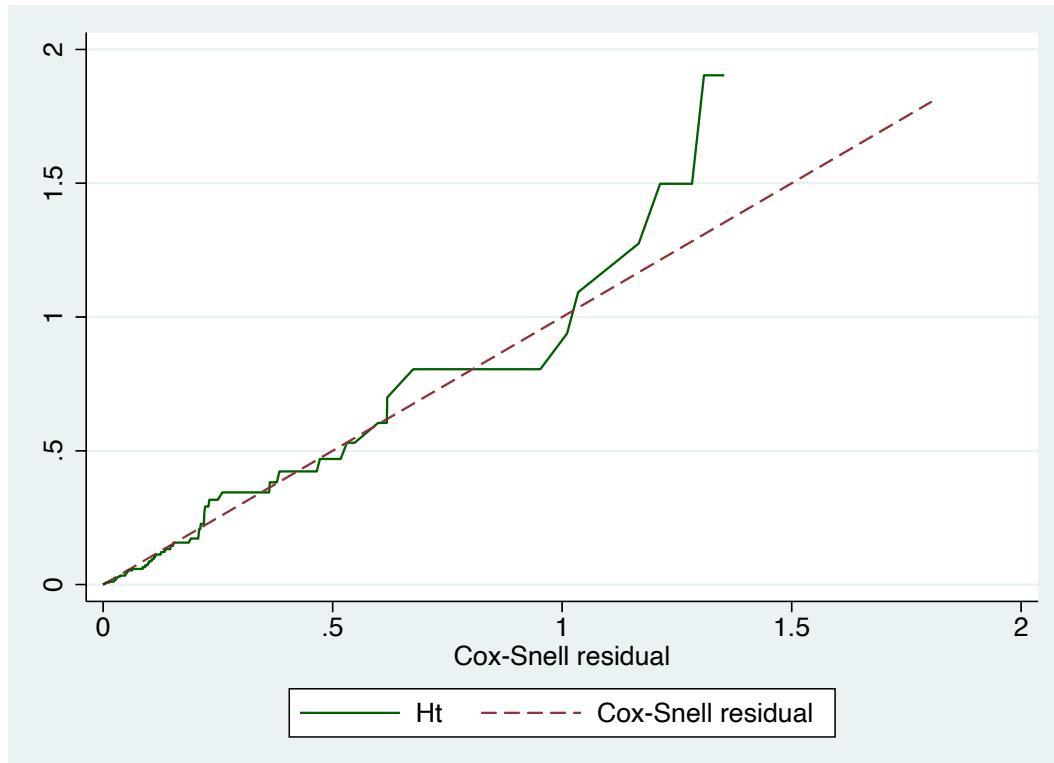
Test of proportional-hazards assumption

Time: Time

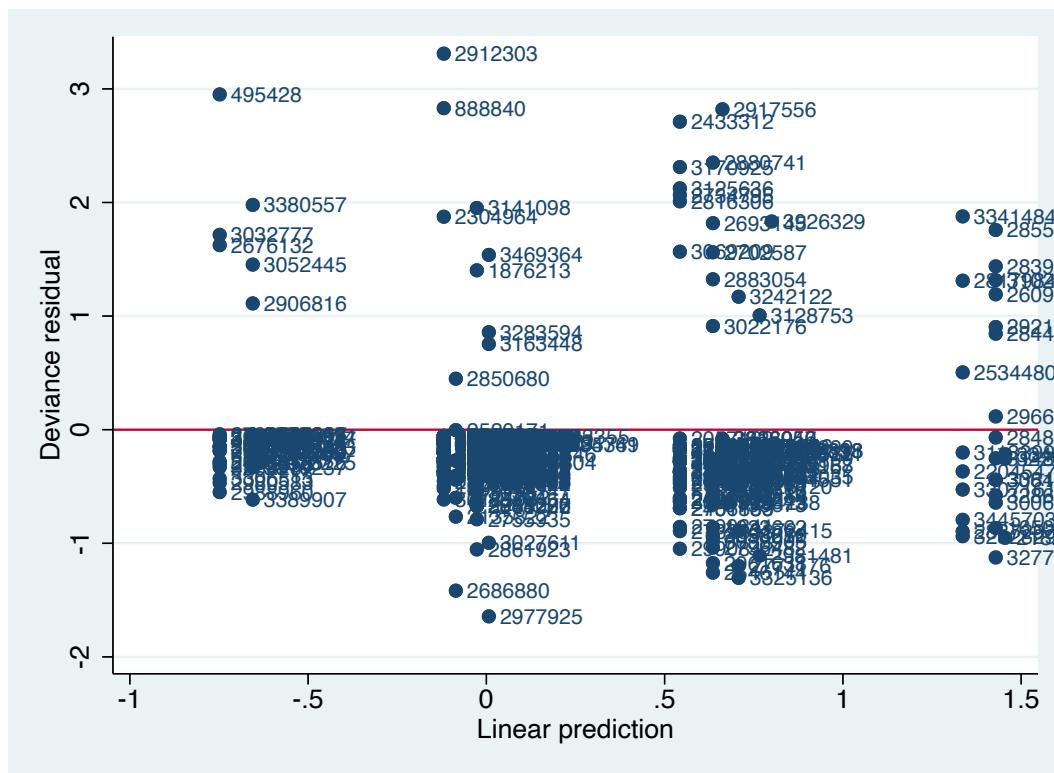
	rho	chi2	df	Prob>chi2
0b.hyper	.	.	1	.
1.hyper	-0.22816	2.20	1	0.1384
1b.csa	.	.	1	.
2.csa	0.12785	0.83	1	0.3608
9.csa	.	.	1	.
0b.highchol	.	.	1	.
1.highchol	0.16769	1.45	1	0.2280
1b.hlamatch	.	.	1	.
2.hlamatch	-0.06231	0.19	1	0.6619
3.hlamatch	-0.07055	0.24	1	0.6237
global test		4.35	5	0.5004

Although it is not statistically significant, none of the coefficients are constant over time. There might be problems with outliers influencing the estimation of coefficients.

4. Assess goodness of fit



5. Identify outliers by deviance



6. Assessing influence of outliers

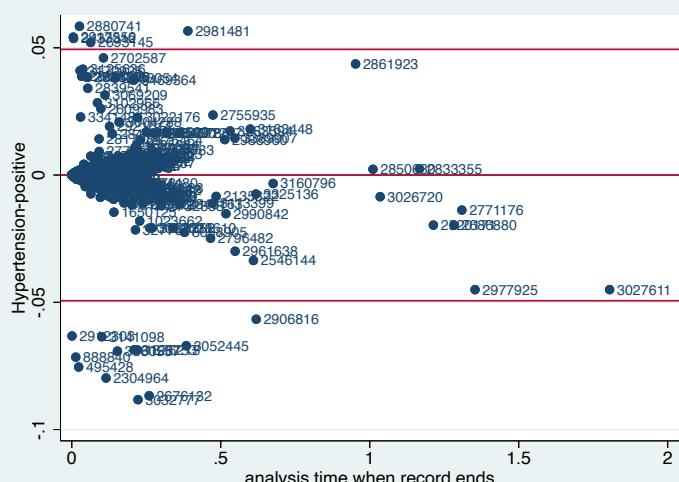
Feature	Plot	Outliers [1]

Feature

Plot

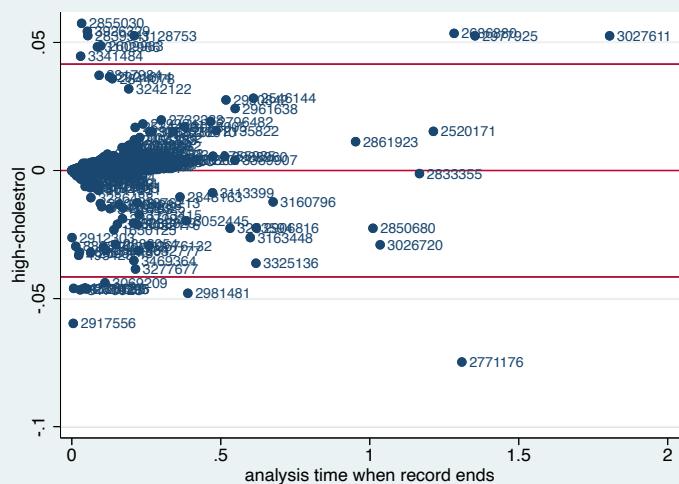
Outliers
1

Hypertension



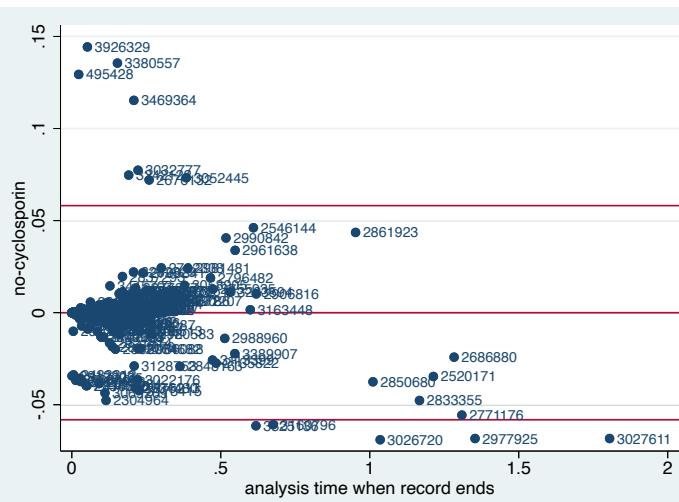
16

Cholesterol equal or more than 240 mg/dl



17

Not taking Cyclosporin



13

Feature	Plot	Outliers [1]
HLA match One haplotype	<p>Scatter plot showing hla-match-one-haplotype on the y-axis versus analysis time when record ends on the x-axis. The y-axis ranges from -1 to 0.5, and the x-axis ranges from 0 to 2. A horizontal red line at approximately 0.15 represents the mean. Two outliers are labeled: 2182325 at x~0.1, y~0.45, and 2917556 at x~0, y~-1.0.</p>	2
HLA mismatch	<p>Scatter plot showing hla-mismatch on the y-axis versus analysis time when record ends on the x-axis. The y-axis ranges from -1 to 0.5, and the x-axis ranges from 0 to 2. A horizontal red line at approximately 0.15 represents the mean. Two outliers are labeled: 2182325 at x~0.1, y~0.45, and 2917556 at x~0, y~-1.0.</p>	2

1. These subjects have values exceeding 3SD and affect the estimated coefficient. ↵