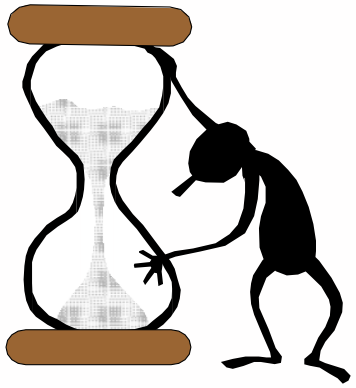
Survival analysis and Cox proportional-hazards model for time-to-event data

What is Survival Analysis?

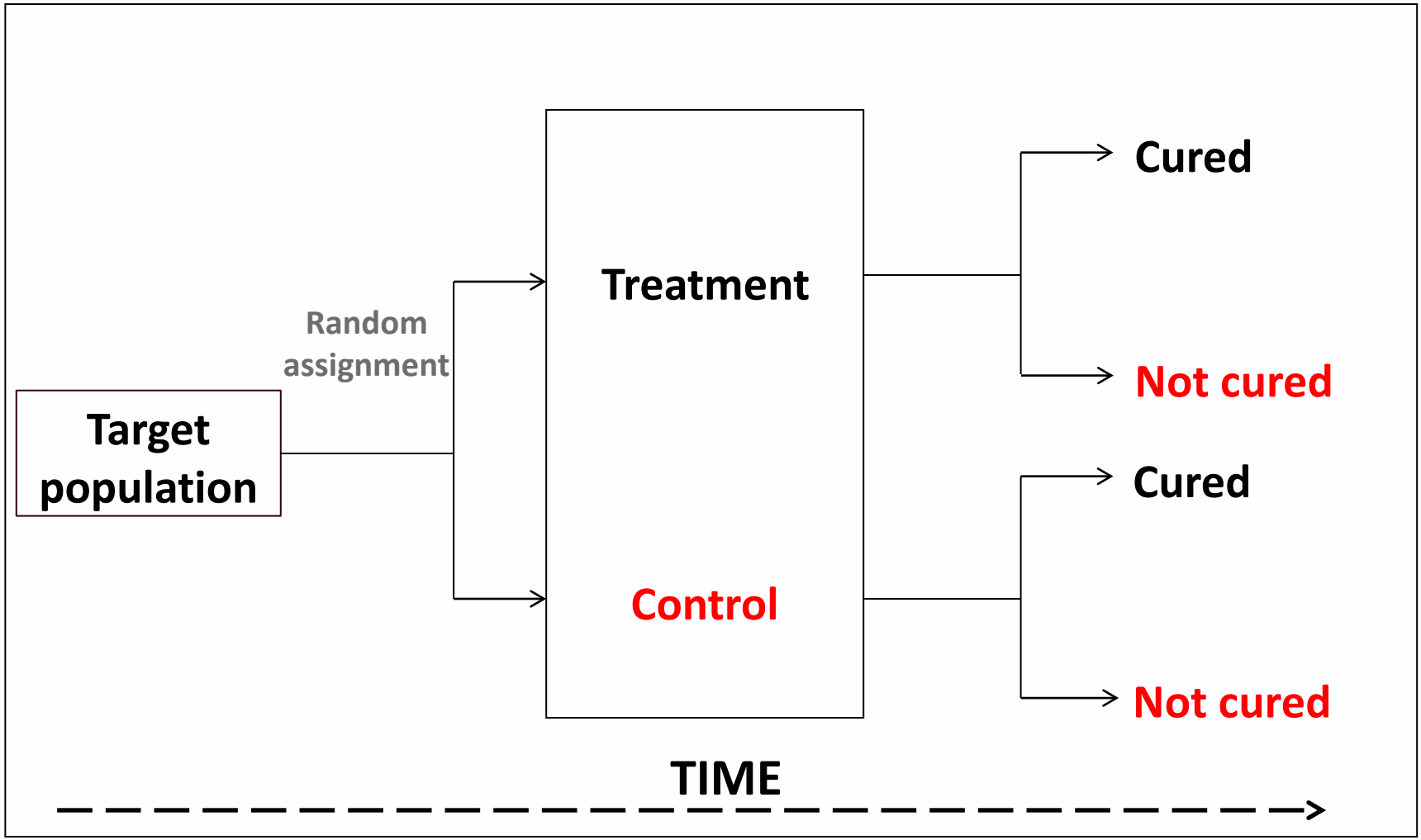
- Survival analysis is a collection of statistical procedures for data analysis for which the **outcome variable** of interest is **time until an event** occurs.
- Also called "time to event analysis"
  - Time to death
  - Time to relapse of a disease
  - Time to recovery from illness
  - Length of stay in a hospital



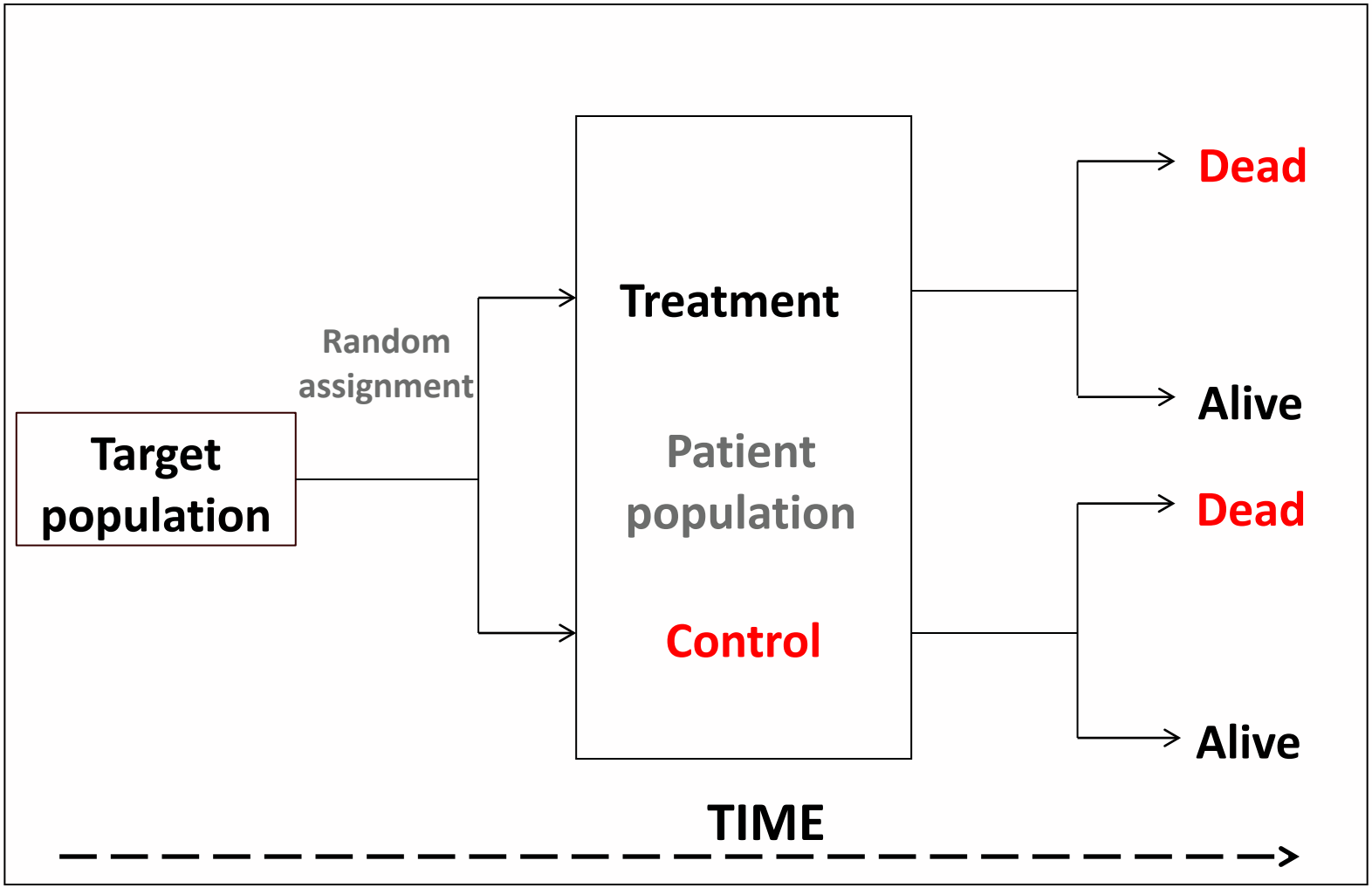
Kind of Survival Studies

- Clinical trials
- Prospective cohort studies
- Retrospective cohort studies

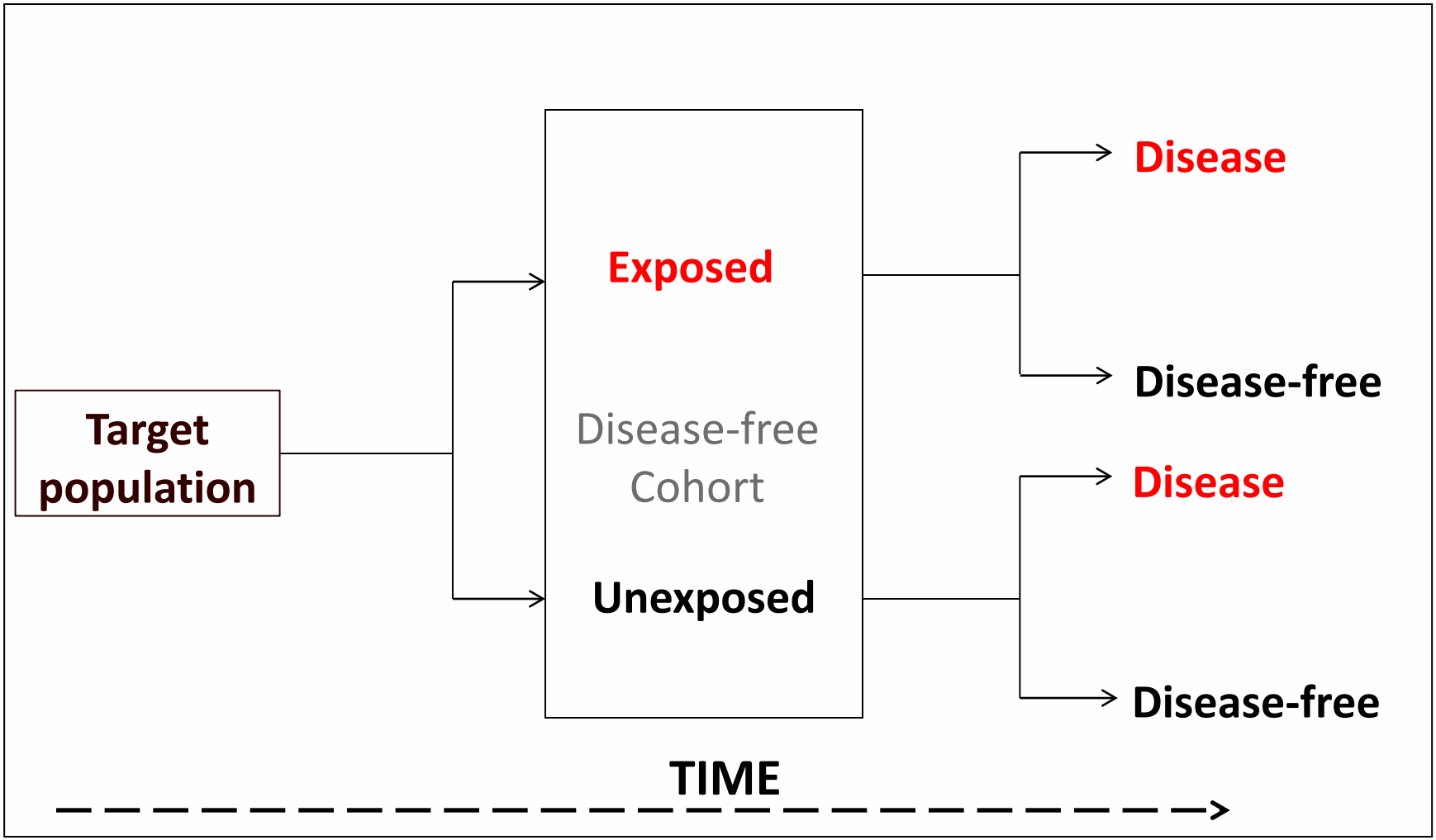
Randomized Clinical Trial (RCT)



Randomized Clinical Trial (RCT)



Cohort Study (Prospective/Retrospective)



Objectives of Survival Analysis

- To estimate time to event for a group of individuals
- To compare time to event between two or more groups
- To assess the relationship between explanatory variables and time to event

Survival Analysis - Advantages

- Why not compare mean time to event between groups using a t-test or linear regression?
  - **ignores censoring**
- Why not compare proportion of events in groups using logistic regression?
  - **ignores censoring**
  - **ignores time**

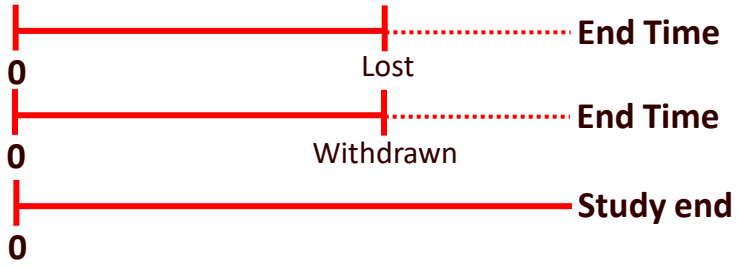
Survival analysis accounts for censored observations as well as time to event.

What is censored data?

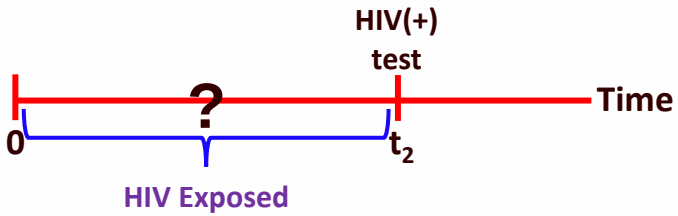
Censored data is any data for which we **do not know** the **exact event time**.

There are three types of censored data –

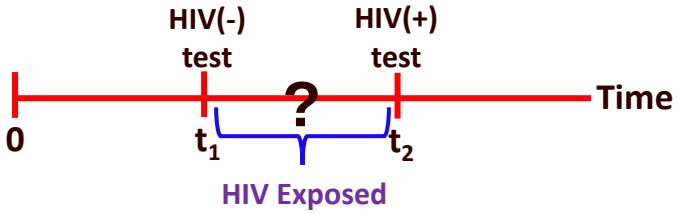
- **Right censored**: true survival time is equal to or greater than observed survival time



- **Left censored**: true survival time is less than or equal to the observed survival time



- **Interval censored**: true survival time is within a known time interval

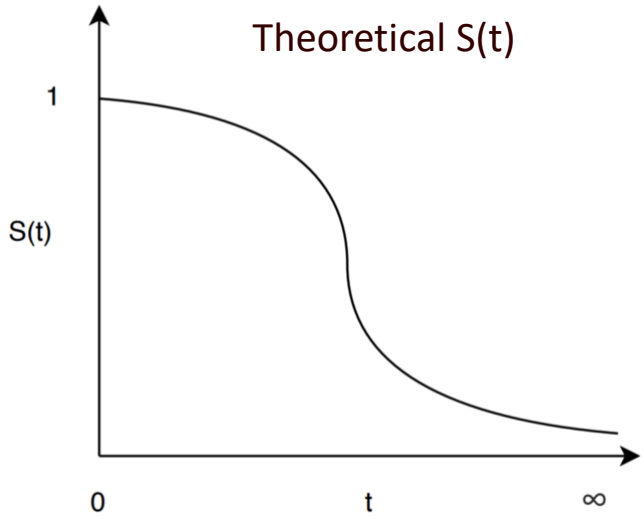


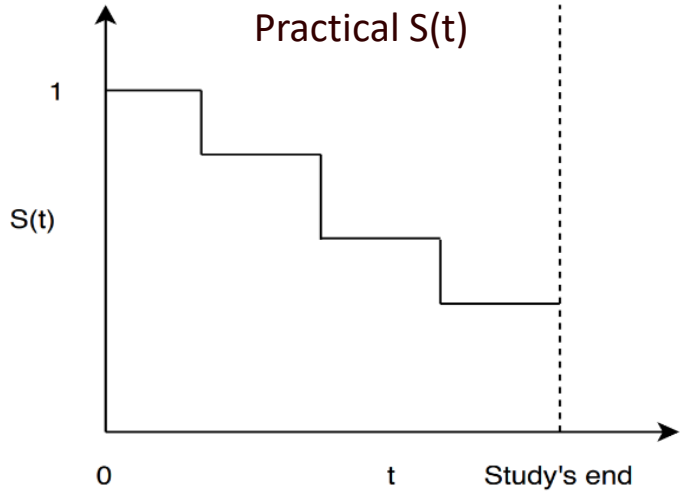
What is a Survival Function?

Survival Function is the probability that an individual survives beyond a specific time T.

It has the following properties:

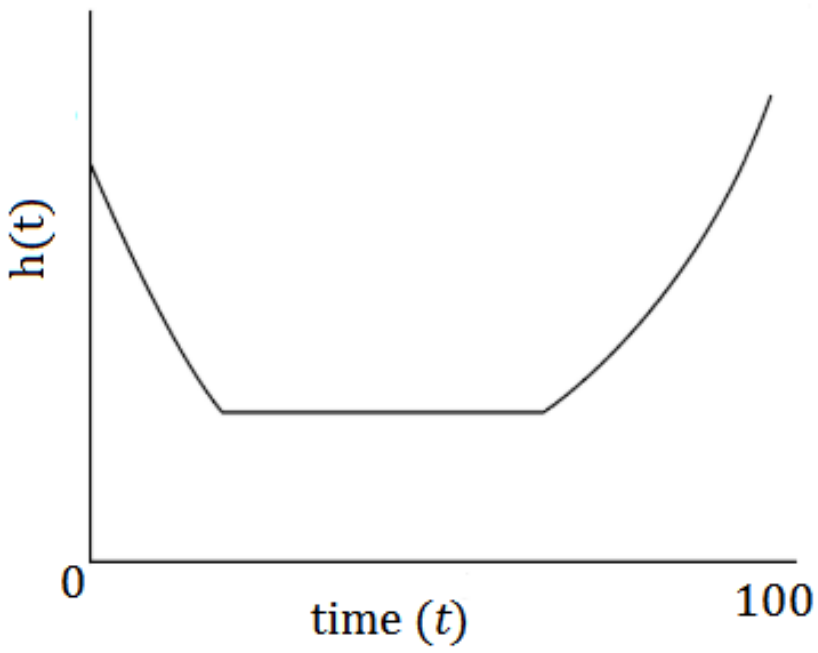
- It is non-increasing
- At time t = 0, S(t) = 1. In other words, the probability of surviving past time 0 is 1
- At time t = ∞, S(t) = S(∞) = 0. As time goes to infinity, the survival curve goes to 0
- In theory, the survival function is smooth
- In practice, events are observed on a discrete time scale





What is a Hazard Function?

The Hazard Function is defined as the instantaneous risk that the event of interest happens, within a very narrow time frame.



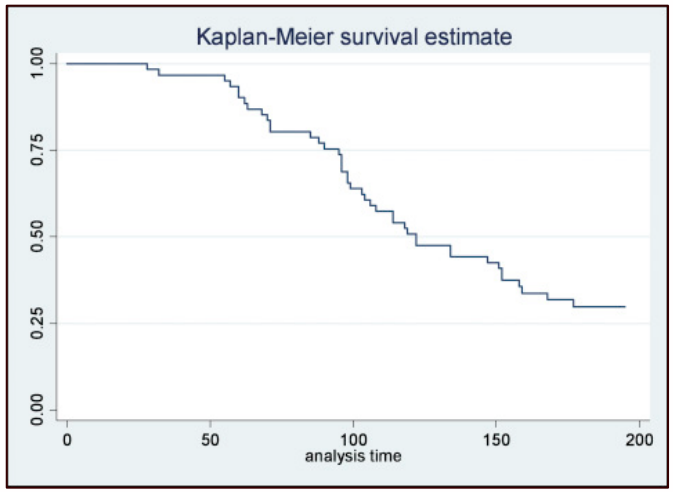
The hazard function h(t) showing the chances of death for a human at any particular age.

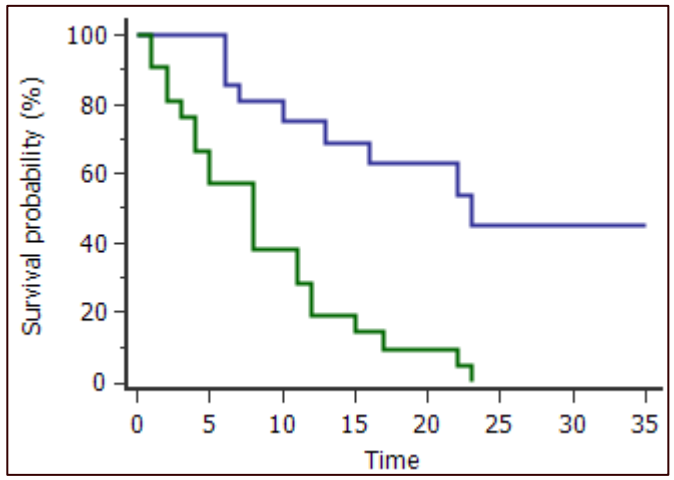
Survival Analysis – methods

- **Non-parametric estimation**
  - Within-group survival: **Kaplan-Meier**
  - Between-group comparison: **Log-rank Test**
- **Semi-parametric estimation model**
  - Cox proportional hazard model (allows **explanatory** variables)
- **Parametric estimation model**
  - Exponential
  - Weibull
  - Gamma

Kaplan-Meier Survival Method

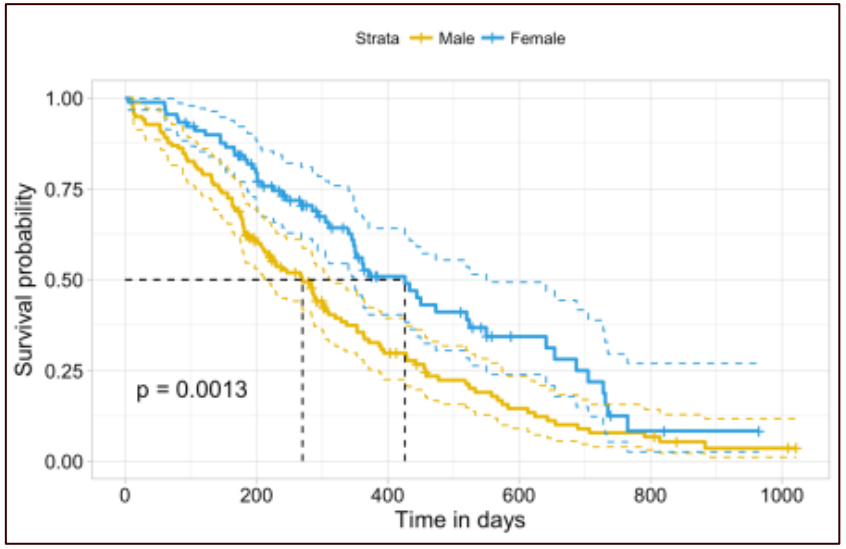
- The Kaplan-Meier (KM) method is a non-parametric method
- Commonly use to estimate the survival probability from observed survival times
- Intuitive graphical presentation
- Cumulative survival characteristics
- Estimation of median survival time
- Commonly use to compare two study populations





Comparison of groups – Log-rank Test

- The log-rank test is the most widely used method of comparing two or more survival curves
- The null hypothesis is that there is no difference in survival between the two groups
- The log rank test is a non-parametric test, which makes no assumptions about the survival distributions



Limitations of KM Curves and Log-Rank Tests

- We can only test one variable at a time
  - We cannot control for potential confounders
  - We cannot control for potential clustering in the data
  - We cannot control for other potential risk factors
  - We cannot include interaction terms
- The log-rank test only provides an estimate of the weight of evidence that the strata are different in their risk, not the magnitude of the difference
- We can not handle continuous exposure variables

Cox - Regression model

The Cox proportional hazard model provides the following benefits:

- ❑ Adjusts for multiple risk factors simultaneously.
- ❑ Allows quantitative (continuous) risk factors, helping to limit the number of strata.
- ❑ Provides estimates and confidence intervals of how the risk changes across the strata and across unit increases in quantitative variables.
- ❑ Can handle data sets with right censoring, staggered entry, etc.; so long as we have adequate data at each time point.

Cox - Regression model

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_n x_{in})$$

$h_i(t)$ is the hazard function for individual *i*

$h_0(t)$ is the baseline hazard function and can take any form
It is estimated from the data (non parametric)

$x_{i1}, x_{i2}, ..., x_{in}$ are the covariates

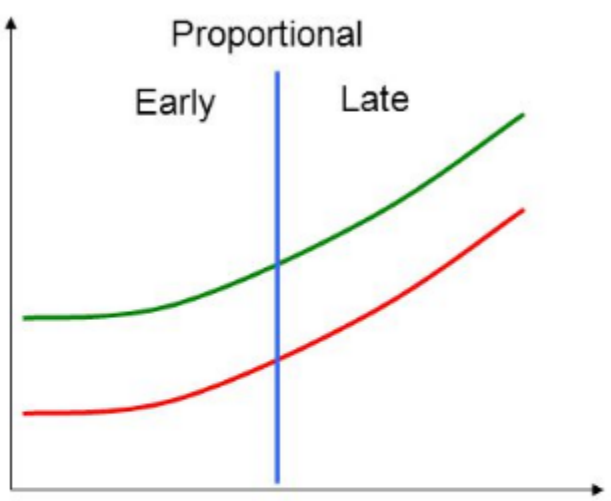
$\beta_1, \beta_2, ..., \beta_n$ are the regression coefficients estimated from the data

Effect of covariates is constant over time (parameterised)
This is the proportional hazards assumption

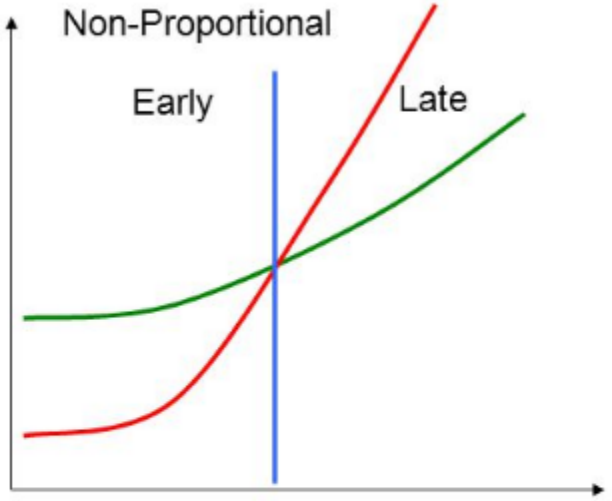
Therefore, Cox regression referred to as a semi-parametric model

Assumptions of Cox-Regression

Assumption 1: The survival curves for two different strata of a risk factor must have hazard functions that are proportional over time



Here, the effect is the same in both time periods



Here, the effect is negative in the early period and positive in the late period

Assumptions of Cox-Regression

Assumption 2: Independent observations

- This assumption means that there is no relationship between the subjects in your data set and that information about one subject's survival does not in any way inform the estimated survival of any other subject.
- This is a key assumption in most statistical models.

Assumption 3: Non-informative or Independent censoring

- This assumption is satisfied when there is no relationship between the probability of censoring and the event of interest.

Checking the Assumptions

The proportional hazards assumption is checked in three main ways:

- Graphical examination of KM curves to confirm they do not cross.
- Graphical examination of log(-log(survival)) versus log(survival time) to confirm the curves are roughly parallel.
- Including time dependent covariates in the model to test for significance. Time dependent covariates take the form of interaction terms between log(time) and the covariate.

Checking the Assumptions

The independent observations assumption:

- This assumption is validated by implementing good experimental design and sampling

The independent censoring assumption:

- This assumption is mainly checked by thinking carefully about the nature of the censoring process and how it is related to the event of interest
- Examples of violations are:
  - Very sick patients are likely to transfer to a different health system.
  - Relatively healthy patients are likely to be unmotivated to complete the study.

Model building with Cox-Regression

Now build a cox-regression model using the following example dataset.

ID	age	ndrugtx	treat	site	time	censor	herco
1	39	1	1	0	188	1	3
2	33	8	1	0	26	1	3
3	33	3	1	0	207	1	2
4	32	1	0	0	144	1	3

Where,

- The variable **time** contains the time until return to drug use
- The **censor** variable indicates whether the subject returned to drug use
  - **censor**=1 indicates return to drug use and **censor**=0 otherwise
- The variable **age** indicates age at enrollment
- The **ndrugtx** variable indicates the number of previous drug treatments
- The **trear** variable indicates two different residential treatment programs that differed in length
  - **treat**=0 is the short program and **treat**=1 is the long program
- The variable **site** indicates the patients were randomly assigned to two different sites (**site=0** is site A and **site=1** is site B)
- **herco** indicates heroin or cocaine use in the past three months (**herco**=1 indicates heroin and cocaine use, **herco**=2 indicates either heroin or cocaine use and **herco**=3 indicates neither heroin nor cocaine use)

Model building with Cox-Regression

Open dataset:

```
use https://stats.idre.ucla.edu/stat/data/uis.dta, clear
```

Declare data to be survival-time data:

```
stset time, failure(censor)
```

```
     failure event:  censor != 0 & censor < .
obs. time interval:  (0, time]
 exit on or before:  failure

--------------------------------------------------------------------------------
        628  total observations
          0  exclusions
--------------------------------------------------------------------------------
        628  observations remaining, representing
        508  failures in single-record/single-failure data
    147,394  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =     1,172
```

Model building with Cox-Regression

Exploring the data: Univariate Analyses

Log-rank test and Kaplan-Meier curve: treat variable

```
sts test treat, logrank
```

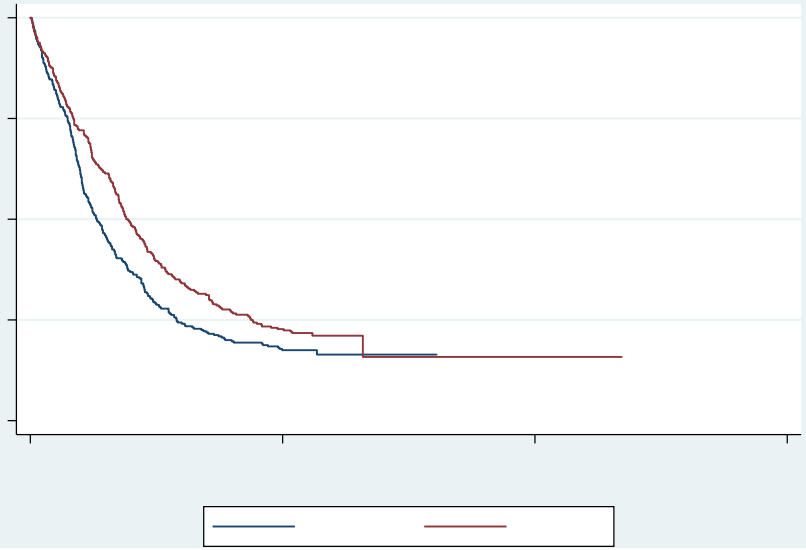
```
        failure _d:  censor
  analysis time _t:  time


Log-rank test for equality of survivor functions

           |  Events       Events
treat      |  observed     expected
-----------+------------------------
0          |       265       235.80
1          |       243       272.20
-----------+------------------------
Total      |       508       508.00

           chi2(1) =        6.80
           Pr>chi2 =      0.0091
```

```
sts graph, by(treat)
```



Model building with Cox-Regression

Exploring the data: Univariate Analyses

Log-rank test and Kaplan-Meier curve: predictor **site**

```
sts test site, logrank
```

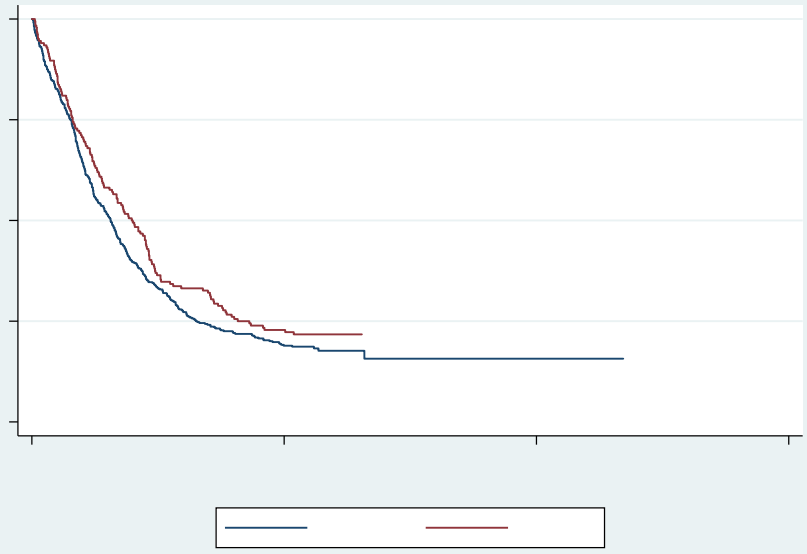
```
        failure _d:  censor
  analysis time _t:  time


Log-rank test for equality of survivor functions

        |  Events       Events
site    |  observed     expected
--------+-------------------------
0       |       364       347.94
1       |       144       160.06
--------+-------------------------
Total   |       508       508.00

          chi2(1) =        2.37
          Pr>chi2 =      0.1240
```

```
sts graph, by(site)
```



Model building with Cox-Regression

Exploring the data: Univariate Analyses

Log-rank test and Kaplan-Meier curve: predictor **herco**

`sts test herco, logrank`

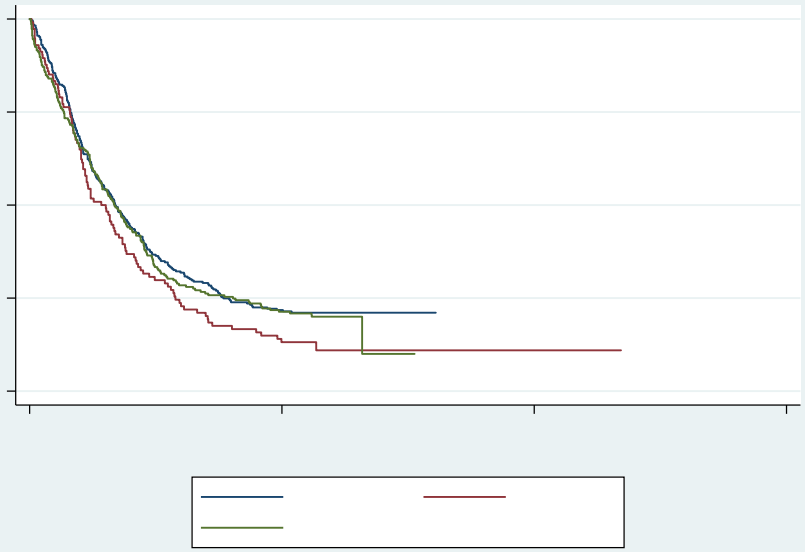
```
        failure _d:  censor
  analysis time _t:  time


Log-rank test for equality of survivor functions

       |  Events       Events
herco  |  observed     expected
-------+--------------------------
1      |       228       242.14
2      |       100        84.19
3      |       180       181.67
-------+--------------------------
Total  |       508       508.00

           chi2(2) =        3.83
           Pr>chi2 =      0.1473
```

`sts graph, by(herco)`



Model building with Cox-Regression

Exploring the data: Univariate Analyses

Cox proportional hazard model: continuous predictors **ndrugtx and age**

`stcox ndrugtx`

```
. stcox ndrugtx

        failure _d:  censor
  analysis time _t:  time

Iteration 0:   log likelihood = -2874.9717
Iteration 1:   log likelihood = -2868.7559
Iteration 2:   log likelihood = -2868.3002
Iteration 3:   log likelihood =  -2868.299
Refining estimates:
Iteration 0:   log likelihood =  -2868.299

Cox regression -- Breslow method for ties

No. of subjects =          611                 Number of obs    =          611
No. of failures =          496
Time at risk    =       143002
                                               LR chi2(1)       =        13.35
Log likelihood  =    -2868.299                 Prob > chi2      =       0.0003

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ndrugtx |   1.029808   .0077214     3.92   0.000     1.014785    1.045053
------------------------------------------------------------------------------
```

`stcox age`

```
. stcox age

        failure _d:  censor
  analysis time _t:  time

Iteration 0:   log likelihood = -2933.1124
Iteration 1:   log likelihood = -2931.4933
Iteration 2:   log likelihood = -2931.4929
Refining estimates:
Iteration 0:   log likelihood = -2931.4929

Cox regression -- Breslow method for ties

No. of subjects =          623                 Number of obs    =          623
No. of failures =          504
Time at risk    =       146816
                                               LR chi2(1)       =         3.24
Log likelihood  =   -2931.4929                 Prob > chi2      =       0.0719

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9872183   .0070969    -1.79   0.074     .9734062    1.001226
------------------------------------------------------------------------------
```

Model building with Cox-Regression

The final model and interpretation of the hazard ratios

```
stcox age ndrugtx i.treat i.site
```

```
No. of subjects =          610                 Number of obs    =          610
No. of failures =          495
Time at risk    =       142994
                                               LR chi2(4)       =        30.64
Log likelihood  =   -2853.2371                 Prob > chi2      =       0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9781141	.0073465	-2.95	0.003	.9638208	.9926194
ndrugtx	1.035645	.0079409	4.57	0.000	1.020198	1.051327
1.treat	.7837396	.0709607	-2.69	0.007	.656301	.9359239
1.site	.8450728	.0848554	-1.68	0.094	.6941021	1.02888

Model building with Cox-Regression

The final model and interpretation of the hazard ratios

```
stcox age ndrugtx i.treat i.site
```

```
No. of subjects =          610                 Number of obs    =          610
No. of failures =          495
Time at risk    =       142994
                                               LR chi2(4)       =        30.64
Log likelihood  =   -2853.2371                 Prob > chi2      =       0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9781141   .0073465    -2.95   0.003     .9638208    .9926194
------------------------------------------------------------------------------
```

The hazard ratio indicates that as the enrollment age increases by one unit, and all other variables are held constant, the rate of relapse decreases by (100%-97.8%)=2.2%.

Model building with Cox-Regression

The final model and interpretation of the hazard ratios

```
stcox age ndrugtx i.treat i.site
```

```
No. of subjects =          610                 Number of obs    =          610
No. of failures =          495
Time at risk    =       142994
                                               LR chi2(4)       =        30.64
Log likelihood  =   -2853.2371                 Prob > chi2      =       0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ndrugtx |   1.035645   .0079409     4.57   0.000     1.020198    1.051327
------------------------------------------------------------------------------
```

The hazard ratio indicates that as the number of previous drug treatment (**ndrugtx**) increases by one unit, and all other variables are held constant, the rate of relapse increases by 3.6%.

Model building with Cox-Regression

The final model and interpretation of the hazard ratios

```
stcox age ndrugtx i.treat i.site
```

```
No. of subjects =          610                 Number of obs    =          610
No. of failures =          495
Time at risk    =       142994
                                               LR chi2(4)       =        30.64
Log likelihood  =   -2853.2371                 Prob > chi2      =       0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.treat |   .7837396   .0709607    -2.69   0.007      .656301    .9359239
------------------------------------------------------------------------------
```

If the treatment length is altered from short to long, while holding all other variables constant, the rate of relapse decreases by (100% – 78.4%) = 21.6%.

Model building with Cox-Regression

The final model and interpretation of the hazard ratios

```
stcox age ndrugtx i.treat i.site
```

```
No. of subjects =          610                 Number of obs    =          610
No. of failures =          495
Time at risk    =       142994
                                               LR chi2(4)       =        30.64
Log likelihood  =   -2853.2371                 Prob > chi2      =       0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.site |   .8450728   .0848554    -1.68   0.094     .6941021    1.02888
------------------------------------------------------------------------------
```

As treatment is moved from site A to site B, and all other variables are held constant, the rate of relapse decreases by (100% – 84.5%) = 15.5%.

Model building with Cox-Regression

Is the model ok?

No. of subjects =	610		Number of obs	=	610
No. of failures =	495				
Time at risk =	142994				
			LR chi2(4)	=	30.64
Log likelihood =	-2853.2371		Prob > chi2	=	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.9781141	.0073465	-2.95	0.003	.9638208	.9926194
ndrugtx	1.035645	.0079409	4.57	0.000	1.020198	1.051327
1.treat	.7837396	.0709607	-2.69	0.007	.656301	.9359239
1.site	.8450728	.0848554	-1.68	0.094	.6941021	1.02888

Model building with Cox-Regression

Is the model ok?

We don’t know without checking those assumptions

Checking the Assumptions

Proportionality Assumption

```
stcox age ndrugtx i.treat i.site, nohr tvc(age ndrugtx treat site) texp(ln(_t))
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
main						
age	-.0253885	.0340036	-0.75	0.455	-.0920344	.0412574
ndrugtx	.0189201	.0319939	0.59	0.554	-.0437868	.0816271
1.treat	-.6606696	.4113948	-1.61	0.108	-1.466989	.1456494
1.site	-.5808523	.4705265	-1.23	0.217	-1.503067	.3413627
tvc						
age	.0006074	.007115	0.09	0.932	-.0133378	.0145526
ndrugtx	.0036057	.0069162	0.52	0.602	-.0099499	.0171613
treat	.0898218	.0862938	1.04	0.298	-.0793109	.2589545
site	.0886499	.0979659	0.90	0.366	-.1033597	.2806596

All of the time-dependent variables are not significant thus supporting the assumption of proportional hazard.

Checking the Assumptions

Proportionality Assumption check by using the Schoenfeld and scaled Schoenfeld residuals

```
quietly stcox age ndrugtx treat site, schoenfeld(sch*) scaledsch(sca*)
stphtest, detail
```

Test of proportional-hazards assumption

Time: Time

	rho	chi2	df	Prob>chi2
age	0.01506	0.11	1	0.7408
ndrugtx	0.02309	0.25	1	0.6165
treat	0.08108	3.34	1	0.0675
site	0.02758	0.39	1	0.5320
global test		3.99	4	0.4074

Since the tests in the table are not significance (p-values over 0.05) then we can not reject proportionality and we assume that we do not have a violation of the proportional assumption

Model building with Cox-Regression

Is the model ok?

Yes, now we rely this model. Because all of the assumptions are satisfied.

Thank you

icddr,b thanks its core donors for their on-going support

Government of the People's Republic of Bangladesh

Canada

SWEDEN

UKaid from the British people