Survival analysis

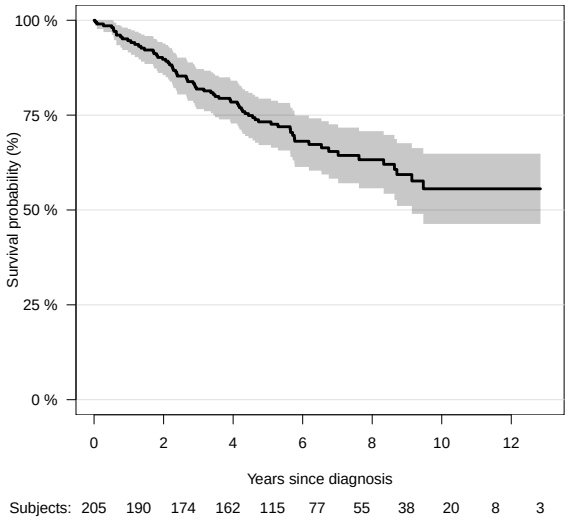
Thomas Alexander Gerds

Department of Biostatistics, University of Copenhagen

Overview

1. Survival analysis and censored data
2. Kaplan-Meier theater
3. The Cox proportional hazard regression model

The Edward L. Kaplan - Paul Meier (1958) plot



Survival and event times

Time between initiation (the time origin) and failure (or other event).

Examples

- Time to death
- Time to relapse or death
- Time to pregnancy
- Time to CVD related death

Characteristic: *right censoring*. For some individuals the event time is not known: the patient was lost to follow-up or the event had not yet happened at the end of the study period.

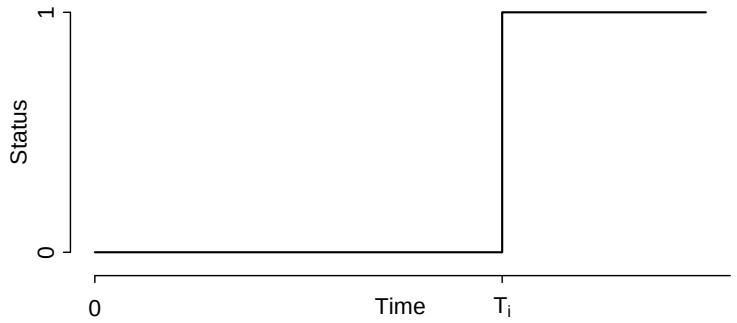
Event time data

Survival outcome consists of a time and a status variable.

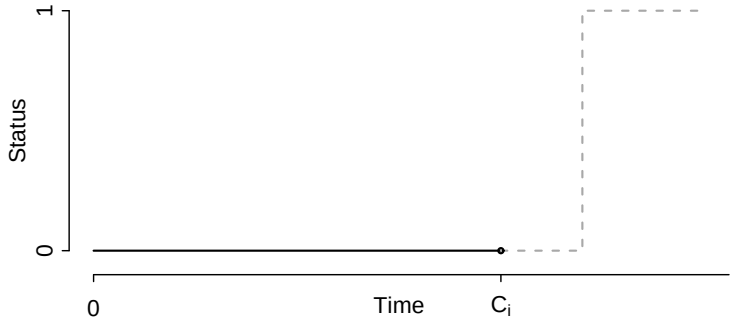
time	Status	Event
35	0	censored
872	1	death (malignant melanoma)
1860	1	death (other causes)
2521	0	censored
2565	1	death (malignant melanoma)
3185	0	censored
3776	0	censored

Here combined event: all cause mortality

Uncensored observations



Right censored observations



Typical research questions and corresponding parameters

What are the survival chances?

- Kaplan-Meier estimate of the survival function

Do treatment, exposure history, phenotype, and genotype change the survival chances?

- Cox regression: hazard rate ratios

Can we predict the survival chances of a new patient?

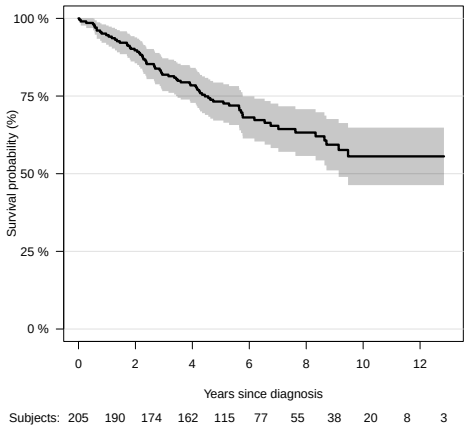
- Prediction model, e.g., derived from Cox regression

Kaplan-Meier theater

- Discover what goes wrong with simple statistics applied naively to censored survival times
- Count events and numbers at risk
- Compute the Kaplan-Meier estimate
- Discover that and how the censored observations enter into the statistic
- Note the assumptions, limitations and interpretation of the Kaplan-Meier method

Kaplan-Meier plot in R

```
km <- prodlim(Hist(Time,Status)~1,data=Melanoma)
plot(km,xlim=c(0,13),
     xlab="Years since diagnosis",
     ylab="Survival probability (%)",
     axis2.las=2)
```



Stratified Kaplan-Meier

```
kmsex <- prodlim(Hist(Time,Status)~sex,data=Melanoma)
plot(kmsex,logrank=TRUE,xlab="Years since diagnosis")
```

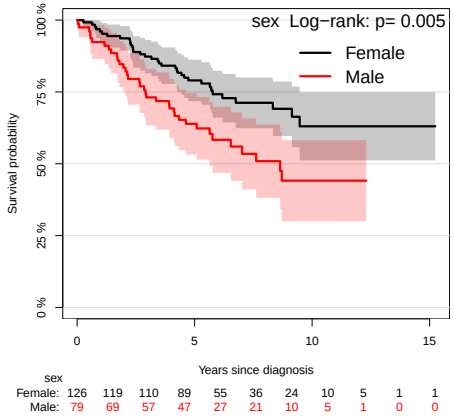
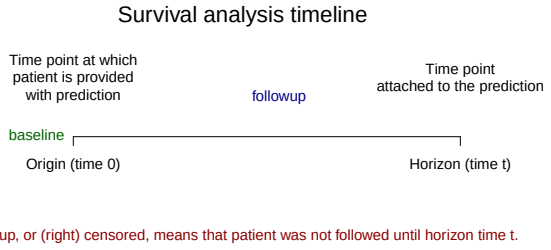


Figure: The log-rank test corresponds to Cox regression with a single categorical explanatory variable

The role of time



Lost to followup, or (right) censored, means that patient was not followed until horizon time t.

Until time t, two (or three) things can happen:

- patient is event-free
- the event of interest has occurred
- (a competing event has occurred)

The status at time t remains *unknown* for all subjects lost to follow-up before time t (event free at the last contact).

This is not a baseline table!

Table 1. **Baseline Characteristics of Participants With and Without Incident AF**

Characteristic	Entire Cohort (n = 1260)	Without AF (n = 917)	Incident AF (n = 343)	P Value*
Median age (IQR), *y*	71 (68–75)	70 (68–74)	71 (68–75)	0.002
Female, *n (%)*	691 (55)	519 (57)	172 (50)	0.041
White, *n (%)*	1200 (95)	873 (95)	327 (95)	0.92
Mean BMI (SD), kg/m^2	26.7 (4.1)	26.6 (4.1)	26.8 (4.2)	0.50
Hypertension, *n (%)*	686 (55)	476 (52)	210 (61)	0.003
Diabetes, *n (%)*	186 (15)	125 (14)	61 (18)	0.066
Heart failure, *n (%)*	31 (2)	16 (2)	15 (4)	0.007
Coronary disease, *n (%)*	245 (19)	161 (18)	84 (25)	0.006
Myocardial infarction, *n (%)*	132 (10)	84 (9)	48 (14)	0.013
Mean PR interval (SD), *ms*	171 (31)	170 (29)	174 (35)	0.036
Median PAC count (IQR), *beats/h*	2.5 (0.8–9.5)	1.8 (0.6–6.1)	5.3 (2.1–18.0)	<0.001

AF = atrial fibrillation; BMI = body mass index; IQR = interquartile range; PAC = premature atrial contraction.
* For the comparison of the indicated characteristic in participants with vs. those without incident AF.

Toy example

sex	age	time	Event	AF	status
female	65	3	Dead	No	0
male	80	4	Censored	No	0
female	50	4	Censored	No	0
male	56	6	AF	Yes	1
male	91	7	AF	Yes	1
male	60	7	Dead	No	0
female	91	8	Dead	No	0
male	87	9	AF	Yes	1
male	55	11	AF	Yes	1
male	78	14	Censored	No	0

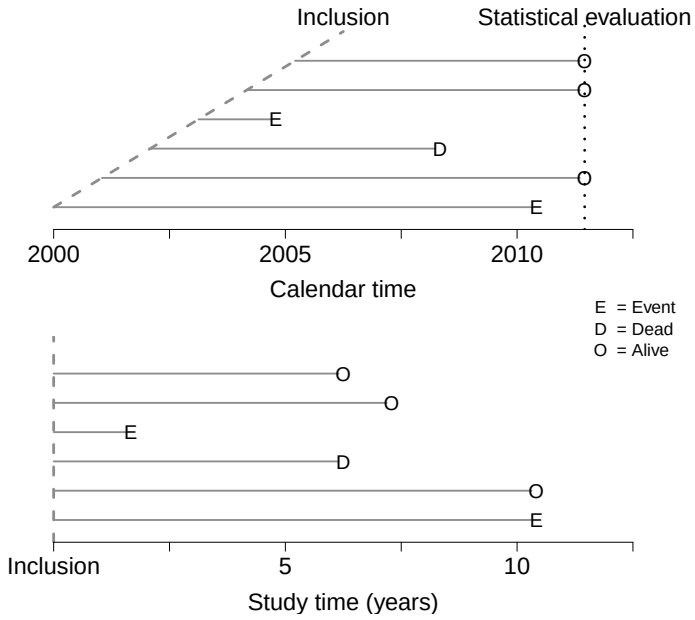
Toy example

sex	age	time	Event	AF	status
female	65	3	Dead	No	0
male	80	4	Censored	No	0
female	50	4	Censored	No	0
male	56	6	AF	Yes	1
male	91	7	AF	Yes	1
male	60	7	Dead	No	0
female	91	8	Dead	No	0
male	87	9	AF	Yes	1
male	55	11	AF	Yes	1
male	78	14	Censored	No	0

Baseline table?

Variable	Level	Incident AF (n=4)	Without AF (n=6)
age	mean (sd)	72.2 (19.4)	70.7 (15.0)
sex	female	0 (0.0)	3 (50.0)
	male	4 (100.0)	3 (50.0)

Time on study



Followup tables

after 5 years

Variable	Level	Incident AF	Without AF	unknown	Event-free
n		0	1	2	7
age	mean (sd)	NaN (NA)	65.0 (NA)	65.0 (21.2)	74.0 (16.6)
sex	female	0 (0.0)	1 (100.0)	1 (50.0)	1 (14.3)
	male	0 (0.0)	0 (0.0)	1 (50.0)	6 (85.7)

10 years

Variable	Level	Incident AF	Without AF	unknown	Event-free
n		3	3	2	2
age	mean (sd)	78.0 (19.2)	72.0 (16.6)	65.0 (21.2)	66.5 (16.3)
sex	female	0 (0.0)	2 (66.7)	1 (50.0)	0 (0.0)
	male	3 (100.0)	1 (33.3)	1 (50.0)	2 (100.0)

Speed and hazard

Speed is the rate at which an object covers distance.

- A fast-moving object has a high speed and covers a relatively large distance in a given amount of time
- A slow-moving object covers a relatively small amount of distance in the same amount of time.

Speed and hazard

Speed is the rate at which an object covers distance.

- A fast-moving object has a high speed and covers a relatively large distance in a given amount of time
- A slow-moving object covers a relatively small amount of distance in the same amount of time.

Hazard rate is the speed at which a person gets a disease or dies.

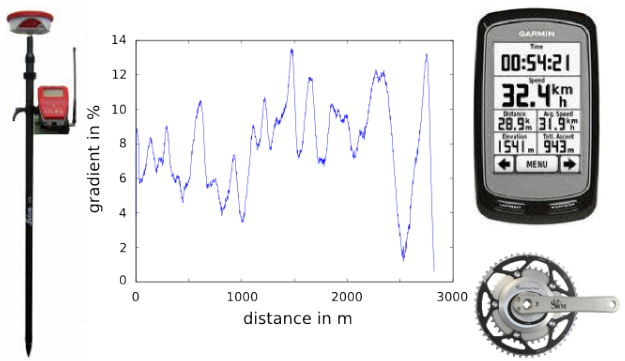
- An exposed person has a high hazard rate and will get diseased with a relatively large probability within a given time period.
- A non-exposed person has a low hazard rate and will get diseased with a relatively small probability within the same time period.

Speed and duration

If we know the (average) speed of a cyclist on the road from place A to place B we can calculate the (expected) duration that the cyclist would need to cycle from A to B.

Speed and duration

If we know the (average) speed of a cyclist on the road from place A to place B we can calculate the (expected) duration that the cyclist would need to cycle from A to B.



The speed changes along the road according to gradient and other things, the duration is a function of the speed and the length of the road.

Hazard and absolute risk

If we know the (average) hazard rate of a person in a time interval between date A and date B we can calculate the (expected) probability that the person would have a stroke between the dates A and B.

Hazard and absolute risk

If we know the (average) hazard rate of a person in a time interval between date A and date B we can calculate the (expected) probability that the person would have a stroke between the dates A and B.

The hazard rate changes over time according to predisposition, exposure and disease, the absolute risk of stroke is a function of the hazard rate and the length of the time interval.

Computing duration

Suppose the speed of a cyclist is piece-wise constant on the road between A and B:

Section	Distance (km)	Speed (km/h)	Duration (min)
A –> a	10	15	40.0
a –> b	20	20	60.0
b –> B	20	22	54.5

- Total distance: 50 km
- Total duration: 154.5 minutes

Conclusion: The cyclist needs 154.5 minutes to cycle from A to B.

Computing absolute risk

Suppose the stroke hazard rate of a patient is piece-wise constant in the time period between A and B:

Interval	Duration (mth)	Hazard rate (%/mth)	Abs. risk (%)
A –> a	10	1.5	13.9
a –> b	20	2.0	33.0
b –> B	20	2.2	35.6

Calculation of Absolute risk: 13.9% = 1-exp(-0.015*10)

- Total duration: 50 months
- Total absolute risk: 62.8%

Calculation of total: 1- (1-0.139)(1-0.33)(1-0.356) = 0.628

Conclusion: The patient will experience a stroke with a probability of 62.8% in the period between date A and date B.

Effect of doping

Suppose the same cyclist as before knows a way to increase the speed by 10%:

Section	Distance (km)	Speed (km/h)	Duration (min)
A –> a	10	16.5	36.4
a –> b	20	22.0	54.5
b –> B	20	24.2	49.6

- Total duration without doping: 154.5 minutes
- Total duration with doping: 140.5 minutes

Conclusion: With doping the cyclist needs 14 minutes (9.1%) less to cycle from A to B.

Effect of aging

Suppose the same patient as before is one year older and this increases the hazard rate of stroke by 10% (hazard ratio=1.1):

Effect of aging

Suppose the same patient as before is one year older and this increases the hazard rate of stroke by 10% (hazard ratio=1.1):

Interval	Duration (mth)	Hazard rate (%/mth)	Abs. risk (%)
A –> a	10	1.65	15.2
a –> b	20	2.20	35.6
b –> B	20	2.42	38.4

- Total absolute risk previous age: 62.8%
- Total absolute risk one year older: 66.4%

calculation: 1- (1-0.152)(1-0.356)(1-0.384) = 0.664

Conclusion: Each year of age increases the risk of stroke in the period between date A and date B by 3.6 %.

MORGAM project (Asplund et al.)

Relative Risks for Stroke by Age, Sex, and Population Based on Follow-Up of 18 European Populations in the MORGAM Project

Kjell Asplund, MD, PhD; Juha Karvanen, DSc(Tech); Simona Giampaoli, MD; Pekka Jousilahti, MD, PhD; Matti Niemelä, MD; Grazyna Broda, MD; Giancarlo Cesana, MD; Jean Dallongeville, MD; Pierre Ducimetriere, MD; Alun Evans, MD; Jean Ferrières, MD; Bernadette Haas, MD; Torben Jorgensen, MD; Abdonas Tamosiunas, MD; Diego Vanuzzo, MD; Per-Gunnar Wiklund, MD, PhD; John Yarnell, MD; Kari Kuulasmaa, PhD; Sangita Kulathinal, PhD; for the MORGAM Project

Background and Purpose—Within the framework of the MOnica Risk, Genetics, Archiving and Monograph (MORGAM) Project, the variations in impact of classical risk factors of stroke by population, sex, and age were analyzed.

Methods—Follow-up data were collected in 43 cohorts in 18 populations in 8 European countries surveyed for cardiovascular risk factors. In 93 695 persons aged 19 to 77 years and free of major cardiovascular disease at baseline, total observation years were 1 234 252 and the number of stroke events analyzed was 3142. Hazard ratios were calculated by Cox regression analyses.

Results—Each year of age increased the risk of stroke (fatal and nonfatal together) by 9% (95% CI, 9% to 10%) in men and by 10% (9% to 10%) in women. A 10-mm Hg increase in systolic blood pressure involved a similar increase in risk in men (28%; 24% to 32%) and women (25%; 20% to 29%). Smoking conferred a similar excess risk in women (104%; 78% to 133%) and in men (82%; 66% to 100%). The effect of increasing body mass index was very modest. Higher high-density lipoprotein cholesterol levels decreased the risk of stroke more in women (hazard ratio per mmol/L 0.58; 0.49 to 0.68) than in men (0.80; 0.69 to 0.92). The impact of the individual risk factors differed somewhat between countries/regions with high blood pressure being particularly important in central Europe (Poland and Lithuania).

Conclusions—Age, sex, and region-specific estimates of relative risks for stroke conferred by classical risk factors in various regions of Europe are provided. From a public health perspective, an important lesson is that smoking confers a high risk for stroke across Europe. **(*Stroke*. 2009;40:2319-2326.)**

Key Words: blood pressure ■ cholesterol ■ cohort studies ■ smoking ■ stroke risk factors

The appraisal of stroke risk in populations or individuals is based on the recognition that all cardiovascular disorders are multifactorial in nature. The most widely used risk score statistical techniques.[1] The Framingham stroke risk score has been used extensively when international and national guide-lines for cardiovascular prevention have been developed

Discussion

Table 2. Comparisons of HRs and their 95% Confidence Limits Between Calculations Based on All Strokes (Fatal+Nonfatal) and Fatal Strokes Only*

	HR (95% Confidence Limits)	
	Fatal and Nonfatal Strokes	Fatal Strokes Only
Men (N=51 703)		
No. of events	1851	545
Age per year	1.09 (1.09–1.10)	1.12 (1.10–1.13)
Blood pressure† per 10 mm Hg	1.28 (1.24–1.32)	1.38 (1.31–1.45)
BMI per unit	1.02 (1.01–1.03)	1.03 (1.00–1.05)
Smoking, yes/no	1.82 (1.66–2.00)	2.00 (1.68–2.38)
HDL cholesterol per mmol/L	0.80 (0.69–0.92)	1.01 (0.78–1.31)
Women (N=41 992)		
No. of events	1291	418
Age per year	1.10 (1.09–1.10)	1.13 (1.11–1.14)
Blood pressure† per 10 mm Hg	1.25 (1.20–1.29)	1.26 (1.19–1.34)
BMI per unit	1.00 (0.99–1.01)	0.99 (0.97–1.01)
Smoking (yes/no)	2.04 (1.78–2.33)	2.56 (2.01–3.24)
HDL cholesterol per mmol/L	0.58 (0.49–0.68)	0.48 (0.36–0.65)

tional risk factors on stroke. Different factors included in the multivariate models may also contribute to the discrepancies. For instance, when we in the MORGAM study adjusted for HDL cholesterol levels, nonlow-density lipoprotein cholesterol no longer remained a significant predictor of stroke.

It should be emphasized that the present results refer only to relative risks for stroke. These relative risks do not necessarily translate into risks of stroke in absolute terms. In the stroke component of the World Health Organization MONICA Project, which included a large number of popu-lations from high-income, middle-income, and low-income countries, differences in population levels of conventional risk factors have explained only a modest part of the variation in stroke incidence rates in cross-sectional comparisons.[28]

Error: a hazard ratio is not a relative risk

The Cox proportional hazard regression model

Cox model

$$\text{hazard rate}(t, X_{i1}, \ldots, X_{iK}) = \alpha_0(t)\, \exp(\beta_1 X_{i1} + \cdots + \beta_K X_{iK})$$

- $X_i = X_{i1}, \ldots, X_{iK}$ are predictor values for individual i
- $exp(\beta_k)$ is the hazard ratio and quantifies if and how different values of X_{ik} change the instantaneous survival chances (similar as in logistic regression)
- $\alpha_0(t)$ is a baseline hazard function which may have an interpretation as the hazard for patients whose predictors are all equal to zero.

Assumption: the hazard ratios do not change over time.

Malignant melanoma data

time	Status	invasion	thick	sex	age
35	0	level.1	1.34	Male	41
621	1	level.2	7.06	Male	72
872	1	level.0	0.97	Female	65
1854	0	level.1	1.62	Female	65
1942	0	level.0	0.81	Female	35
2165	0	level.1	5.64	Male	62
2256	1	level.1	2.26	Female	43
2521	0	level.0	1.29	Female	45
2565	1	level.1	3.54	Male	34
3185	0	level.0	0.48	Female	64
3776	0	level.2	7.09	Male	12
4479	0	level.0	1.13	Female	19

Cox regression in R

```
cfit = coxph(Surv(time, Status) ~ age+sex+invasion+thick, data
    = Melanoma)
summary(cfit)
```

```
Call:
coxph(formula = Surv(time, Status) ~ age + sex + invasion + thick,
    data = Melanoma)

  n= 205, number of events= 71

                   coef exp(coef) se(coef)     z Pr(>|z|)
age            0.018875  1.019054 0.008103 2.329  0.01984 *
sexMale        0.508403  1.662634 0.241126 2.108  0.03499 *
invasionlevel.1 0.402718  1.495886 0.297284 1.355  0.17553
invasionlevel.2 0.099005  1.104072 0.483583 0.205  0.83778
thick          0.133516  1.142840 0.044941 2.971  0.00297 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
age                1.019     0.9813    1.0030     1.035
sexMale            1.663     0.6015    1.0365     2.667
invasionlevel.1    1.496     0.6685    0.8353     2.679
invasionlevel.2    1.104     0.9057    0.4279     2.849
thick              1.143     0.8750    1.0465     1.248

Concordance= 0.712  (se = 0.036 )
Rsquare= 0.164   (max possible= 0.967 )
Likelihood ratio test= 36.61  on 5 df,   p=0.0000007177
Wald test            = 39.18  on 5 df,   p=0.0000002188
Score (logrank) test = 44.04  on 5 df,   p=0.00000002277
```


Example conclusions

```
Melanoma[,Age:=age/10]
cfit = coxph(Surv(time, Status) ~ Age+sex+invasion+thick, data
    = Melanoma)
publish(cfit,org=TRUE,digits=2)
```

Variable	Units	HazardRatio	CI.95	p-value
Age		1.21	[1.03;1.42]	0.020
sex	Female	1.00	[1.00;1.00]	1.000
	Male	1.66	[1.04;2.67]	0.035
invasion	level.0	1.00	[1.00;1.00]	1.000
	level.1	1.50	[0.84;2.68]	0.176
	level.2	1.10	[0.43;2.85]	0.838
thick		1.14	[1.05;1.25]	<0.01

- The all cause mortality is signifantly increasing with age, a 10 year age difference is associated with a 21% (3%-42%) increase of the hazard rate.
- The survival chances were significantly higher in females compared to males (HR=1.66, [1.04;2.67], p<0.05).
- The hazard rate was 66% (4%, 167%) higher in males compared to females.

Example conclusions

```
Melanoma[,Age:=age/10]
cfit = coxph(Surv(time, Status) ~ Age+sex+invasion+thick, data
    = Melanoma)
publish(cfit,org=TRUE,digits=2)
```

Variable	Units	HazardRatio	CI.95	p-value
Age		1.21	[1.03;1.42]	0.020
sex	Female	1.00	[1.00;1.00]	1.000
	Male	1.66	[1.04;2.67]	0.035
invasion	level.0	1.00	[1.00;1.00]	1.000
	level.1	1.50	[0.84;2.68]	0.176
	level.2	1.10	[0.43;2.85]	0.838
thick		1.14	[1.05;1.25]	<0.01

- The all cause mortality is signifantly increasing with age, a 10 year age difference is associated with a 21% (3%-42%) increase of the hazard rate.
- The survival chances were significantly higher in females compared to males (HR=1.66, [1.04;2.67], p<0.05).
- The hazard rate was 66% (4%, 167%) higher in males compared to females.

- The probability that a (randomly selected) men had a longer survival time than a (randomly selected) woman was 37.6% (CI=27.3%-49.1%; p=0.035).

The probabilistic index

There is a one-to-one relation between the hazard ratio and the probabilistic index:

$$\mathrm{P}(T_i < T_j \mid X_i, X_j, z, z) = \frac{1}{1 + e^{\beta(X_j - X_i)}}$$

- A hazard ratio of $e^{\beta} = 1$ corresponds to a probabilistic index of 50%.
- The results of Asplund et al., male HR=1.09; CI-95%: (1.09; 1.10) can be reformulated as:

The probability that a male patient who is one year older than an otherwise similar patient will have a longer lifetime is 47.8%, CI-95%: (47.6; 47.8).

Take home messages

- Censored data cannot be summarized with simple statistics such as mean, median, standard deviation, or linear regression.

- The Kaplan-Meier method and Cox regression assume that every subject who does not experience the event during the followup time will experience the event at a later time point.

- A hazard rate is an event probability per time unit.
- A hazard ratio is the ratio between two hazard rates.
- A hazard ratio is not a risk ratio.