Survival analysis I

Research Methods for Quantitative Data 2018

Camille Lassale
Department of Epidemiology and Population Health

c.lassale@ucl.ac.uk

Objectives

By the end of this session you should be able to:

- Explain what censoring is and give examples of why it may occur
- Interpret a Kaplan-Meier survival graph
- Test the equality of two survival curves using the log rank test

Types of outcome

Continuous	Linear regression
Binary	Logistic regression
Count data	Poisson regression
Time to event data	Survival analysis

Survival (Time to Event) data

Survival data arises in many fields of application: medical, actuarial, physical sciences, social sciences

The key requirements of survival data are that

- the time of origin is unambiguously defined (e.g. date of birth, date of surgery, date of randomisation)

- we know the scale for the passage of ‘time’ (e.g. days, months, years)

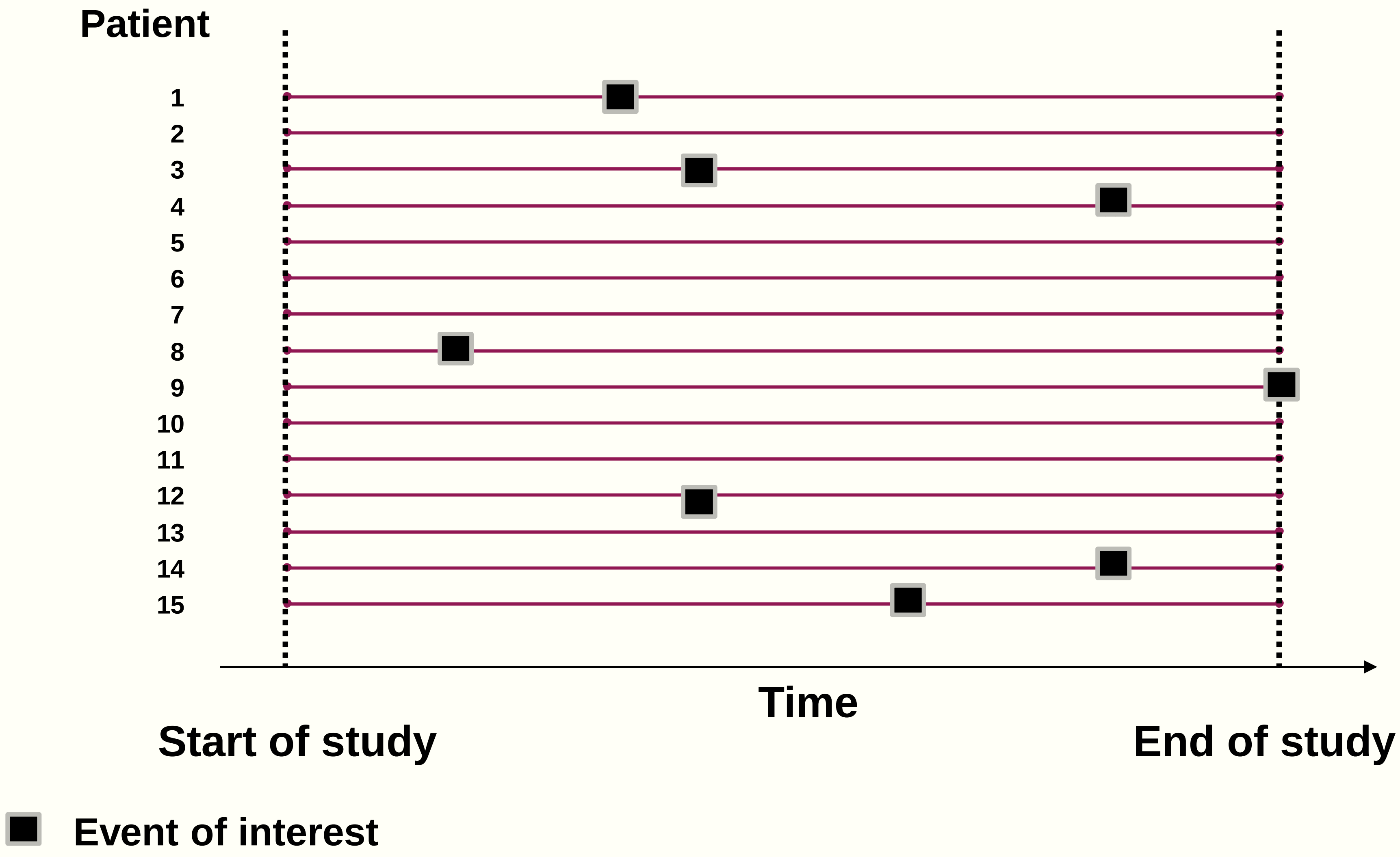
- failure is unambiguously defined (e.g. death, resolution of symptoms)

Purpose of survival analysis

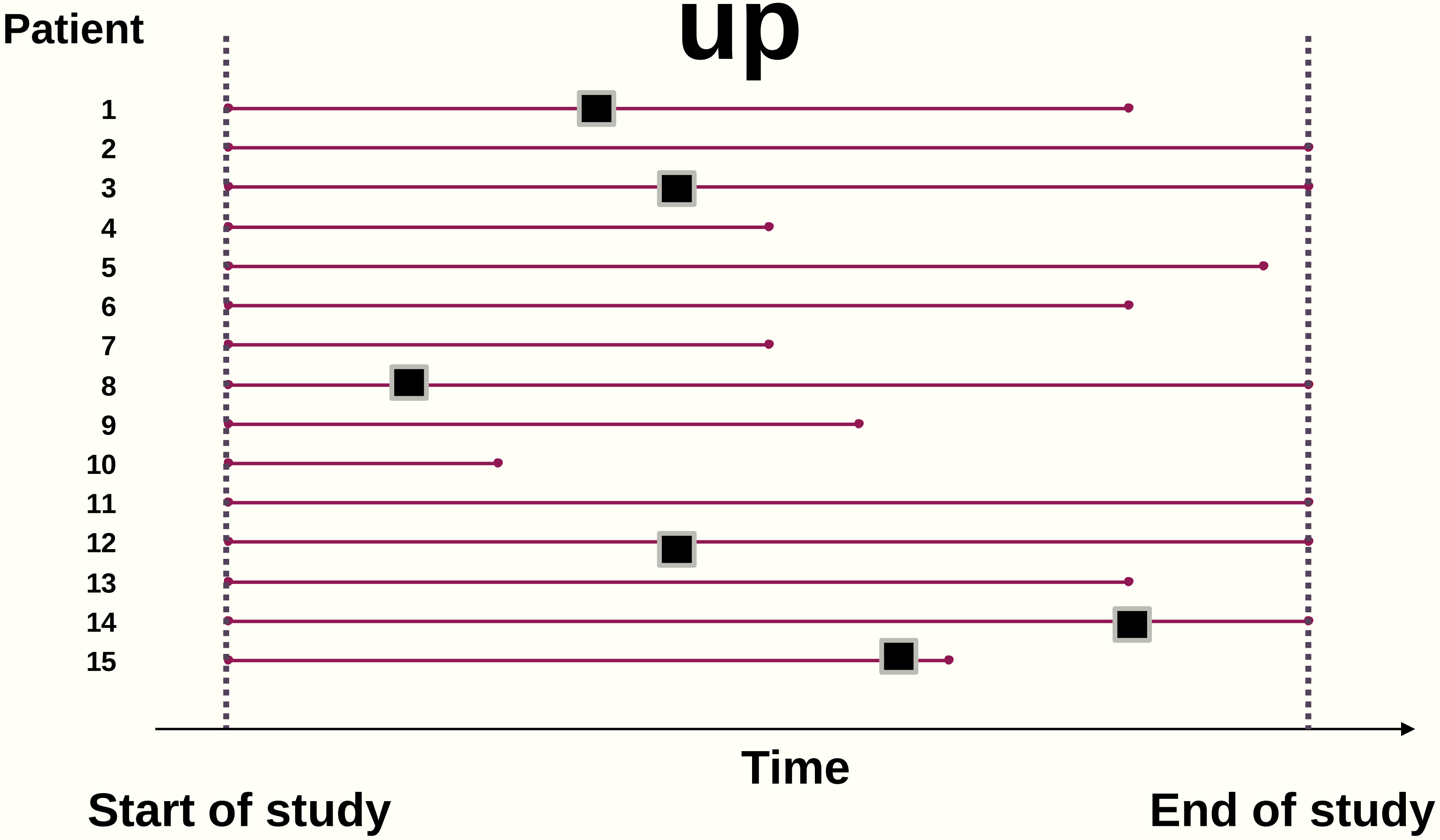
The main features of survival analysis are

- the study of distributions
- the comparison of distributions
- the effect of explanatory variables

Ideal situation – Equal follow-up



Real situation – Unequal follow-up



■ Event of interest

Differing amounts of follow-up

- Six of the patients experienced the event of interest during the study

- However, some patients were followed for longer periods than others

- Follow-up on patients who did not experience the event before dropping out of the study is **censored** – all we know is that they had not experienced the event by the time they were lost to follow-up

Why take this into account?

- Patients followed for the whole study period had a greater chance of experiencing the event, simply because they were followed for longer

- Estimates of the event incidence are likely to be underestimated if you make the assumption all patients are followed for the same length of time

- If the pattern of censoring differs between groups, there is the potential for comparisons to be seriously biased

Time to event data – Examples

- Time to death
- Time to treatment response
- Time to onset of disease
- Time to find a job (amongst the unemployed)
- Time to quit smoking (amongst smokers)

Time to event data

- Here, we analyse the **length of time** until occurrence of event

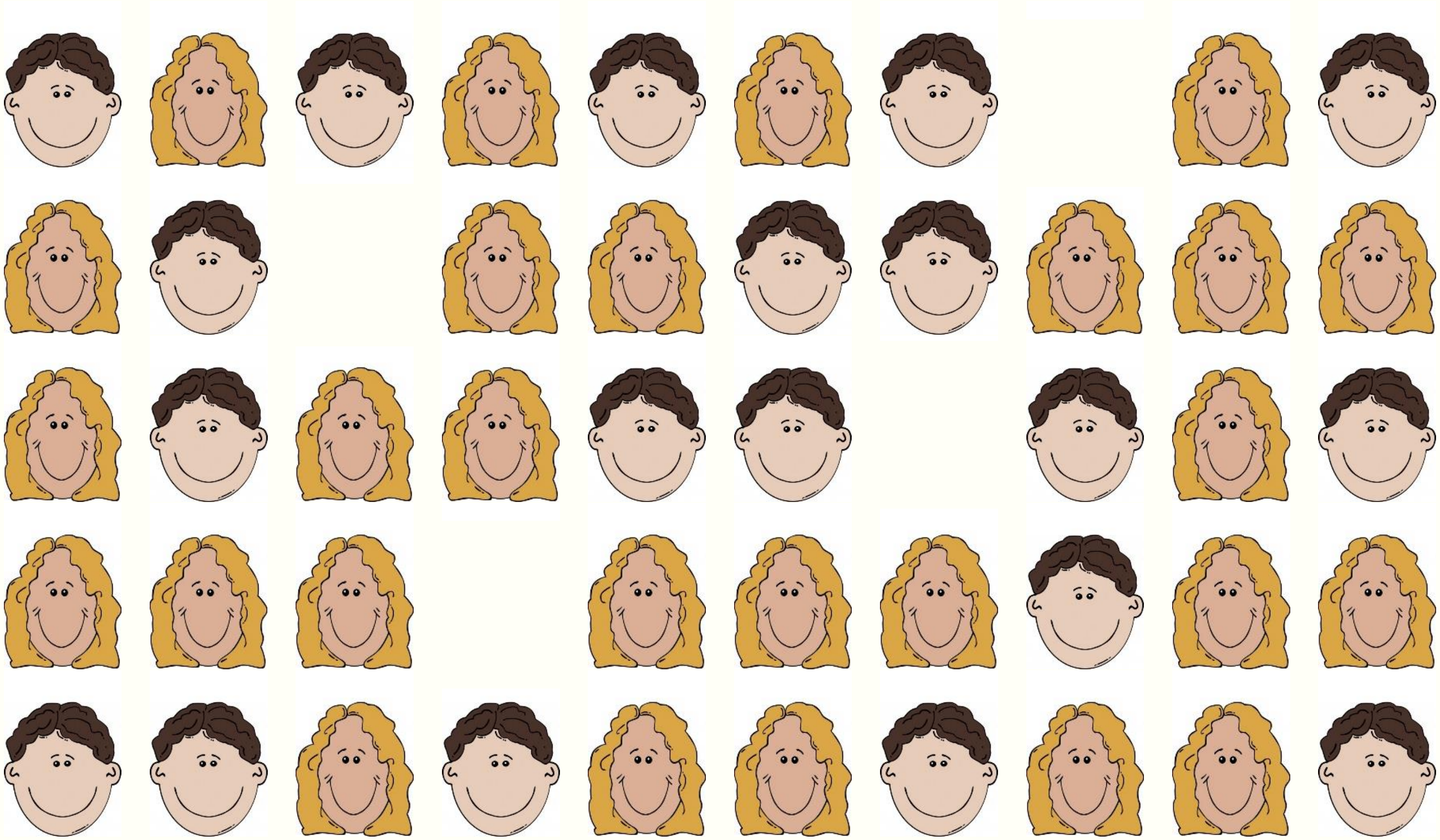
- The data are positive (i.e. the time to an event cannot be <0) and the distribution is usually skewed

- Data are usually **censored**

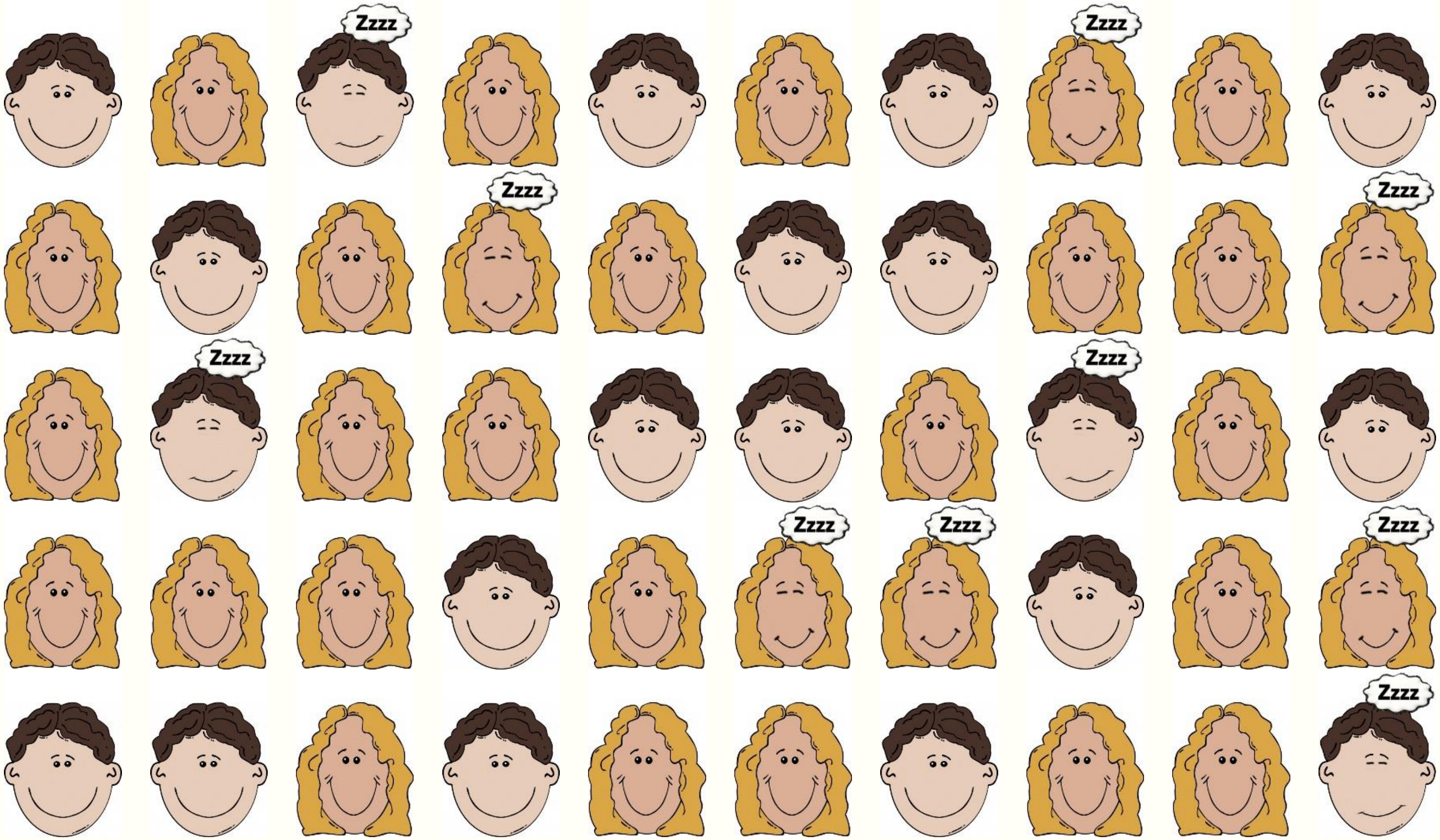
Person-time

- Sum of the time periods spent in the cohort by each individual
- Can be computed in both open and closed cohorts
- Persons contribute person-time during the time they could have developed an event that would have been counted as a case (until diagnosis, death or loss to follow-up)
- Various units of person-time
  - person-year, person-day, person-hour etc.

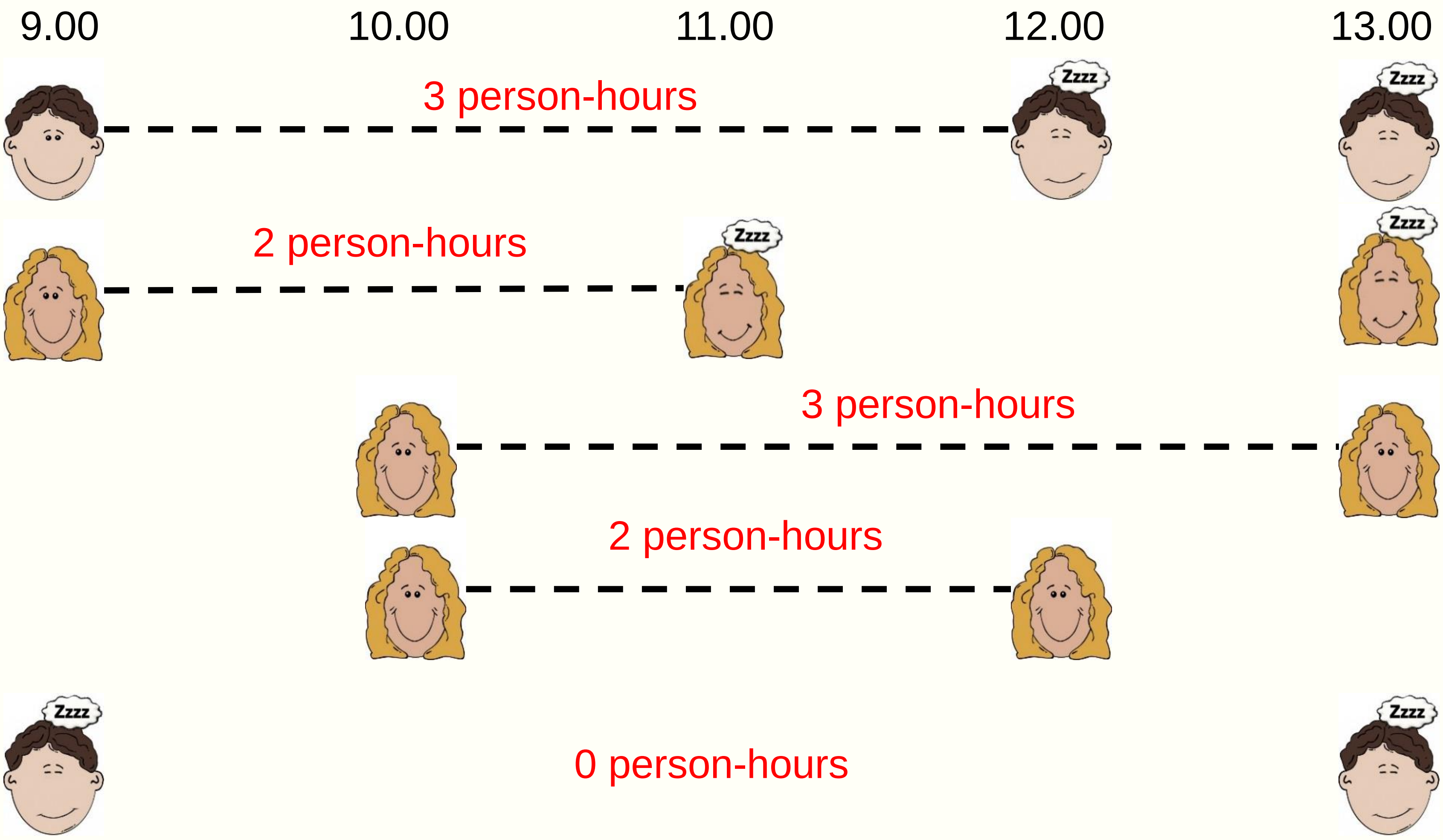
Epidemiology classroom 9.00am



Epidemiology classroom 13.00



Accrual of person-time



9.00 10.00 11.00 12.00 13.00

3 person-hours

2 person-hours

3 person-hours

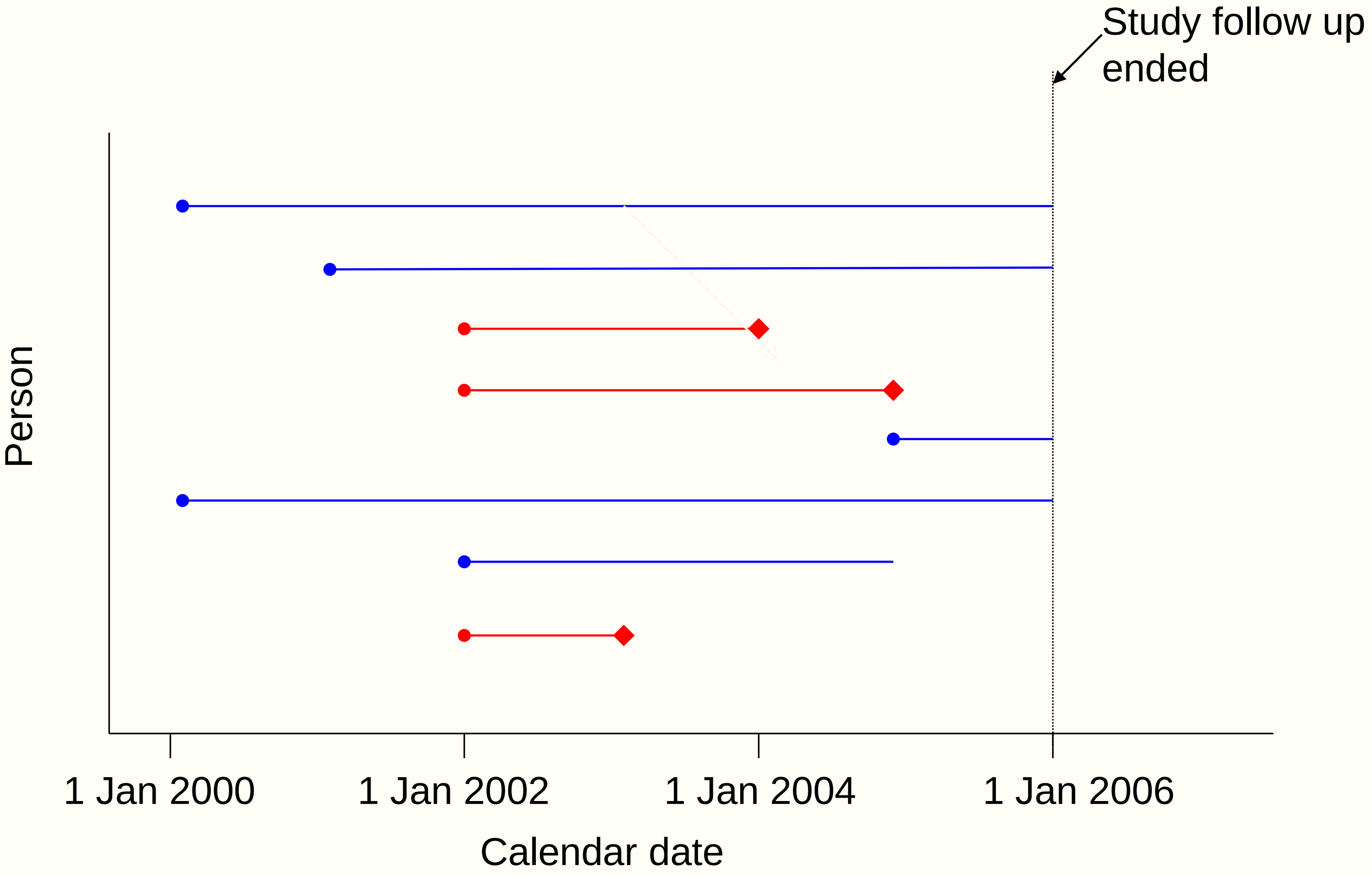
2 person-hours

0 person-hours

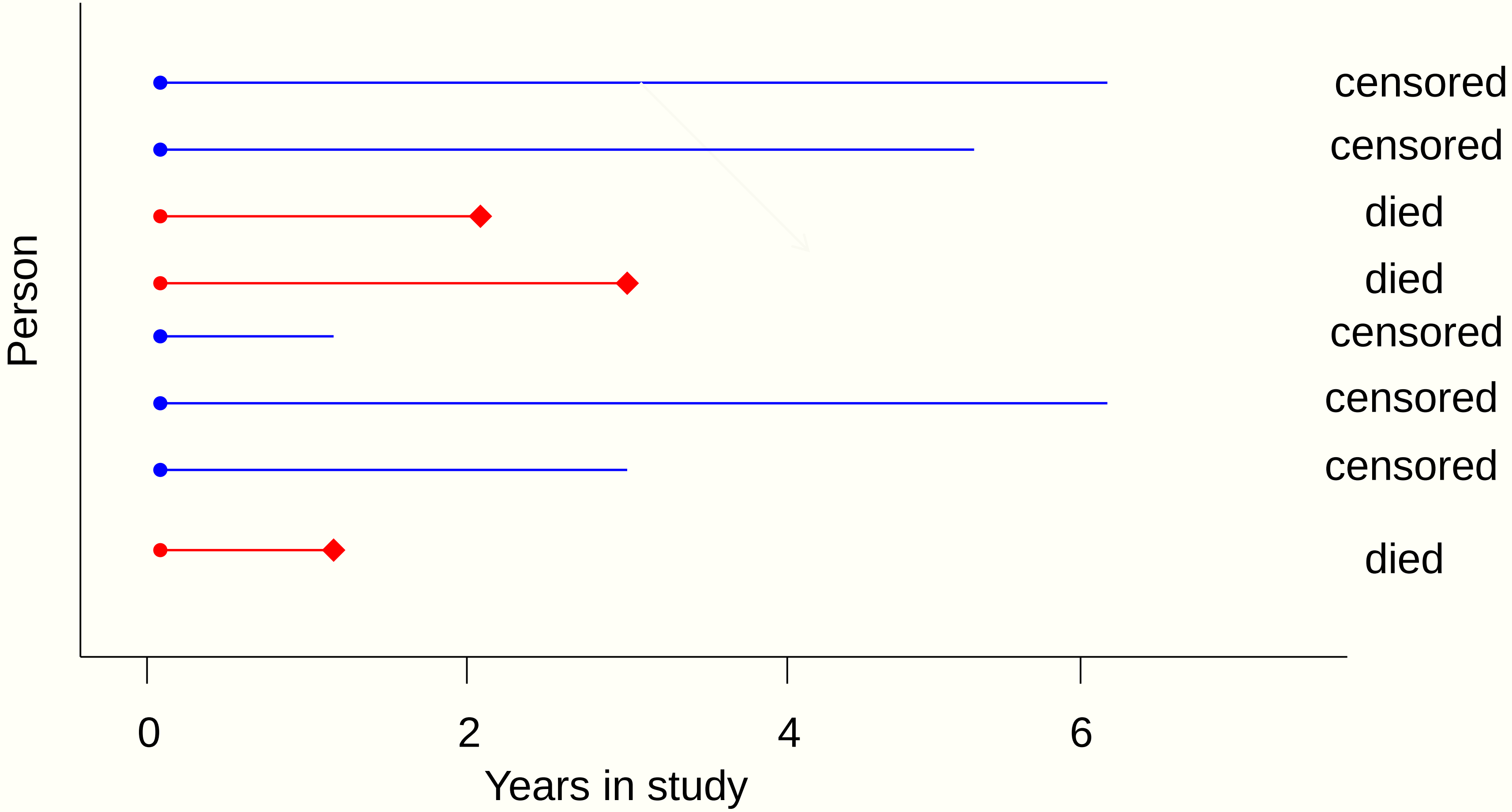
Example (more serious) data

ID	Entry_date	Death_date	End_date	time	died
1	1Jan2000	.	1Jan2006	6.0	0
2	1Jan2001	.	1Jan2006	5.0	0
3	1Jan2002	1Jan2004	1Jan2004	2.0	1
4	1Jan2002	1Jan2005	1Jan2005	3.0	1
5	1Jan2005	.	1Jan2006	1.0	0
6	1Jan2000	.	1Jan2006	6.0	0
7	1Jan2002	.	1Jan2005	3.0	0
8	1Jan2002	1Jan2003	1Jan2003	1.0	1

Calendar time



Time since entry to study



Censoring

- A censored observation is one with incomplete information

- Although different types are possible, we will only consider **right censoring**
  - subject did not have an event during the time that they were under study

- Right censoring examples:
  - Study follow-up ends
  - Person drops out of study
  - Person emigrates

Survival Data

Survival data is characterised by two variables:

1. Time variable
2. Failure/Censoring indicator:

    = 1 failed/event

    = 0 censored

Time (years)	Failure
6.4	0
4.5	1
2.5	0
2.9	0
3.8	1

Summarising time to event data

- People are followed for different lengths of time
- Therefore, we cannot calculate prevalence/ incidence risk (% people who have an event)
- Can use **rates** (death rates or incidence rates) that take account of **person-time** at risk

Summarising time to event data

Numerator = number of events (**d**)

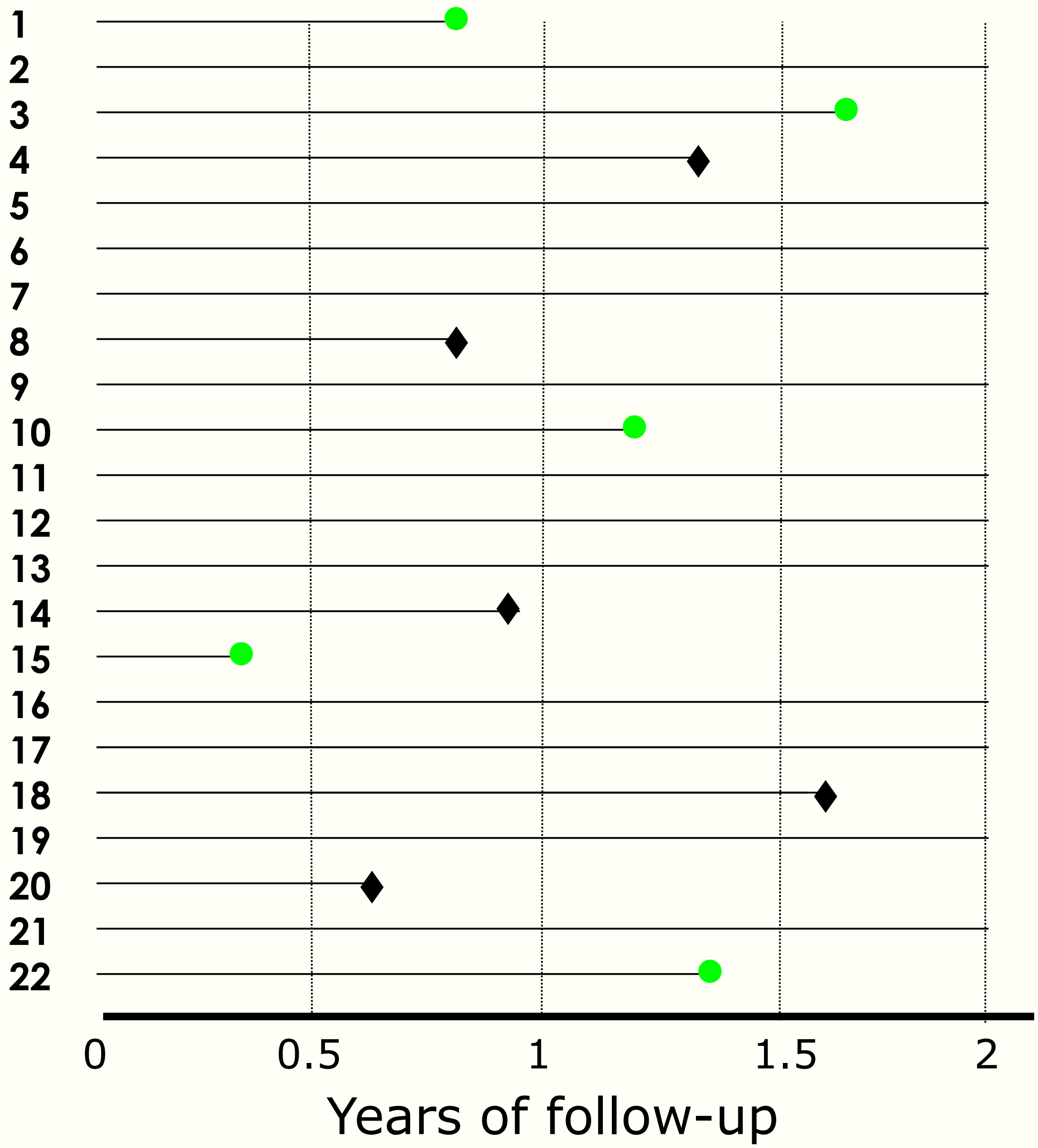
Denominator = total person-time at risk
e.g. person-years at risk (**py**)

$$\text{Rate per person-year} = \frac{\text{number of events } (\mathbf{d})}{\text{person-years at risk } (\mathbf{py})}$$

$$\text{Rate (/100 person-years)} = \frac{d}{py} * 100$$

Person-time

- Sum of the time periods spent in the cohort by each individual
- Can be computed in both open and closed cohorts
- Persons contribute person-time during the time they could have developed an event that would have been counted as a case (until diagnosis, death or loss to follow-up)
- Various units of person-time
  - person-year, person-day, person-hour etc.



Person-years at risk

0.8
2.0
1.7
1.3
2.0
2.0
2.0
0.8
2.0
1.2
2.0
2.0
2.0
0.9
0.4
2.0
2.0
1.6
2.0
0.6
2.0
1.4

34.7

♦ = event

● = censored

Rate:

$$\frac{5}{34.7}$$

=0.144 per py

=14.4 per 100 pys

The rate

- Rate may be expressed relative to any period of time (e.g. per 100 patient-years, per 1000 patient-years, per patient-month, etc), depending on frequency of event

→ Can compare rates in two groups with the **(incidence) rate ratio** (rate in group 1 divided by rate in group 2) –interpreted in a similar manner to RR

- Can calculate confidence intervals and p-values for the rate ratio

Incidence Rate Ratio (IRR)

- Incidence rate for men = 2.5 per 100 pyrs
- Incidence rate for women = 5 per 100 pyrs

- IRR (women vs. men) = 5/2.5 = 2

Kaplan-Meier plots

- The Kaplan Meier method estimates cumulative probability (percentage) experiencing an event by a certain time point

- These estimates are usually presented by constructing a Kaplan Meier plot

- Stata command
  - First declare survival time:

```
stset time,failure(event)
```

  - In this example:

```
stset survtime,fail(allcause)
```

survtime = time from beginning of study to death or end study follow up

allcause = 1 if died, 0 if censored

  - Draw graph: `sts graph`

Censored data: Kaplan-Meier estimates

Prob (dying between *t* and *t+1*)

$$= \frac{\text{no of deaths in interval}}{\text{no 'at risk' at start of interval}}$$

Prob (surviving from *t* to *t+1*) =

1-P (dying between *t* and *t+1*)

Kaplan-Meier estimator

It provides the probability of surviving to a particular time point *t* .

One should multiply all the probabilities for intervals before that time point *t* .

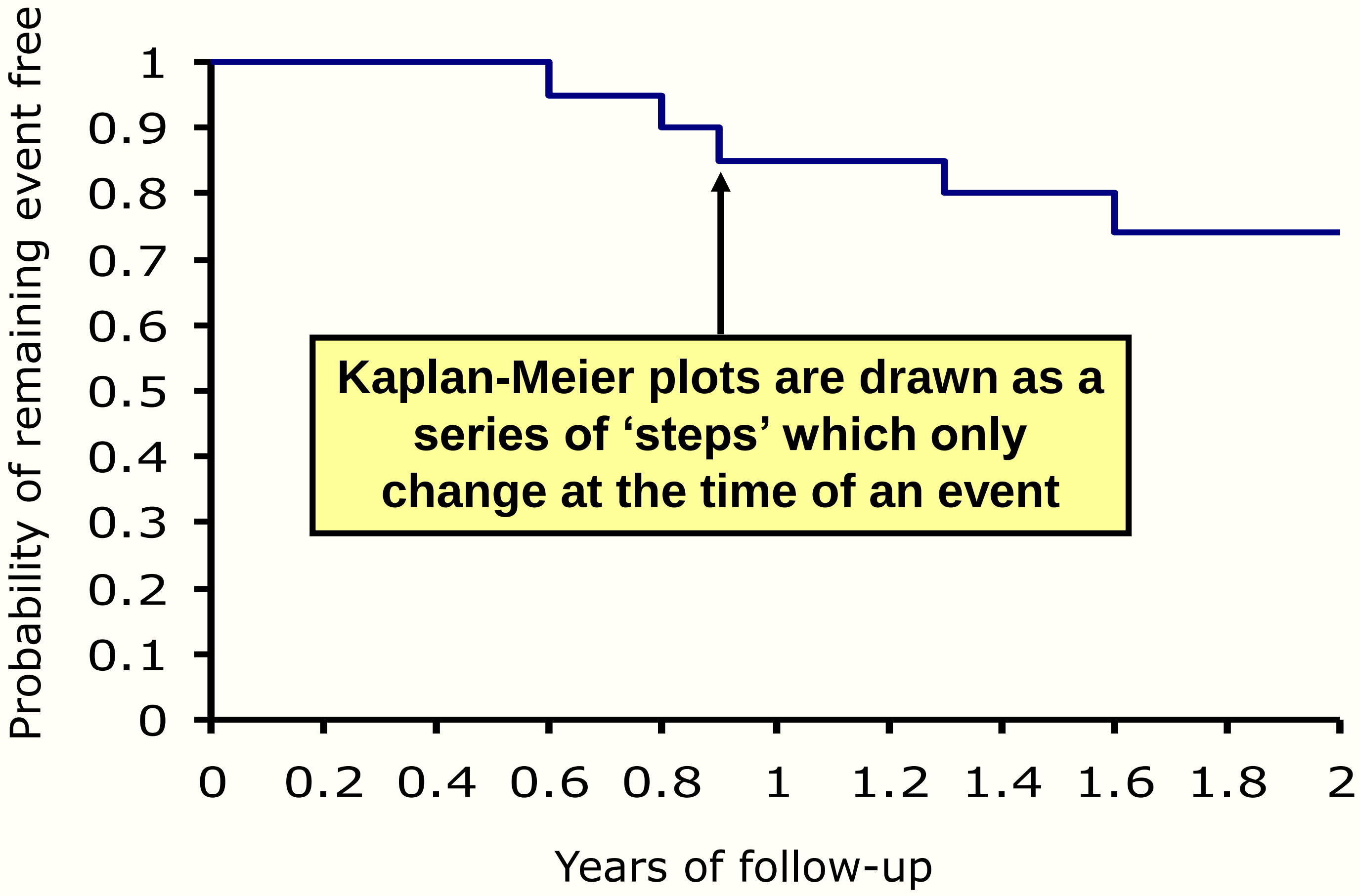
Kaplan-Meier estimates

Year	Number at risk	Number events	Number censored	Prob. event at this time	Prob. no event at this time (p(t))	Prob. remaining free of an event up to and including this time
0.1	22	0	0	0	1.00	1.00
0.2	22	0	0	0	1.00	1.00x1.00=1.00
…	…	…	…	…	…	…
0.4	22	0	1	0	1.00	1.00x1.00=1.00
…	…	…	…	…	…	…
0.6	21	1	0	1/21 = 0.0048	0.952	1.00x0.952=0.952
…	…	…	…	…	…	…
0.8	20	1	1	1/20 =0.050	0.950	0.952x0.950=0.904
0.9	18	1	0	1/18=0.056	0.944	0.904x0.944=0.853
…	…	…	…	…	…	…

First, calculate conditional probability of event occurrence

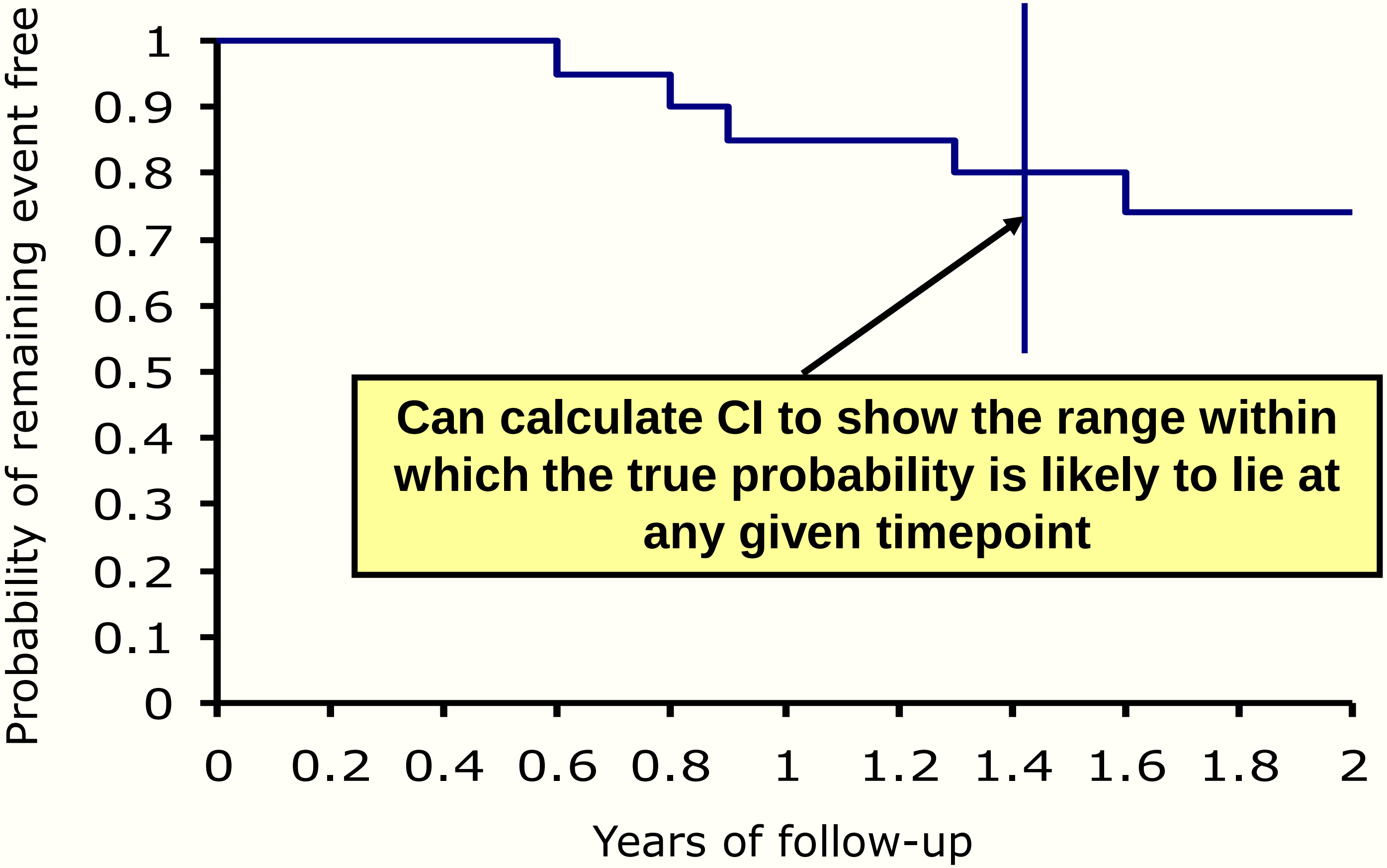
Then, multiply complements of these probabilities $1*(1-p(t_j))*\ (1-p(t_j))$

Kaplan-Meier plot



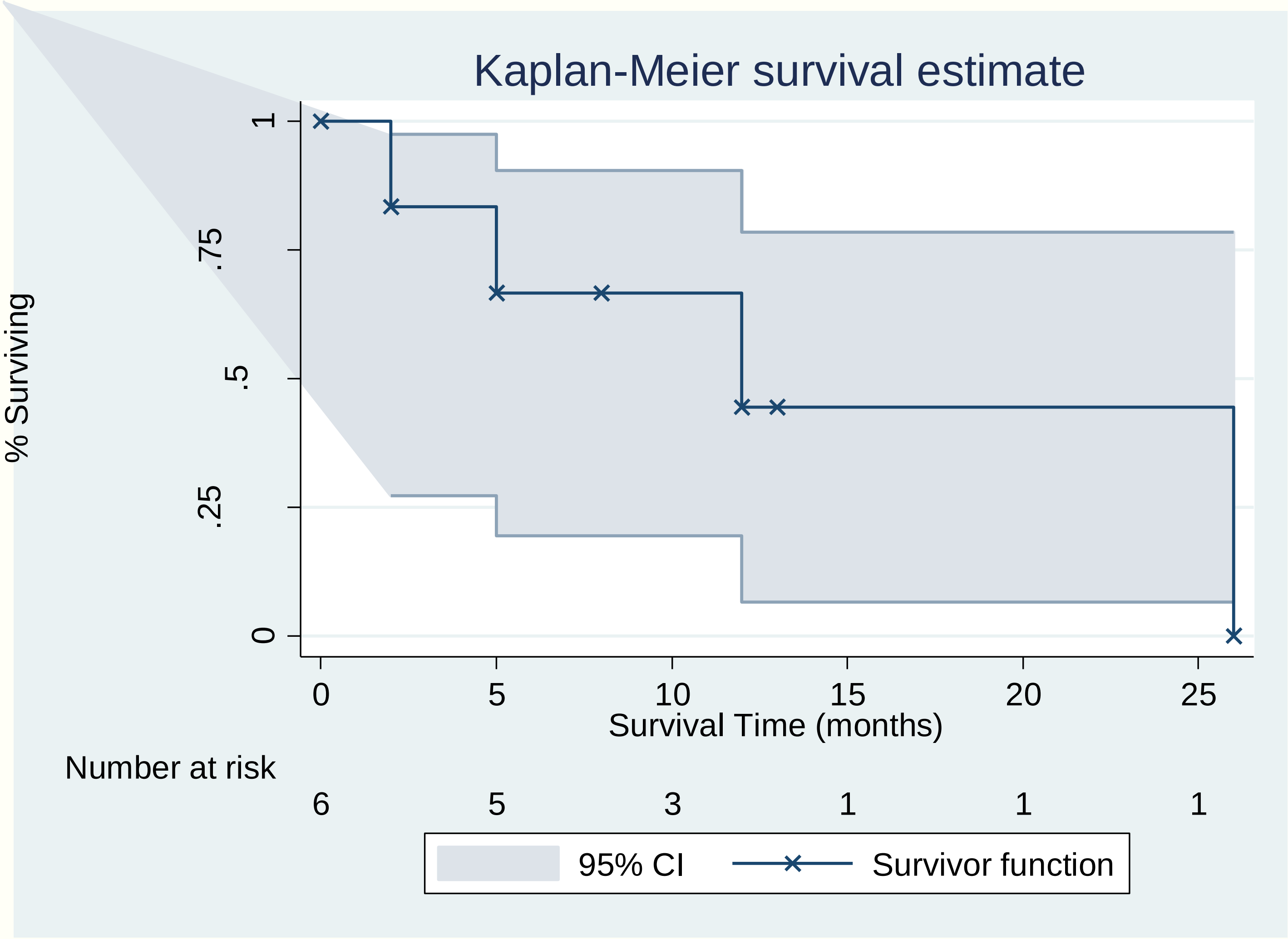
Kaplan-Meier plots are drawn as a series of ‘steps’ which only change at the time of an event

Kaplan-Meier plot

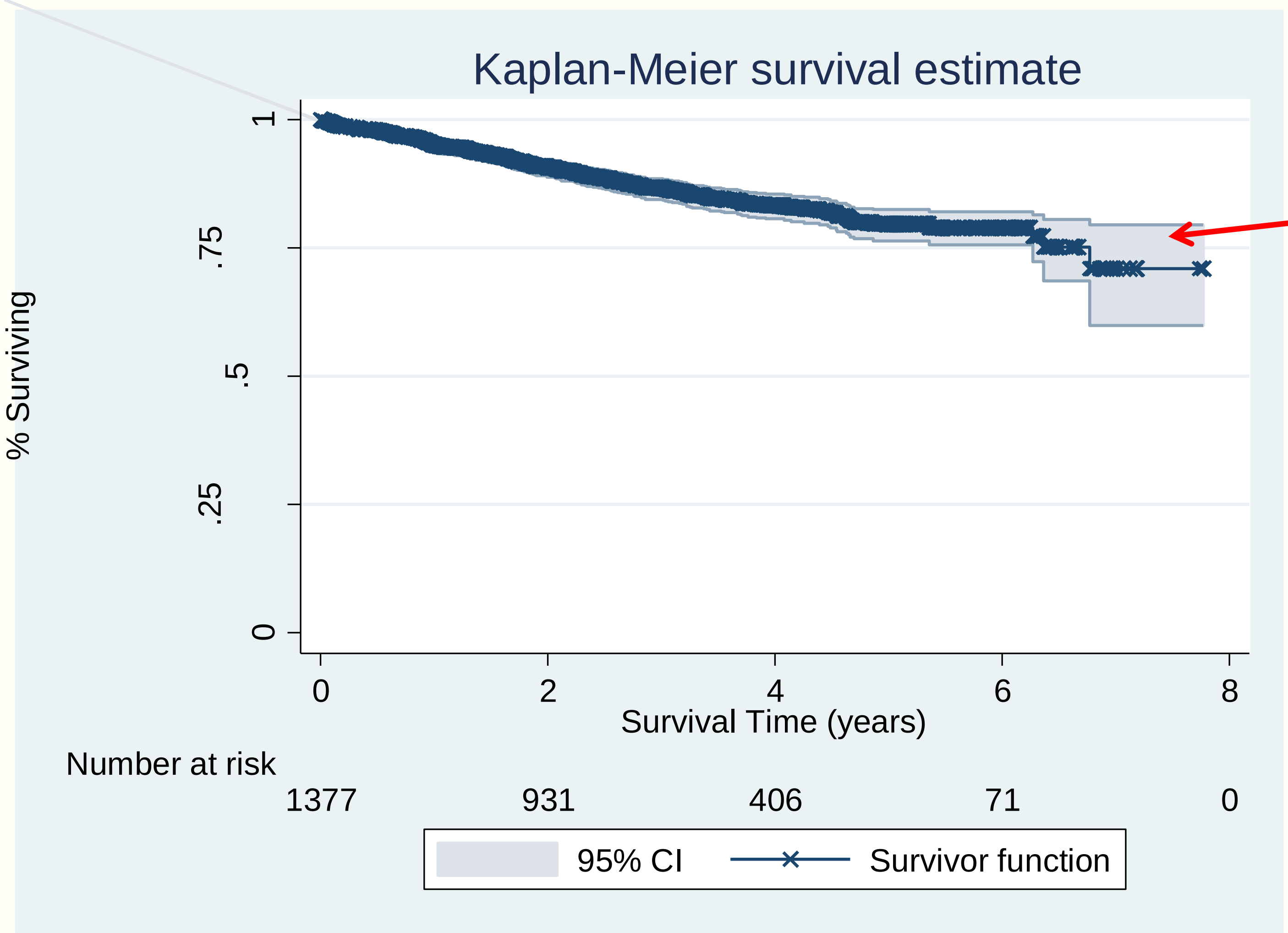


Can calculate CI to show the range within which the true probability is likely to lie at any given timepoint

Kaplan-Meier plot with confidence interval

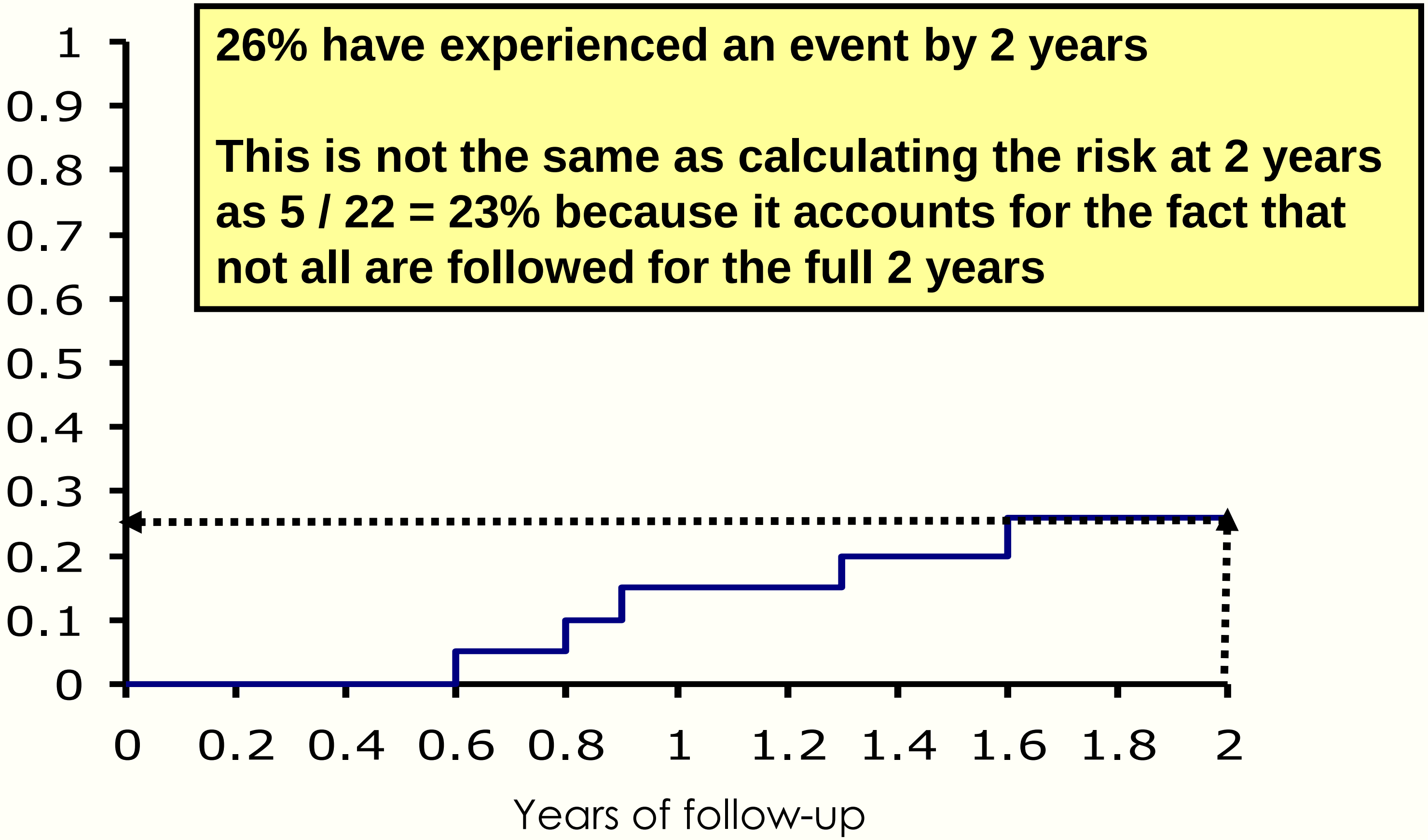


Kaplan-Meier plot with confidence interval



Confidence intervals are narrow due to large numbers at risk and long follow-up

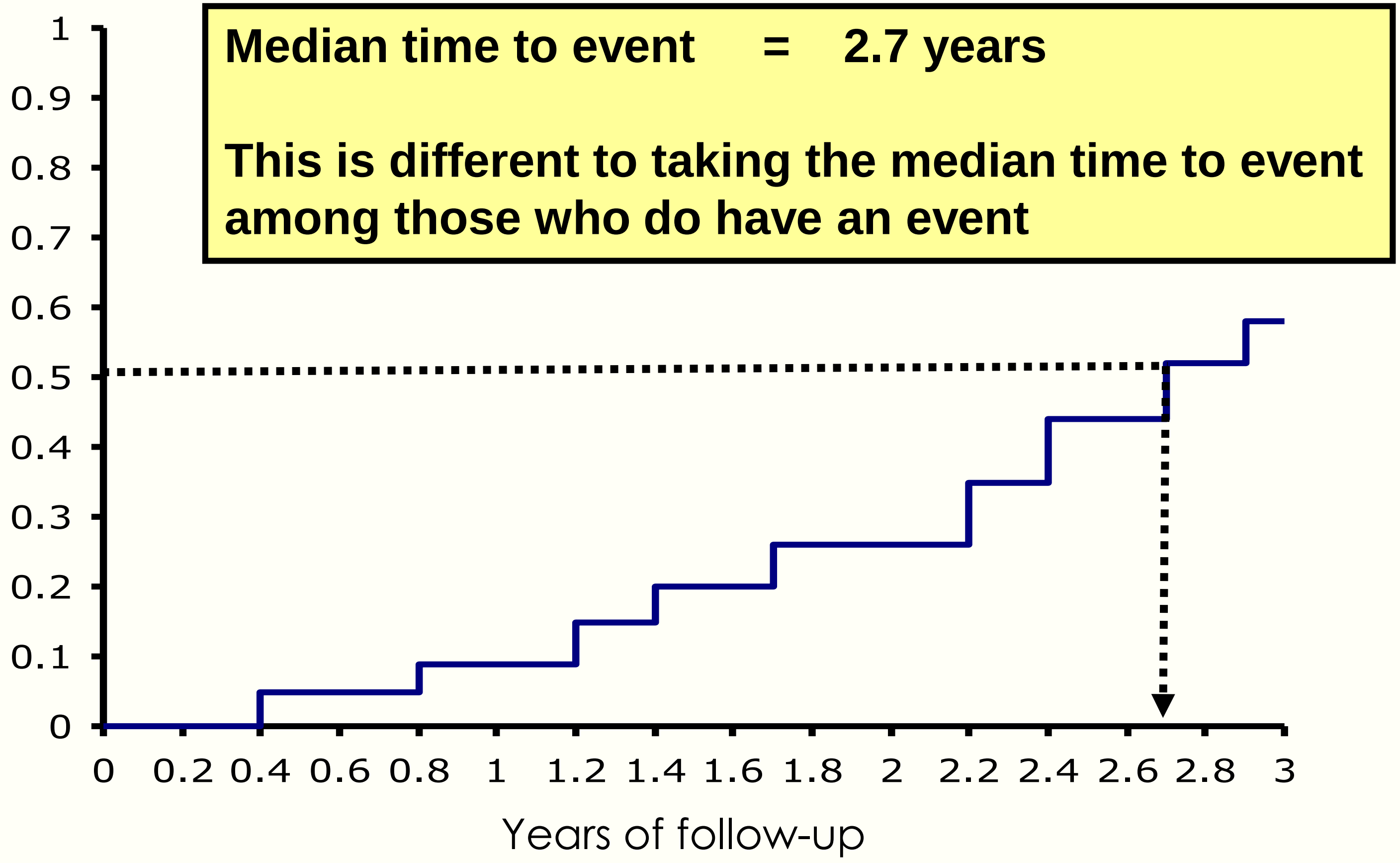
Alternative presentation



26% have experienced an event by 2 years

This is not the same as calculating the risk at 2 years as 5 / 22 = 23% because it accounts for the fact that not all are followed for the full 2 years

Median time to an event

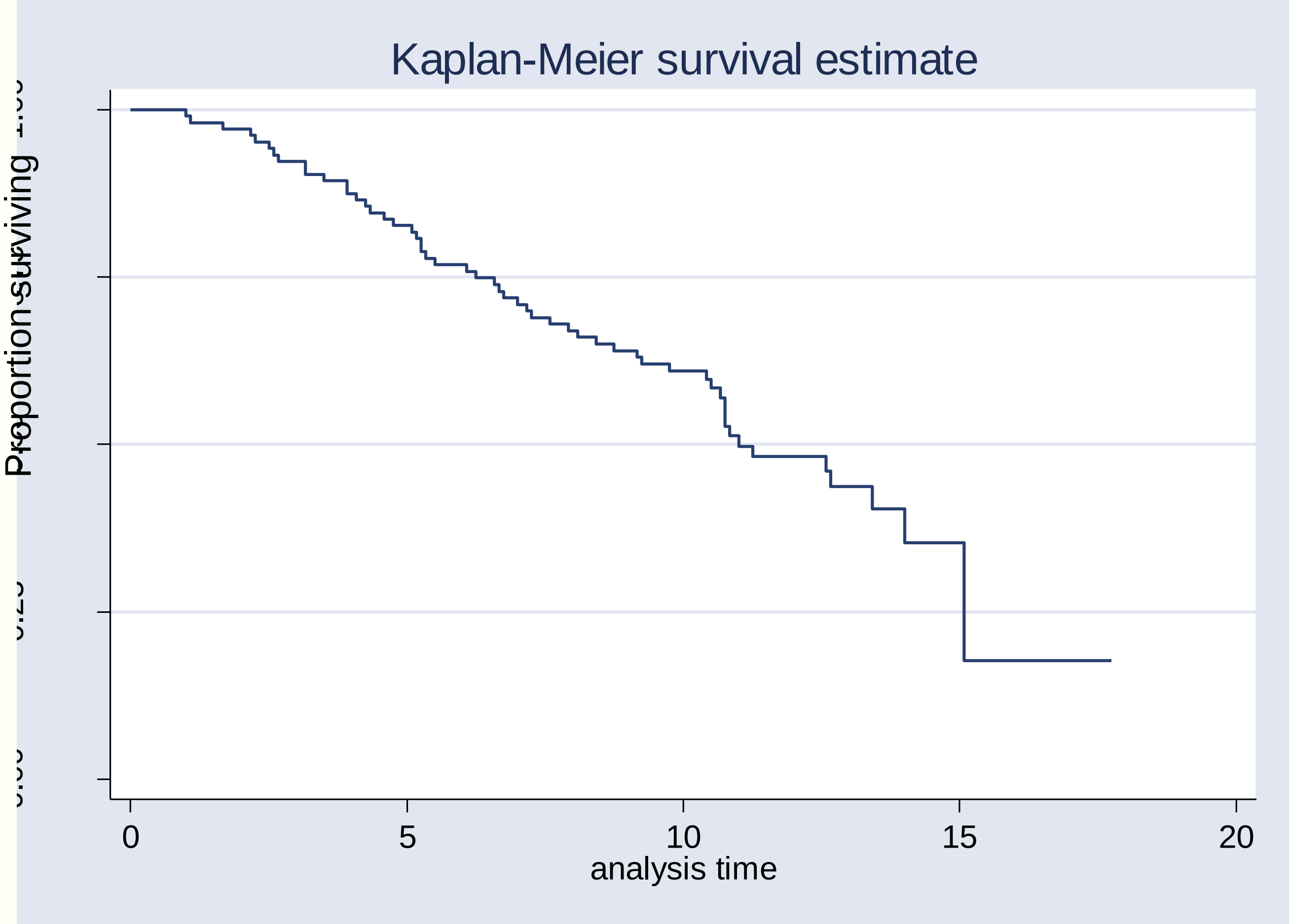


Median time to event = 2.7 years

This is different to taking the median time to event among those who do have an event

Kaplan Meier in Stata - Example

Survival time (years) after diagnosis with Parkinson’s disease



Kaplan Meier in Stata - Example

Survival time (years) after diagnosis with Parkinson’s disease

Output from **`sts list`**

```
Time    Total   Fail   Lost           Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
 .0833       520     0      1              1.0000         .          .         .
 .1667       519     1      0              0.9981    0.0019     0.9864    0.9997
 .3333       518     1      1              0.9961    0.0027     0.9847    0.9990
 .4167       516     1      0              0.9942    0.0033     0.9822    0.9981
    .5       515     2      0              0.9904    0.0043     0.9770    0.9960
 .5833       513     1      0              0.9884    0.0047     0.9744    0.9948
   .75       512     1      0              0.9865    0.0051     0.9719    0.9935
 .8333       511     1      0              0.9846    0.0054     0.9694    0.9923
 .9167       510     0      1              0.9846    0.0054     0.9694    0.9923
     1       509     1      0              0.9826    0.0057     0.9669    0.9909
```

Notes on Kaplan Meier plots 1

- Can either plot the probability of having an event (curve goes upwards) or of remaining event-free (curve goes downwards)

- Plots should be stopped once the number of patients remaining under follow-up and free of an event is small (<10?)

- The number remaining at risk in each group should be shown at regular intervals under the x-axis

Notes on Kaplan Meier plots 2

- Sometimes not enough data to estimate the median (e.g. only 5% experienced event by end of study)

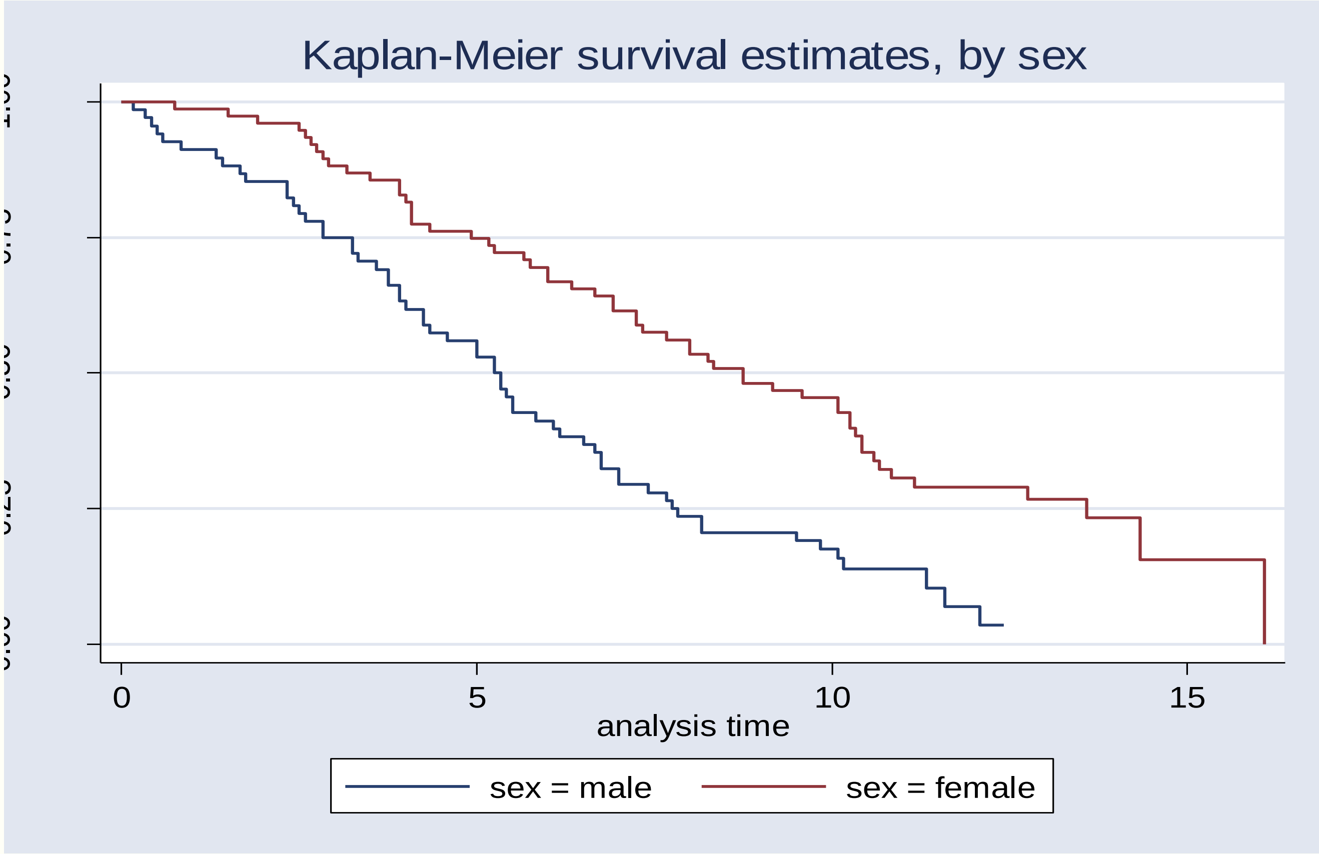
- By treating observations as censored, we implicitly assume that, were people to have been followed after censoring, they would have experienced the same event rate as those not censored

- This may not be the case if censoring is due to some other event that happened to the patient

Kaplan Meier in Stata - Example

Survival time (years) after diagnosis with Parkinson’s disease – by gender

Output from `sts graph, by(sex)`



Comparing two groups

Hypothesis test:

Null hypothesis : no difference in survival between two groups

Test statistic : the **log-rank test**. Non-parametric

Stata command (after ***`stset`*** data):

```
        sts test varname
```

Comparing two groups

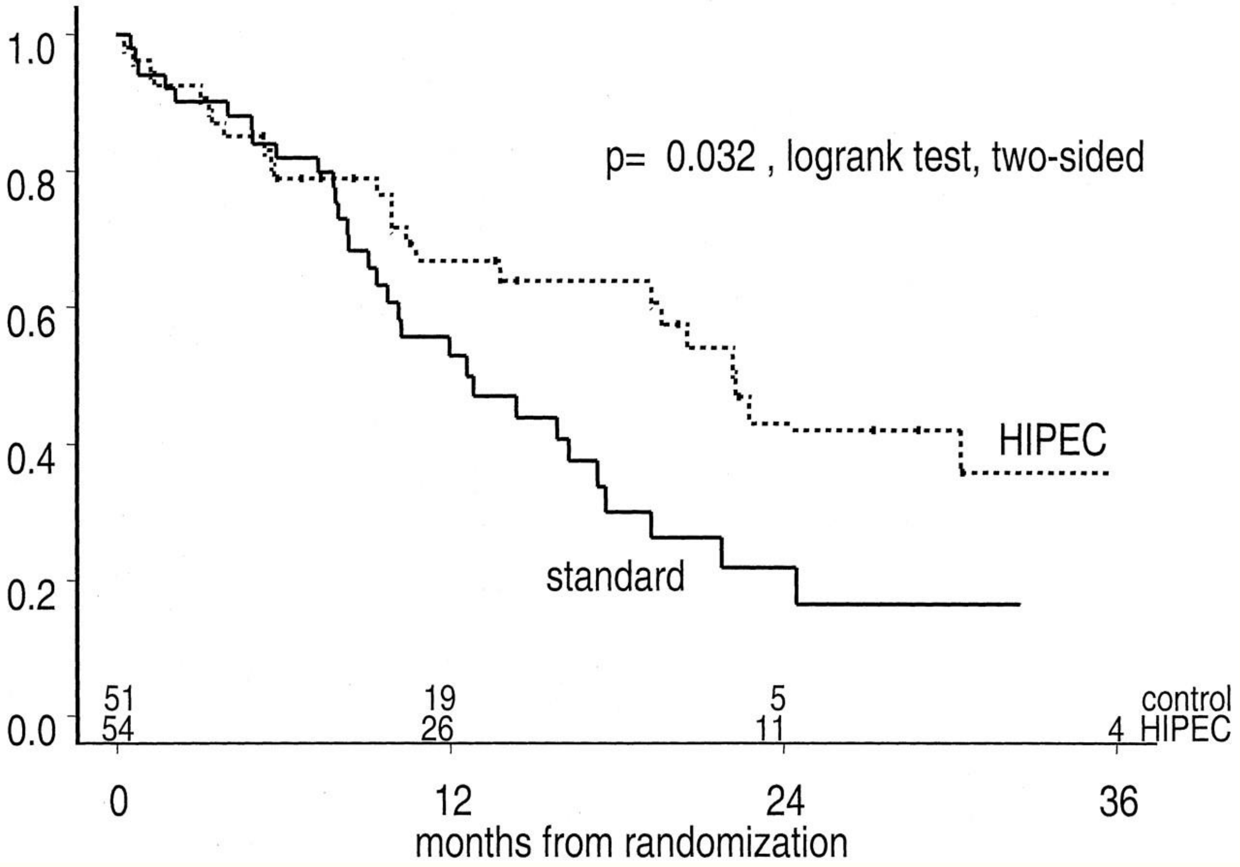
Calculate the Kaplan-Meier survival curves for each group separately

Can perform hypothesis test to assess whether difference between these curves is statistically significant (log-rank test)

The “log-rank test” tests for a difference between the curves across the whole range of observed values

Therefore better than simply comparing proportion alive at a particular time point (12 months, say)

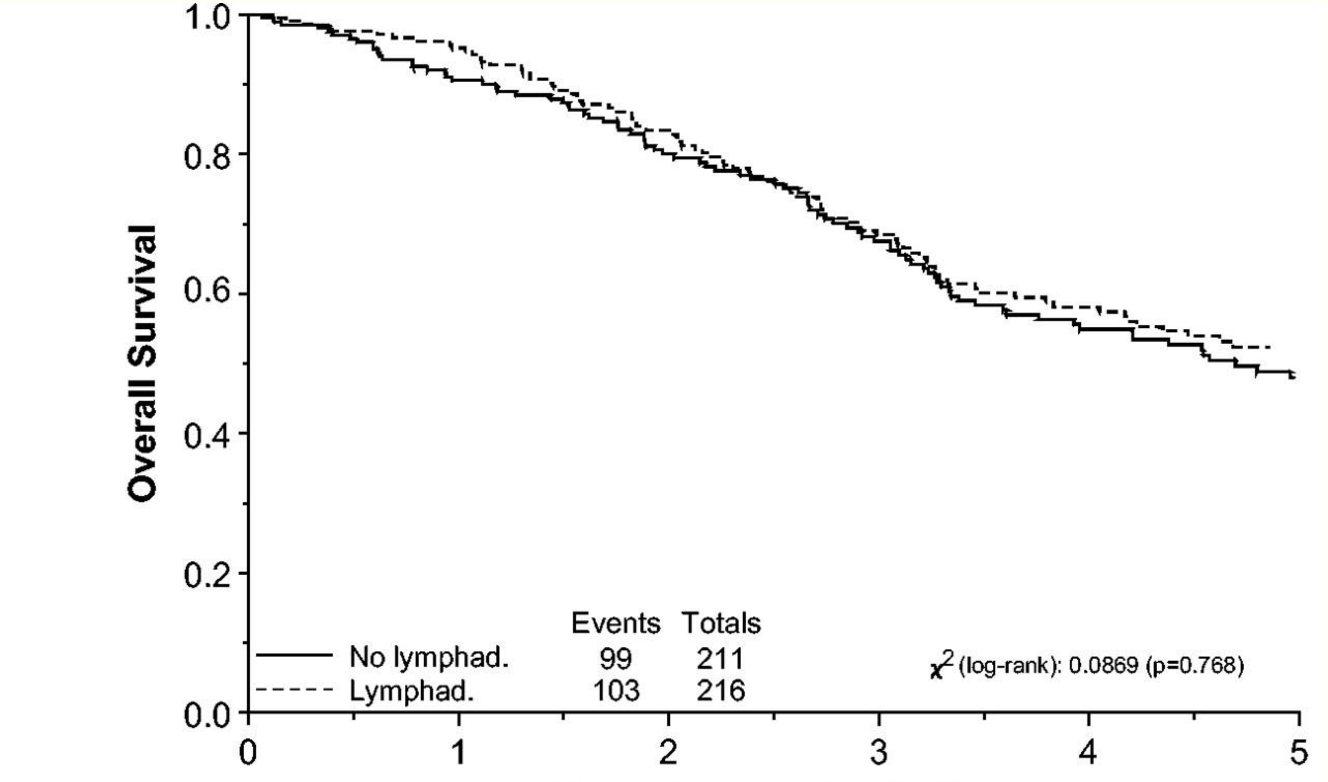
Example: Kaplan-Meier survival curve, comparing standard treatment to hyperthermic intraperitoneal chemotherapy (HIPEC).



Verwaal V J et al. JCO 2003;21:3737-3743

JOURNAL OF CLINICAL ONCOLOGY

Example: Overall survival (OS) for patients with optimally debulked advanced ovarian carcinoma undergoing systematic aortic and pelvic lymphadenectomy (Lymphad.) versus resection of bulky nodes only (No lymphad.).



Patients at Risk

	0	1	2	3	4	5
No lymphad.	211	177	136	104	78	56
Lymphad.	216	195	153	110	84	64

Panici P B et al. JNCI J Natl Cancer Inst 2005;97:560-566

Summary

- Survival (or time-to-event) data are common in the medical research field
- Survival data commonly contains censored data
- Kaplan Meier methods are frequently used to ascertain the cumulative percentage that have experienced an event by a particular time point
- We can compare survival between groups using the log rank test

Example: Log-rank test

Deaths times, t	$d_{(t)}$	n_{At}	n_{Bt}	e_{At}
2	2	22	22	1.0
3	1	21	21	0.50
4	1	20	21	0.49
6	1	19	21	0.48
.				
.				
.				
96	2	6	15	0.57
.				
.				
168	1	1	3	0.25

e.g For *t* = 96

$$e_{At} = \frac{n_{At}}{n_{At} + n_{Bt}} \times d_t = \frac{6}{6+15} \times 2 = 0.57$$

$$e_{Bt} = \frac{n_{Bt}}{n_{At} + n_{Bt}} \times d_t = \frac{15}{6+15} \times 2 = 1.53$$

Expected number of death in each group

E_A= 1.0+0.5+..... +0.57+..+0.25 = 10.62

E_B = 16+11 - 10.62 = 16.38

$$\chi^2 = \frac{(O_A - E_A)^2}{E_A} + \frac{(O_B - E_B)^2}{E_B} = 4.49$$

This gives P<0.05, so if null hypothesis of no survival difference between the treatments were true, we would only see such an extreme difference less than one time in twenty.

Hence we reject the null hypothesis.

However, note that this gives no estimate of the magnitude of the risk.