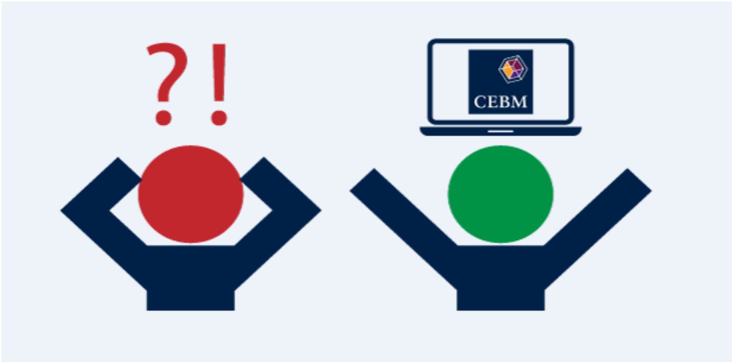
Tip for data extraction for meta-analysis – D10



Estimating a hazard ratio from a Kaplan curve and follow-up information

Kathy Taylor

Previously (post D9), I highlighted the paper by Tierney et al which describes how to estimate hazard ratios (HRs) from Kaplan Meier (K-M) curves and other time-to-event data. I also showed an example of the use of their spreadsheet calculator with the FLOT4 trial data. In this post I'll going to look at the underlying equations for the case of K-M curves reported with information about follow-up and work through equations the FLOT4 trial data.

I'd like to thank David Fisher (MRC Clinical Trials Unit, UCL) for his help in deriving the equations.

Table 1. Data for the FLOT4 trial

Time at start of interval (months)	Survival (event-free) %		Reported numbers at risk	
	FLOT	ECF/ECX	FLOT	ECF/ECX
0	100	100	356	360
2	99	99		
4	98	97		
6	93	91		
8	91	90		
10	87	83		
12	84	80	297	287
14	80	75		
16	78	73		
18	76	69		
21	72	63		
24	69	58	231	202
27	65	55		
30	61	54		
33	60	51		
36	57	49	140	126
39	55	47		
42	54	46		
45	53	45		
48	50	44	87	83

54	49	40		
60	45	36	39	33
66	43	35		
72	43	32	5	9

Table 1 shows my extracted data with the reported numbers at risk of mortality for the FLOT4 trial. This was a trial of two different peri-operative chemotherapy regimes in patients with gastric or gastro-oesophageal cancer. The treatment groups are abbreviated FLOT (for the research, intervention group) and ECF/ECX (for the comparator group).

The spreadsheet estimates for each time interval and each treatment arm (Figure):

1. Numbers of patients at risk (without events) at the start of the current interval
2. Numbers censored during the current interval
3. Numbers at risk during the current interval, adjusted for censoring
4. Numbers of (patients with) events during the current interval
5. O-E, V and the HR for the current interval

These steps are repeated across all intervals and finally these statistics are combined to calculate:

6. O-E, V and the HR for the whole survival curve.

Note that for the intervals up to the minimum follow-up time, no patients are censored.

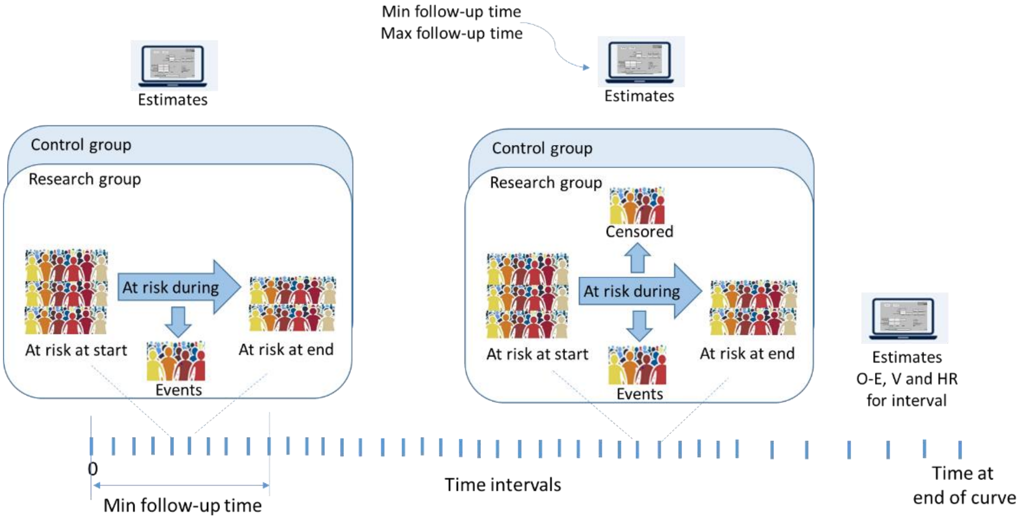


Figure. Spreadsheet calculations

Calculations are made from interval to interval, along all the time intervals which will include those reported and those chosen by the data extractor. This differs from the case of KM curves with numbers at risk (see my next post) where numbers at risk 'anchors' the estimates at particular times.

In my trial example, I estimated the follow-up range of 15 to 80 months. We're dealing with months as blocks of time so a minimum follow-up of 15 months means that all patients had complete follow-up and no patients were censored up to the end of month 15, which is the end of the time interval 14-16 months. Censoring will occur from the beginning of month 16 onwards, starting in the interval 16-18. This is why I said previously that intervals should be chosen so that the assumed minimum follow-up period falls at the end of an interval.

I will look at 16-18 months, so this will be the **current interval** and 14-16 months will be the **prior interval**.

The equations for the prior interval are simpler to those in the current interval where censoring applies.

Equations for the prior interval (14-16 months)

Numbers at risk at the start of the prior interval is

$$Number\ randomised\ \times Survival\ \%\ at\ start\ of\ prior$$

i.e.
356 x 0.80 = 284.8 in the research group
360 x 0.75 = 270.0 in the control group.

Numbers censored during the prior interval is assumed to be zero in both groups

Numbers of events in the prior interval is
$$Number\ randomised\ \times (Survival\ \%\ at\ start\ of\ prior - \ Survival\ \%\ at\ end\ of\ prior\ )$$
i.e.
356 x (0.80 – 0.78) = 7.12 in the research group
360 x (0.75 – 0.73) = 7.20 in the control group

Equations for the current interval (16-18 months)

Step 1: Numbers at risk at the start of the current interval
These are the numbers at risk at the end of the prior interval.

$$At\ risk\ at\ start\ of\ current = \ At\ risk\ at\ start\ of\ prior - Events\ in\ prior - Censored\ in\ prior$$
i.e.
284.8 – 7.12 – 0 = 277.68 in the research group
270.0 – 7.20 – 0 = 262.80 in the control group

STEP 2: Numbers of patients censored during the current interval
Assuming non-informative censoring (patients drop out for reasons unrelated to the study and at random), that censoring occurs at a constant rate within a given time interval, and using a simple

estimate based on similar triangles described in the appendix of Parmar et al (and which also shows the maths!):

$$Censored\ during\ current = At\ risk\ at\ start\ of\ current\ \times \frac{1}{2} \times \left(\frac{End\ of\ interval - Start\ of\ interval}{Maximum\ followup - Start\ of\ interval}\right)$$

i.e.

277.68 x 0.5 x (18-16)/(80-16) = 4.34 in the research group

262.80 x 0.5 x (18-16)/(80-16) = 4.11 in the control group

STEP 3: Numbers of patients at risk during the current interval, adjusted for censoring

The estimated number of censored patients are removed from those who are at risk at the start of the interval to calculate the "effective" numbers of patients at risk:

$$At\ risk, adjusted = At\ risk\ at\ start\ of\ current\ - censored\ in\ current$$

equation 1

i.e.

277.68 – 4.34 = 273.33 in the research group

262.80 – 4.11 = 258.69 in the control group

STEP 4: Numbers of patients with events during the current interval

A bit of maths (see below if you're interested) shows that

$$Events\ during\ current = At\ risk, adjusted \times \left(\frac{Survival\ \%\ at\ start\ of\ interval - Survival\ \%\ at\ end\ of\ interval}{Survival\ \%\ at\ start\ of\ interval}\right)$$

equation 2

i.e.

273.33 x (0.78 – 0.76)/0.78 = 7.01 in the research group

258.69 x (0.73 – 0.69)/0.73 = 14.17 in the control group

STEP 5: O-E, V and the HR for the current interval

The HR is calculated as a relative risk as both time to event and censoring have been accounted for.

$$HR = \left(\frac{\frac{Events\ for\ research}{At\ risk, adjusted\ for\ research}}{\frac{Events\ for\ control}{At\ risk, adjusted\ for\ control}}\right)$$

i.e.

7.01/273.33 divided by 14.17/258.69 = 0.468

A bit of maths (see below if you're interested) shows that

$$V = \frac{1}{\left(\frac{1}{Events\ for\ research} - \frac{1}{At\ risk, adjusted\ research} + \frac{1}{Events\ for\ \ control} - \frac{1}{At\ risk, adjusted\ control}\right)}$$

equation 3

$$V = \frac{1}{\left(\frac{1}{7.01} - \frac{1}{273.33} + \frac{1}{14.17} - \frac{1}{258.69}\right)} = 4.86$$

A direct method to calculate the HR is

$$HR = \exp\left(\frac{O - E}{V}\right)$$

Taking natural logs (post G8) of both sides and rearranging gives

$$O - E = ln(HR) \times V$$

i.e

ln(0.468) x 4.86 = -3.69

STEP 6: O-E, V and the HR for the whole survival curve.

Accounting for all intervals, the HR for the whole curve is calculated

$$HR = \exp\left(\frac{\sum O - E}{\sum V}\right)$$

i.e.

$$= exp\left(\frac{(0) + (-1.67) + (-1.99) + (1.63) + (-5.50) + (-0.26) + (-2.29) + (-0.24) + (-3.69) + \ etc}{(1.81) + (2.41) + (10.34) + (2.43) + (9.64) + (5.57) + (8.41) + (3.67) + (4.86) + etc}\right)$$

$$= exp\left(\frac{-23.47}{94.02}\right)$$

$$= 0.78$$

With V = 94.02, O-E = -23.47 and 95% CI of the HR is 0.64 to 0.95.

Here's a tip…

You can calculate a hazard ratio from a survival curve reported with follow-up information but make sure, when extracting curve data, that the minimum follow-up time lies at the end of an interval.

In my next blog post, I'm going to look at the equations underlying the spreadsheet calculations for estimating a HR from a Kaplan Meier curve reported with numbers of patients at risk.

Where did the equations come from?

(You can skip this if you are only interested in carrying out the calculations)

To derive equation 2:

I'll use shorter names of variables so the equations fit on a single line and only consider the equations for a single arm of a trial to simplify the notation. The same equations will apply to both treatment arms.

The standard K-M limit formula is

$$S_2 = S_1(1 - d_2/n_2)$$

Where

S_1 is the survival (at risk) proportions at the start of adjacent time-points t_1

S_2 is the survival (at risk) proportions at the start of adjacent time-point t_2

There are no events nor patients censored between, t_1 and t_2

n_2 is the number at risk just before time t_2,

d_2 is the number of events at time t_2.

As we are not observing events or censoring directly, the equation becomes

$$S_2^* = S_1^*(1 - d_2^*/n_2^*)$$

equation 5

Stars indicate that the quantities were not observed at time-points corresponding to changes in the risk set. Also,

d_2^* is the number of events since time t_1

c_2^* the number censored since time t_1

n_2^* is now the number of patients at risk since time t_1, adjusted for censoring.

Rearranging equation 5 becomes

$$d_2^* = n_2^* \left(\frac{S_1^* - S_2^*}{S_1^*} \right)$$

which is equation 2.

To derive equation 3:

As the HR can be calculated as a relative risk, we can use the formula for the standard error of the log relative risk, SE(ln(RR)) i.e.

$$SE(ln(HR)) = \sqrt{\frac{1}{Events\ for\ research} - \frac{1}{At\ risk, adjusted\ research} + \frac{1}{Events\ for\ control} - \frac{1}{At\ risk, adjusted\ control}}$$

$$Variance\ of\ ln(HR) = V^* = SE(ln(HR))^2$$

$$V = \frac{1}{V^*}$$

Therefore,

$$V = \frac{1}{\left(\frac{1}{Events\ for\ research} - \frac{1}{At\ risk, adjusted\ research} + \frac{1}{Events\ for\ control} - \frac{1}{At\ risk, adjusted\ control}\right)}$$

Dr Kathy Taylor teaches data extraction in Meta-analysis. This is a short course that is also available as part of our MSc in Evidence-Based Health Care, MSc in EBHC Medical Statistics, and MSc in EBHC Systematic Reviews.

Follow updates on this blog, related news, and to find out about other examples of statistics being made more broadly accessible on Twitter @dataextips