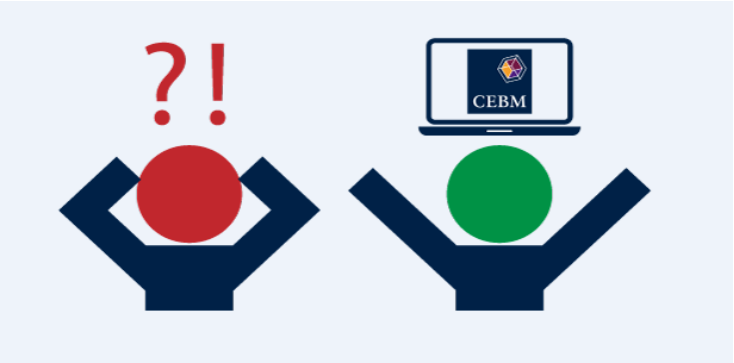
Tip for data extraction for meta-analysis – D11



Estimating a hazard ratio from a Kaplan curve and numbers at risk

Kathy Taylor

Previously (post D10), using trial data, I worked through the equations that underlie the spreadsheet calculator of Tierney et al to estimate a hazard ratio (HR) from Kaplan Meier (K-M) curves reported with information about follow-up. In this post I'll look at the equations for the case of a K-M curves reported with numbers at risk.

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

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

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
12	84	80	297	287
24	69	58	231	202
36	57	49	140	126
48	50	44	87	83
60	45	36	39	33
72	43	32	5	9

The authors highlight the advantage of these data offering a more direct way to assess censoring but the disadvantage is that there are fewer data points.

The spreadsheet calculations, which follow an actuarial life-table approach (Figure 1), estimate for each time interval and each treatment arm:

1. Numbers at risk during the current interval
2. Numbers of (patients with) events during the current interval
3. Numbers censored in the current interval (involving a fractional equation)
4. O-E, V and the HR for the current interval (2 estimations are given)

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

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

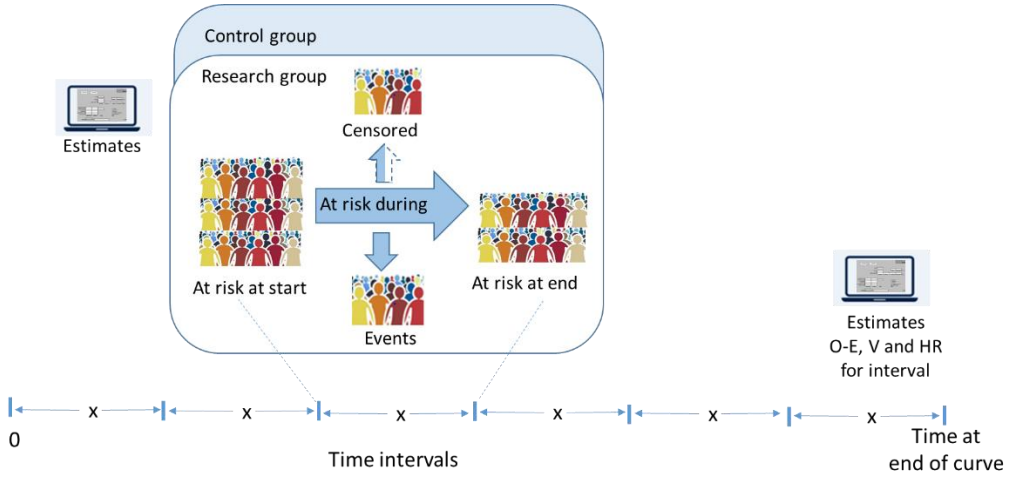


Figure 1. Spreadsheet calculations

Having the numbers at risk 'anchors' the estimates at particular times, unlike the case of estimates made for K-M curves with information about follow-up where the time intervals are those reported and those chosen by the data extractor <link>.

I will illustrate the use of the equations by showing the calculations for the first interval of the trial data (0-12 months).

A bit of maths shows (see below if you are interested)

STEP 1: Numbers at risk during the current interval

$$At\ risk\ during\ the\ interval = \frac{(At\ risk\ at\ start + At\ risk\ at\ end) \times Survival\ \%\ at\ start}{(Survival\ \%\ at\ start + Surical\ \%\ at\ end)}$$

equation 1

i.e.

(356+297) x 1.00/(1.00+0.84)=354.89 in the research group

(360+287) x 1.00/(1.00+0.80)=359.44 in the control group

STEP 2: Numbers of patients with events during the current interval

$$Events\ during\ current = \frac{(At\ risk\ at\ start + At\ risk\ at\ end)(Survival\ \%\ at\ start - Survival\ \%\ at\ end)}{(Survival\ \%\ at\ start + Survival\ \%\ at\ end)}$$

equation 2

i.e.

(356+297) x (1.00-0.84)/(1.00+0.84) = 56.78 in the research group

(360+287) x (1.00-0.80)/(1.00+0.80) = 71.89 in the control group

STEP 3: Numbers censored in the current interval

$$Censored\ during\ current = \frac{2\ \times (At\ risk\ at\ start\ \times Survival\ \%\ at\ end - At\ risk\ at\ end\ \times Survival\%\ at\ start)}{(Survival\ \%\ at\ start + Survival\ \%\ at\ end)}$$

equation 3

i.e.

2x(356x0.84-297x1.00)/(1.00+0.84)=2.22 in the research group

2x(360x0.80-287x1.00)/(1.00+0.80)=1.11 in the control group

It is interesting to see that approximately 3 patients were censored in the period 0-12 months but these were not accounted for previously (post D10) by estimating a minimum followup period of 15 months.

STEP 4a: Estimate HR and V for the current interval using the steps given previously <link>

i.e. The HR is calculated as a relative risk 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. (56.78/354.89)/(71.89/359.44)=0.80

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

equation 4

A direct method to calculate the HR is

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

Taking natural logs of both sides and rearranging gives

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

STEP 4b: Estimate E and then O-E

The number of expected events in the current interval for the research group is estimated as the fraction with events multiplied by the number at risk in the research group:

$$Expected\ events\ during\ for\ research = (Events\ research + Events\ control) \times \frac{At\ risk\ during, research}{At\ risk\ during, research + At\ risk, during, control}$$

equation 5

i.e.

E=(56.78+71.89) x 354.89/(354.89+359.64)=63.91

The difference between the observed and expected events in the research group is

O-E=56.78-63.91= -7.13

If the randomisation ratio is 1:1, estimate

$$V = \frac{Total\ observed\ events}{4}$$

equation 6

i.e.

(56.78+71.89)/4=32.17

If randomisation ratio is not 1:1 or the reported numbers at risk (Event free at start of intervals) are very different, use

$$V = \frac{Total\ observed\ events\ \times Analysed\ research\ \times Analysed\ control}{(Analysed\ research + Analysed\ control)^2}$$

equation 7

i.e.

$$\frac{(56.78 + 71.89) \times 356 \times 360}{(356 + 360)^2} = 32.17$$

The HR can be calculated directly as

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

i.e.

exp(-7.13/32.17)=0.80

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.

$$= exp\left(\frac{(-7.1) + (-13.8) + (1.4) + (0.9) + (-2.4) + (-0.7)}{(32.17) + (32.43) + (15.48) + (6.9) + (4.37) + (0.75)}\right)$$

$$= exp\left(\frac{-21.7}{92.1}\right)$$
$$= 0.79$$

With V = 92.1, O-E= -21.7 and 95% CI of the HR is 0.64 to 0.97.

Here's a tip…
You can calculate a hazard ratio from a survival curve reported with numbers at risk

Where did the equations come from?

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

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 (event-free) proportion at the start of time-points t_1
S_2 is the survival (event-free) proportion at the start of the 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.

We are not observing events or censoring directly so, the equation becomes

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

equation 8

where the stars indicate that the quantities were *not* observed at time-points corresponding to changes in the risk set.

d_2^* is the number of events since time t_1
c_2^* is the number of censorings since time t_1
n_2^* is now the number of patients at risk since time t_1, adjusted for censoring.

The method assumes that numbers at risk are known at a regularly spaced set of timepoints (an actuarial, life-table approach) and there is a constant rate of censoring within each time period. As those censored are only at risk for part of the interval it is assumed that censoring has the equivalent effect of half of them being at risk for the whole period:

$$n_2^* = n_1 - \frac{1}{2}c_2^*$$

equation 9

Therefore, 'fractional' numbers at risk across the intervals are estimated rather than "effective" numbers as described previously (post D10).

The number at risk at the *next* timepoint, $n_2$is obtained by subtracting *all* the censored patients *plus* the estimated number of events since time t_1:

$$n_2 = n_1 - c_2^* - d_2^*$$

equation 10

Rearrange equation 8 to calculate

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

equation 11

Rearrange equation 10 to calculate d_2^*
Rearrange equation 9 for n_2^*
Substitute d_2^* and n_2^* into equation 11 and then rearrange produces

$$c_2^* = 2\frac{n_1 S_2^* - n_2 S_1^*}{S_1^* + S_2^*}$$

which is equation 3.

Substitute equation 3 into equation 10 and rearrange produces

$$d_2^* = \frac{(n_1 + n_2)(S_1^* - S_2^*)}{S_1^* + S_2^*}$$

which is equation 2.

Substitute equation 2 into equation 11 and rearrange produces

$$n_2^* = \frac{(n_1 + n_2)S_1^*}{S_1^* + S_2^*}$$

which is equation 1.

To derive equations 6 and 7:

To simplify the notation, I'll consider the whole follow-up period.

E_1 is the expected number of events in the research group
E_2 is the expected number of events in the comparator group
D_1 is the number of events in the research group
D_2 is the number of events in the comparator group
N_1 is the number in the research group
N_2 is the number in the comparator group

Tierney et al refer to a direct method of calculating the variance as

$$V = \frac{1}{\frac{1}{E_1} + \frac{1}{E_2}} = \frac{E_1 E_2}{E_1 + E_2}$$

equation 12

As in equation 5

$$E_1 = (D_1 + D_2)\frac{N_1}{N_1 + N_2}$$

equation 13

$$E_2 = (D_1 + D_2)\frac{N_2}{N_1 + N_2}$$

equation 14

Substitute equations 13 and 14 into equation 12

$$V = (D_1 + D_2)\frac{N_1 N_2}{(N_1 + N_2)^2}$$

This is equation 6.

If the randomisation ratio is 1:1 then $N_1 = N_2 = N$ and equation 7 becomes

$$V = (D_1 + D_2)\frac{N^2}{4N^2} = \frac{(D_1 + D_2)}{4}$$

This is equation 8.

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