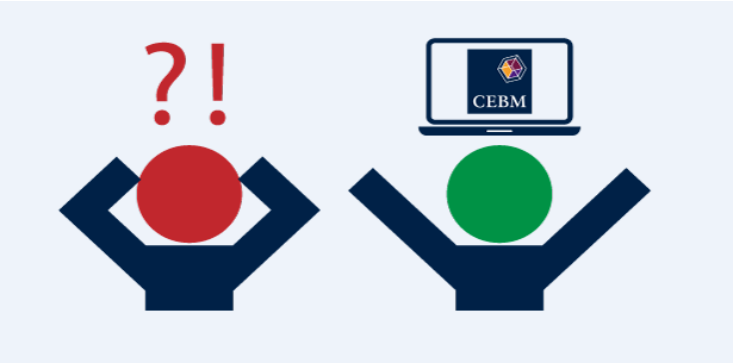
Tip for data extraction for meta-analysis – D9



Estimating a hazard ratio from time-to-event data – updated May 2025

Kathy Taylor

Here I return to extracting hazard ratios (HRs), but this time it's about making estimates from time-to-event data (survival data). Guyot et al use image extraction software to extract the co-ordinates of Kaplan-Meier (K-M) curves, also known as survival curves. They apply an algorithm to pseudo-reconstruct individual patient data, which they then re-analyse to estimate the HR. Guyot et al highlight other approaches, which use fewer data points from K-M curves including the methods of Parmar et al and Williamson et al. Tierney et al revisit these methods and make them more accessible, by providing simpler notation, step-by-step instructions, equations, worked examples from a couple of published trials, and a very useful spreadsheet that does all the calculations. Tierney et al have recently published updated guidance.

What does the new guidance add?

- Some clarifications – some are highlighted below **in bold**
- Two new scenarios – highlighted below **in bold**
- A figure presenting the various possibilities about how data may be reported
- Updated spreadsheet tool with enhanced features and a user guide
- Outline of best practices when extracting data from K-M curves and derivation of equations
- Multiple tips, such as recognising hazard ratios may be reported by another name, choosing between adjusted and unadjusted hazard ratios, estimating absolute effects, considering non-proportional hazards situations and choosing the most appropriate time intervals
- Recommendation that intervals are chosen such that a change in survival proportion is seen on both arms between timepoints (to avoid division-by-zero errors).

The guidance uses the following notation:

HR	Hazard ratio
V	logrank Variance
O	Observed number of events

E Expected number of events

SE Standard error

KM Kaplan Meier

log HR Natural logarithm (see post G8) of the hazard ratio (I will use the notation lnHR)

To generate a HR and V from a published report, the updated guidance explains what to do when the published report presents each of the following scenarios (**with updates in bold**):

1. O and E for the intervention and control arms
2. Any two of HR, O-E and logrank V (or logHR and SE)
3. HR (or O-E) and a confidence interval
4. HR (or O-E) and the events in each arm and **any randomisation ratio**
5. HR (or O-E) and the total events and randomisation ratio must be 1:1
6. HR (or O-E), total events and the numbers analysed in each arm and **any randomisation ratio**
7. **HR (or O-E) and a p value or chi-squared statistic (new scenario)**
8. p-value or chi-squared statistic and the events in each arm **for any randomisation ratio ***
9. p-value or chi-squared statistic and total events and randomisation ratio must be 1:1
10. p-value or chi-squared statistic, total events and numbers randomised to each arm and **any randomisation ratio**
11. **p-value or chi-squared statistic and confidence interval (a new scenario)**
12. Kaplan-Meier curves
    a. Reported with information about follow-up
    b. Reported with numbers at risk

*** Also clarifies that the p value should be divided by 2, not the z-score**

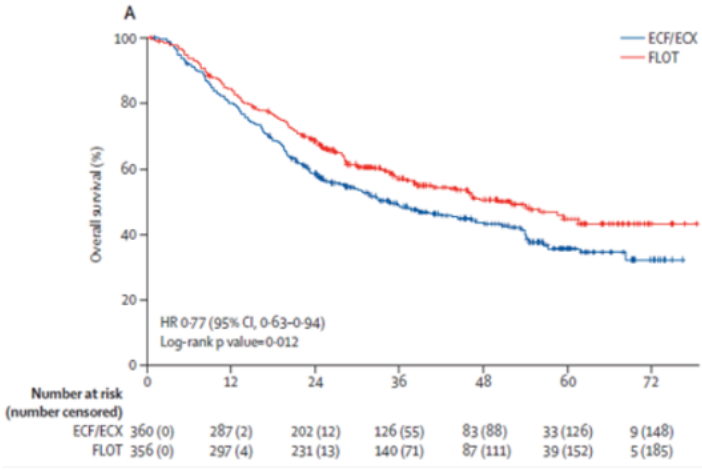
The spreadsheet can be used for all of the above, although the underlying equations for 1 to 11 are straightforward. For 12a and 12b, the inputs required for the spreadsheet include extracted curve data, and to estimate the numbers censored, either the reported maximum and minimum follow-up times (if these are not reported, Tierney et al offer advice on how these data may be estimated), or the reported numbers at risk. We say that a patient is censored if they leave the study before they've experienced the event of interest.

For 12a, the survival curve needs to be divided into a number of time intervals and the times and survival proportions extracted. These intervals should be chosen to give a good representation of the event rates over time, so when the event rate is high, you need to use closer intervals, and when the event rate is low, you can space out the intervals. You should also ensure that the minimum follow-up lies at the end of an interval (I'll explain why in the next blog post). For 10b, only the survival proportions at the times of the reported numbers at risk need to be extracted.

I'm going to illustrate the use of the updated spreadsheet by working through an example based on the FLOT4 trial. This was a trial of two different peri-operative chemotherapy regimes –

fluorouracil plus leucovorin, oxaliplatin and docetax (FLOT group) and epirubicin, cisplatin, fluorouracil or capecitabine (ECF/ECX comparator group) in patients with gastric or gastro-oesophageal cancer.

The reported HR for overall survival is 0.77 (95% CI 0.63 to 0.94) and here are the K-M curves.:



Source: Al-Batran et al. Lancet. 2019 May 11; 393(10184):1948-1957. Epub 2019 Apr 11

Here are the extracted data (which I extracted using the software that I demonstrated in my video post) tabulated with the reported numbers of patients at risk:

Table. 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

The 1st worksheet of the spreadsheet calculator (Figure 1) provides background information about the spreadsheet, including guidance on citing the calculator, and it also lists the other worksheets.

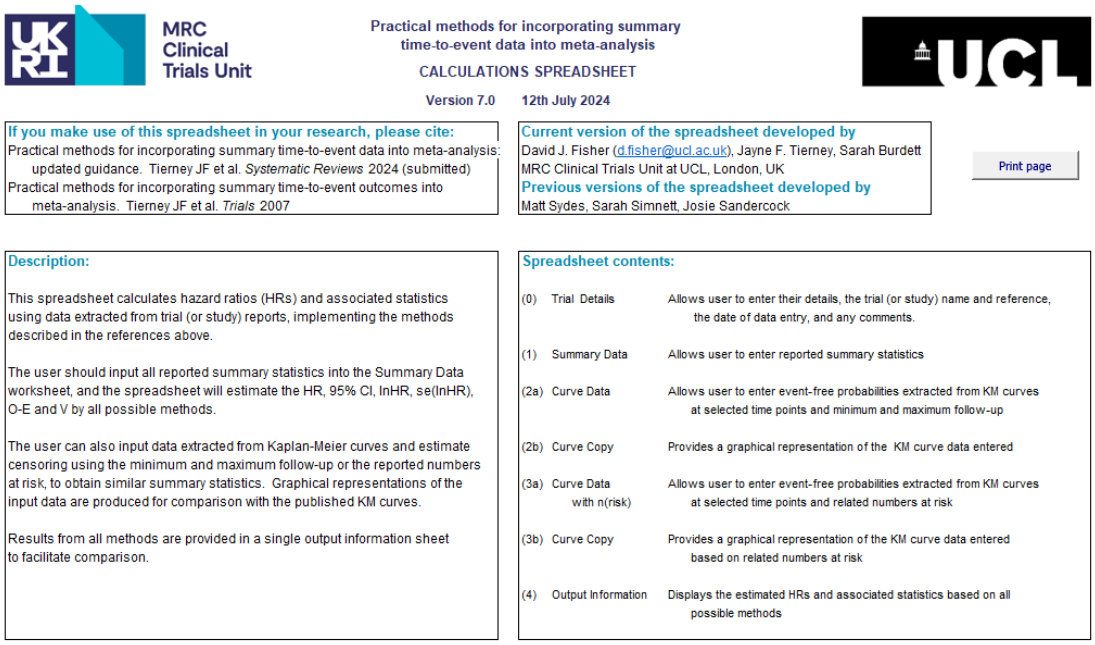


Figure 1. Introducing the spreadsheet

The 2nd spreadsheet (Figure 2) includes boxes to enter some basic data that are used for labelling subsequent spreadsheets, and it also provides licensing information.

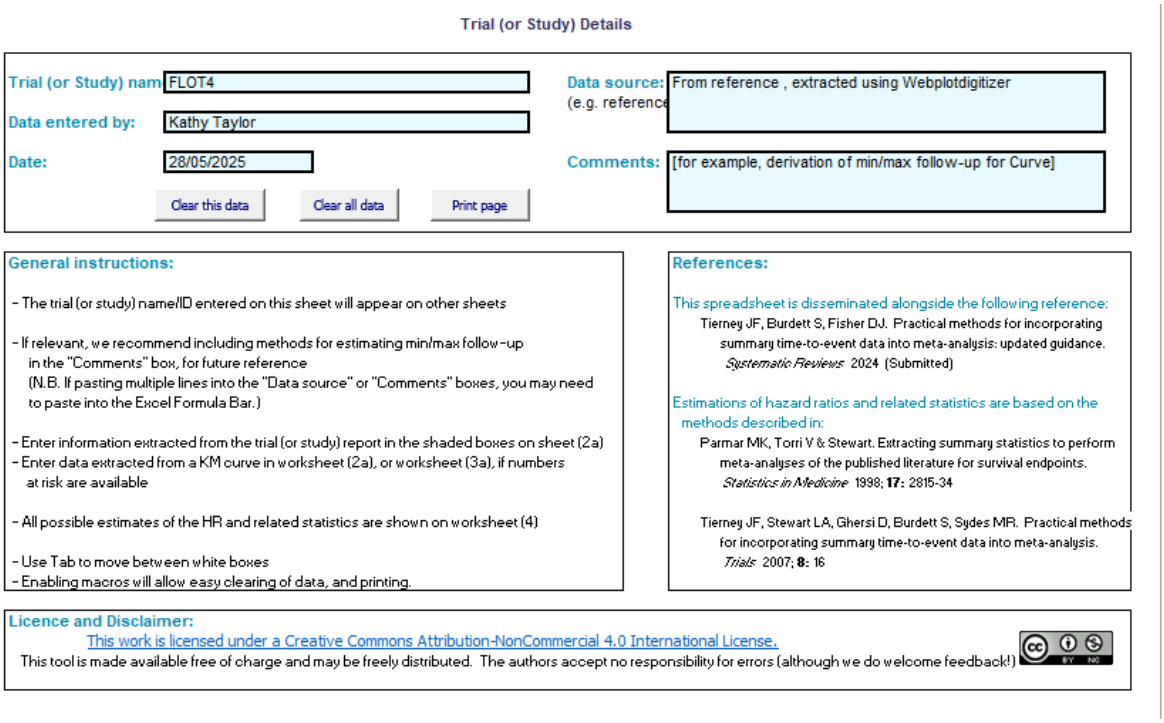


Figure 2. Inputting labels

The 3rd spreadsheet is the summary input data screen (Figure 3). This shows the time-to-event data that was reported for the FLOT4 trial.

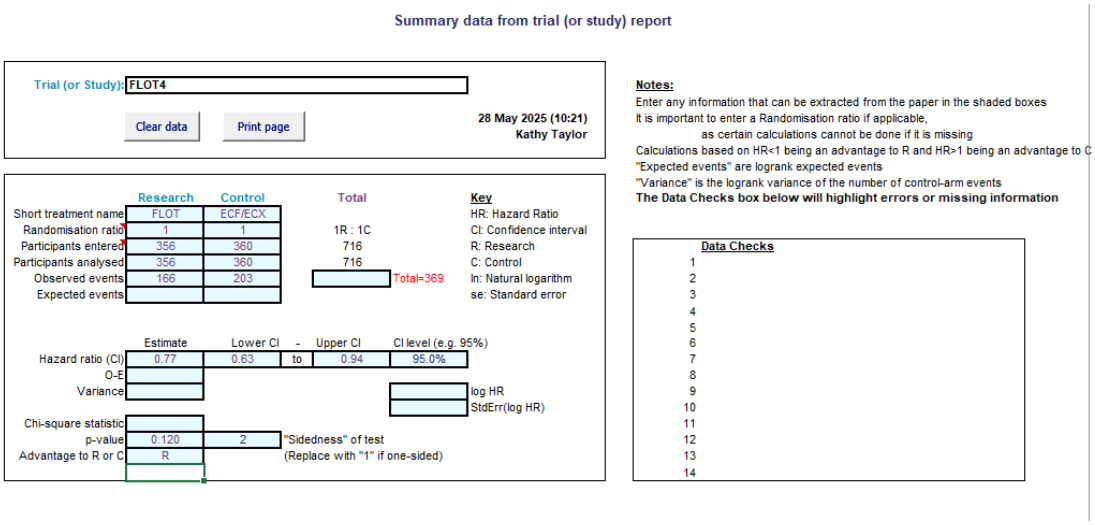


Figure 3. Summary input screen

The 4th worksheet (Figure 4) shows the extracted curve and followup data. The followup data was not reported and I estimated the minimum follow-up to be 15 months and the maximum follow-up to be 80 months. Note that using the data-extraction software produces numbers to many decimal places, but here I use integer times. I also entered the survival curves as integers so that the calculated numbers in my worked examples in the next two posts match exactly the calculated numbers in the spreadsheet.

The figure in the right-hand corner gives the estimated HR as 0.78 (the reported HR is 0.77). The accuracy of the calculated HR is pretty good but it could be improved by making the intervals smaller and extracting more data points.

Trial (or Study) FLOT4

Data extracted from Kaplan-Meier curve with min and max follow-up

28 May 2025 (10:26)
Kathy Taylor

Key

A: Checks
B: Time interval
C: Start of t
D: Survival (event-free) probability at start of t (%)
E: Effective number event-free at start of t
F: Effective number at risk during t
G: Effective number of events during t
H: Effective number censored during t
I: Survival (event-free) probability at start of t (%)
J: Effective number event-free at start of t
K: Effective number at risk during t
L: Effective number of events during t
M: Effective number censored during t
N: Log hazard ratio (HR) for t
O: Standard Error of log HR for t
P: Difference in observed and expected events (O-E) for t
Q: Logrank variance (V) for t

Clear data

Print page

Follow-up: (months)
15 80

Final survival (event-free) proportions:
Research 43 $S_r(t_e)$
Control 32 $S_o(t_e)$

Results (by Parmar method):
HR 0.78
V 94.57

			Research(1): FLOT					Control(1): ECF/ECX					Outcomes			
	Total					179.1	118.6				214.4	101.7			-23.11	94.57
A Checks	B t	C t_s	D $S_r(t_s)$	E $R_r(t_s)$	F $R_r(t)$	G $D_r(t)$	H $C_r(t)$	I $S_o(t_s)$	J $R_o(t_s)$	K $R_o(t)$	L $D_o(t)$	M $C_o(t)$	N $logHR_t$	O $StdErr_t$	P O_t-E_t	Q V_t
	0 to 2	0	100	356.0	356.0	3.6	0.0	100	360.0	360.0	3.6	0.0	0.00	0.74	0.00	1.81
	2 to 4	2	99	352.4	352.4	3.6	0.0	99	356.4	356.4	7.2	0.0	-0.69	0.64	-1.67	2.41
	4 to 6	4	98	348.9	348.9	17.8	0.0	97	349.2	349.2	21.6	0.0	-0.19	0.31	-1.99	10.34
	6 to 8	6	93	331.1	331.1	7.1	0.0	91	327.6	327.6	3.6	0.0	0.67	0.64	1.63	2.43
	8 to 10	8	91	324.0	324.0	14.2	0.0	90	324.0	324.0	25.2	0.0	-0.57	0.32	-5.50	9.64
	10 to 12	10	87	309.7	309.7	10.7	0.0	83	298.8	298.8	10.8	0.0	-0.05	0.42	-0.26	5.57
	12 to 14	12	84	299.0	299.0	14.2	0.0	80	288.0	288.0	18.0	0.0	-0.27	0.34	-2.29	8.41
	14 to 16	14	80	284.8	284.8	7.1	0.0	75	270.0	270.0	7.2	0.0	-0.06	0.52	-0.24	3.67
	16 to 18	16	78	277.7	273.3	7.0	4.3	73	262.8	258.7	14.2	4.1	-0.76	0.45	-3.69	4.86
	18 to 21	18	76	266.3	259.9	13.7	6.4	69	244.5	238.6	20.7	5.9	-0.50	0.34	-4.43	8.83
	21 to 24	21	72	246.2	240.0	10.0	6.3	63	217.9	212.3	16.9	5.5	-0.64	0.39	-4.28	6.65
	24 to 27	24	69	230.0	223.8	13.0	6.2	58	195.5	190.2	9.8	5.2	0.11	0.41	0.67	5.92
	27 to 30	27	65	210.8	204.9	12.6	6.0	55	180.4	175.3	3.2	5.1	1.22	0.62	3.19	2.61
	30 to 33	30	61	192.2	186.5	3.1	5.8	54	172.1	166.9	9.3	5.2	-1.22	0.65	-2.88	2.36
	33 to 36	33	60	183.4	177.6	8.9	5.9	51	157.7	152.6	6.0	5.0	0.24	0.52	0.91	3.74
	36 to 39	36	57	168.7	162.9	5.7	5.8	49	146.6	141.6	5.8	5.0	-0.15	0.58	-0.45	2.99
	39 to 42	39	55	157.2	151.5	2.8	5.8	47	135.9	130.9	2.8	5.0	-0.16	0.84	-0.22	1.41
	42 to 45	42	54	148.7	142.8	2.6	5.9	46	128.1	123.1	2.7	5.1	-0.16	0.86	-0.22	1.36
	45 to 48	45	53	140.2	134.2	7.6	6.0	45	120.4	115.2	2.6	5.2	0.93	0.71	1.85	1.98
	48 to 54	48	50	126.6	114.7	2.3	11.9	44	112.7	102.1	9.3	10.6	-1.51	0.72	-2.88	1.90
	54 to 60	54	49	112.4	99.5	8.1	13.0	40	92.8	82.1	8.2	10.7	-0.20	0.47	-0.91	4.49
	60 to 66	60	45	91.3	77.6	3.5	13.7	36	73.9	62.8	1.7	11.1	0.47	0.91	0.56	1.20
	66 to 72	66	43	74.2	58.3	0.0	15.9	35	61.1	48.0	4.1	13.1	na	na	na	na
	72 to 80	72	43	58.3	58.3	0.0	0.0	32	43.9	43.9	0.0	0.0	na	na	na	na
				na	na	na	na		na	na	na	na	na	na	na	na

Figure 4. Curve and followup data

The spreadsheet plots the extracted data in the 5th worksheet (Figure 5).

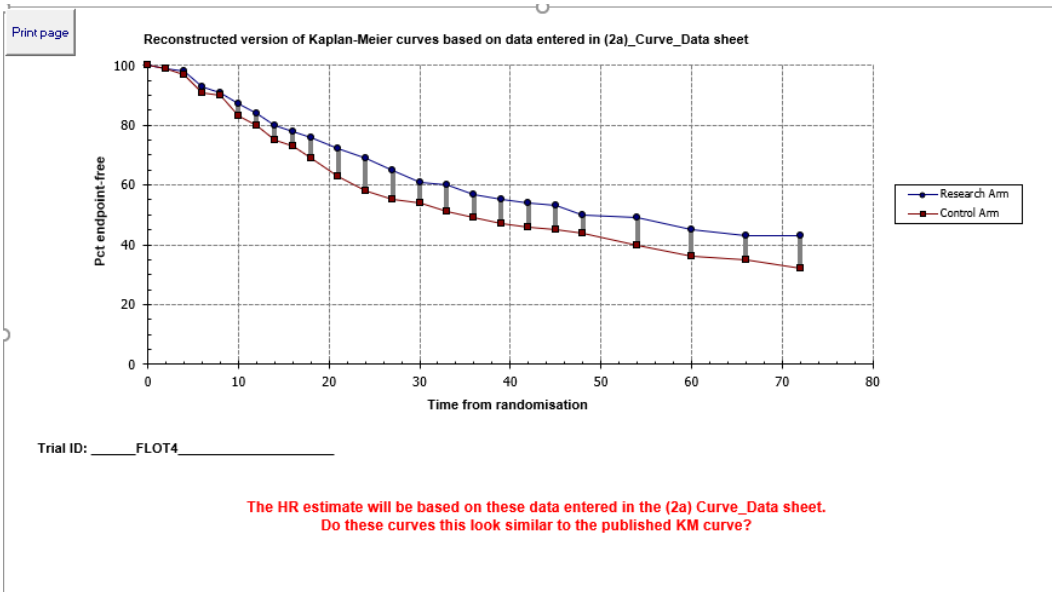


Figure 5. Plotted data corresponding with data shown in Figure 3

The 6th worksheet (Figure 6) includes the numbers at risk and corresponding survival fractions.

rial (or Study) FLOT4

Data extracted from Kaplan-Meier curve with related numbers of risk

28 May 2025 (10:33)
Kathy Taylor

Key

A: Checks
B: Time interval
C: Start of t
D: Survival (event-free) probability at t_{i-1} (%)
E: Number event-free at t_{i-1}
F: Effective number at risk during t_{i-1},t
G: Effective number of events during t_{i-1},t
H: Effective number censored during t_{i-1},t
I: Survival (event-free) probability at t_{i-1} (%)
J: Effective number event-free at t_{i-1}
K: Effective number at risk during t_{i-1},t
L: Effective number of events during t_{i-1},t
M: Effective number censored during t_{i-1},t
N: Logrank expected number of events in research arm for t_i
O: Difference in observed and expected events (O-E) for t_i
P: Logrank variance (V) for t_i
Q: Log hazard ratio (log HR) for t_i
R: Standard Error of log HR for t_i

Clear data

Print page

Results (by Parmar method with numbers at risk): HR 0.79, V 92.18

			Research(2): FLOT					Control(1): ECF/ECX					Outcomes				
	Total					166.4	184.6				202.7	148.3		-21.4	92.2		
A Checks	B t_{i-1}	C t_{i-1}	D $s_{2,i}$	E $n_{2,i}$	F $n^*_{2,i}$	G $d^*_{2,i}$	H $c^*_{2,i}$	I $s_{1,i}$	J $n_{1,i}$	K $n^*_{1,i}$	L $d^*_{1,i}$	M $c^*_{1,i}$	N $e^*_{2,k}$	O $d^*_{2,k}-e^*_{2,k}$	P v_k	Q $\log HR_k$	R $StdErr_k$
	0 to 12	0	100	356	354.9	56.8	2.2	100	360	359.4	71.9	1.1	63.9	-7.1	32.17	-0.22	0.18
	12 to 24	12	84	297	289.9	51.8	14.2	80	287	283.5	78.0	7.0	65.6	-13.8	32.43	-0.43	0.2
	24 to 36	24	69	231	203.2	35.3	55.7	58	202	177.8	27.6	48.4	33.6	1.8	15.66	0.11	0.3
	36 to 48	36	57	140	120.9	14.9	38.1	49	126	110.1	11.2	31.8	13.7	1.2	6.51	0.18	0.4
	48 to 60	48	50	87	66.3	6.6	41.4	44	83	63.8	11.6	38.4	9.3	-2.7	4.56	-0.58	0.5
	60 to 72	60	45	39	22.5	1.0	33.0	36	33	22.2	2.5	21.5	1.7	-0.7	0.87	-0.86	1.1
	72 onwards	72	43	5	na	na	na	32	9	na	na	na	na	na	na	na	na
					na	na	na			na	na	na	na	na	na	na	na
					na	na	na			na	na	na	na	na	na	na	na
					na	na	na			na	na	na	na	na	na	na	na
					na	na	na			na	na	na	na	na	na	na	na
					na	na	na			na	na	na	na	na	na	na	na

Figure 6. Curve data and reported numbers at risk

For this case, the calculated HR, shown in the upper right hand corner, is 0.79, which with the plotted curve (Figure 7) indicates the lower accuracy with less data.

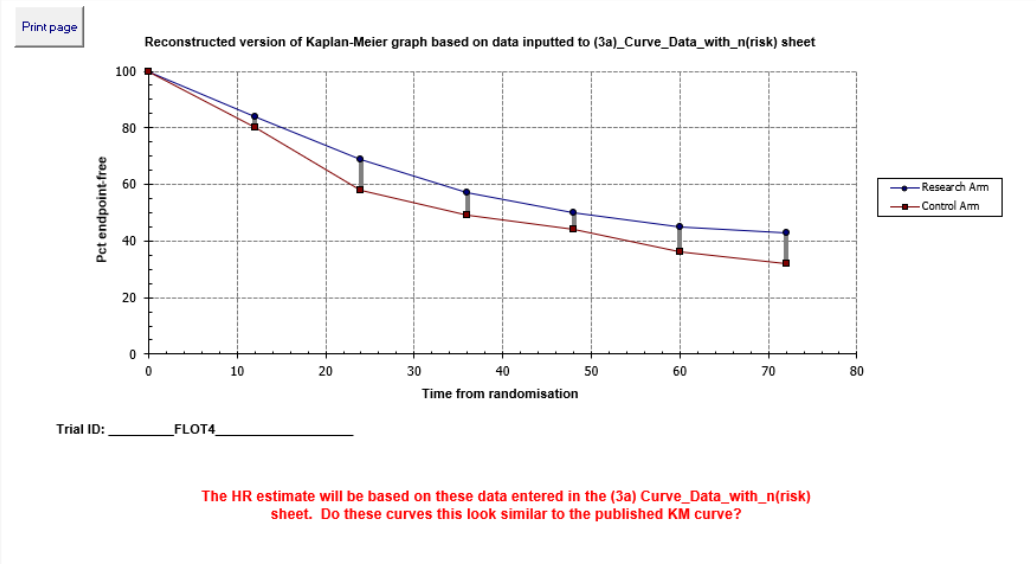
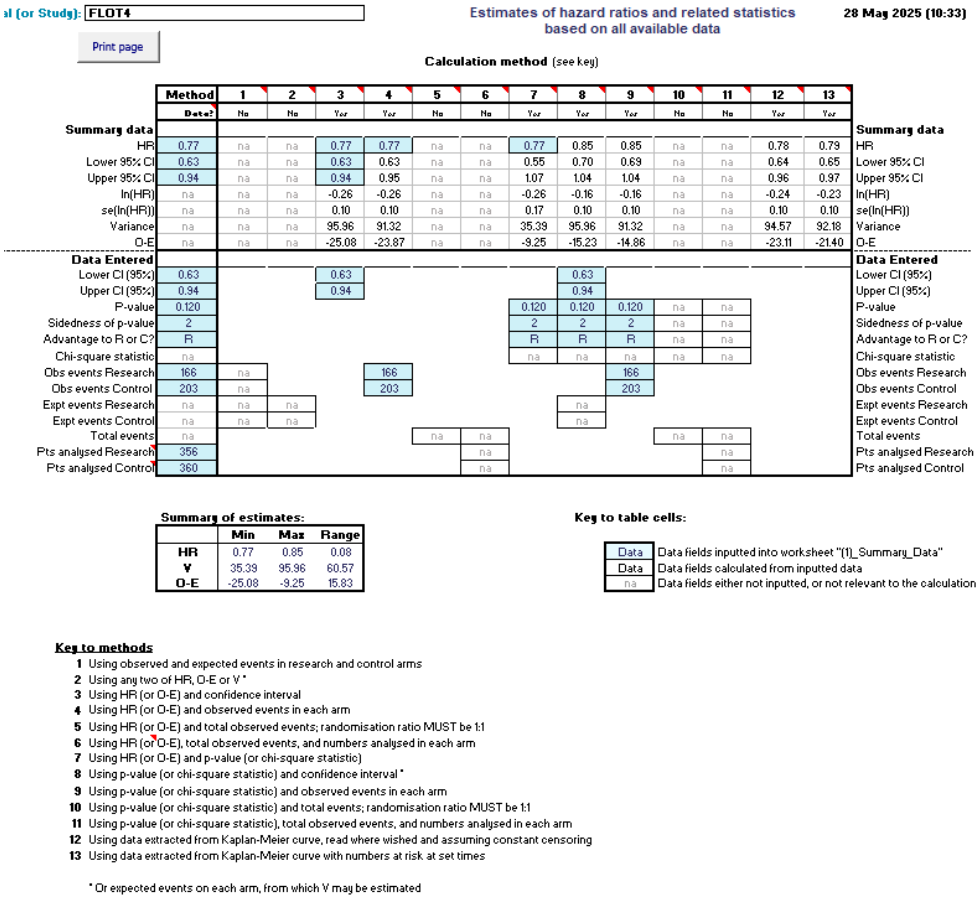


Figure 7. Plotted data corresponding with screen shot shown in Figure 4

The output screen (Figure 8) provides the estimated HRs with their confidence intervals. The estimated HR using the survival curve and follow-up data is 0.78 (0.64 to 0.96) – shown in column 12 – and the estimated HR using the survival curve and the numbers at risk is 0.79 (0.65 to 0.97) – shown in column 13. Both these estimates are very close to the actual HR of 0.77 (0.63 to 0.94).



Estimates of hazard ratios and related statistics based on all available data

28 May 2025 (10:33)

al (or Study): FLOT4

Print page

Calculation method (see key)

	Method	1	2	3	4	5	6	7	8	9	10	11	12	13	
	Data?	No	No	Yes	Yes	No	No	Yes	Yes	Yes	No	No	Yes	Yes	
Summary data															**Summary data**
HR	0.77	na	na	0.77	0.77	na	na	0.77	0.85	0.85	na	na	0.78	0.79	HR
Lower 95% CI	0.63	na	na	0.63	0.63	na	na	0.55	0.70	0.69	na	na	0.64	0.65	Lower 95% CI
Upper 95% CI	0.94	na	na	0.94	0.95	na	na	1.07	1.04	1.04	na	na	0.96	0.97	Upper 95% CI
ln(HR)	na	na	na	-0.26	-0.26	na	na	-0.26	-0.16	-0.16	na	na	-0.24	-0.23	ln(HR)
se(ln(HR))	na	na	na	0.10	0.10	na	na	0.17	0.10	0.10	na	na	0.10	0.10	se(ln(HR))
Variance	na	na	na	95.96	91.32	na	na	35.39	95.96	91.32	na	na	94.57	92.18	Variance
O-E	na	na	na	-25.08	-23.87	na	na	-9.25	-15.23	-14.86	na	na	-23.11	-21.40	O-E
Data Entered															**Data Entered**
Lower CI (95%)	0.63			0.63					0.63						Lower CI (95%)
Upper CI (95%)	0.94			0.94					0.94						Upper CI (95%)
P-value	0.120							0.120	0.120	0.120	na	na			P-value
Sidedness of p-value	2							2	2	2	na	na			Sidedness of p-value
Advantage to R or C?	R							R	R	R	na	na			Advantage to R or C?
Chi-square statistic	na							na	na	na	na	na			Chi-square statistic
Obs events Research	166	na			166					166					Obs events Research
Obs events Control	203	na			203					203					Obs events Control
Expt events Research	na	na	na						na						Expt events Research
Expt events Control	na	na	na						na						Expt events Control
Total events	na					na	na				na	na			Total events
Pts analysed Research	356						na					na			Pts analysed Research
Pts analysed Control	360						na					na			Pts analysed Control

Summary of estimates:

	Min	Max	Range
HR	0.77	0.85	0.08
V	35.39	95.96	60.57
O-E	-25.08	-9.25	15.83

Key to table cells:

Data	Data fields inputted into worksheet "(1)_Summary_Data"
Data	Data fields calculated from inputted data
na	Data fields either not inputted, or not relevant to the calculation

Key to methods

1 Using observed and expected events in research and control arms
2 Using any two of HR, O-E or V *
3 Using HR (or O-E) and confidence interval
4 Using HR (or O-E) and observed events in each arm
5 Using HR (or O-E) and total observed events; randomisation ratio MUST be 1:1
6 Using HR (or O-E), total observed events, and numbers analysed in each arm
7 Using HR (or O-E) and p-value (or chi-square statistic)
8 Using p-value (or chi-square statistic) and confidence interval *
9 Using p-value (or chi-square statistic) and observed events in each arm
10 Using p-value (or chi-square statistic) and total events; randomisation ratio MUST be 1:1
11 Using p-value (or chi-square statistic), total observed events, and numbers analysed in each arm
12 Using data extracted from Kaplan-Meier curve, read where wished and assuming constant censoring
13 Using data extracted from Kaplan-Meier curve with numbers at risk at set times

* Or expected events on each arm, from which V may be estimated

Figure 8. Output screen

In my next two blog posts, I'm going to look more closely at the equations underlying these spreadsheet calculations. I will first deal with the case of estimating a HR from K-M curves reported with follow-up information and in the next post I will look at the case of estimating the HR with reported numbers at risk.

Here's a tip…

There are equations you can use to convert time-to-event data into a suitable form for meta-analysis and there's a very useful spreadsheet available to do the calculations.

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.