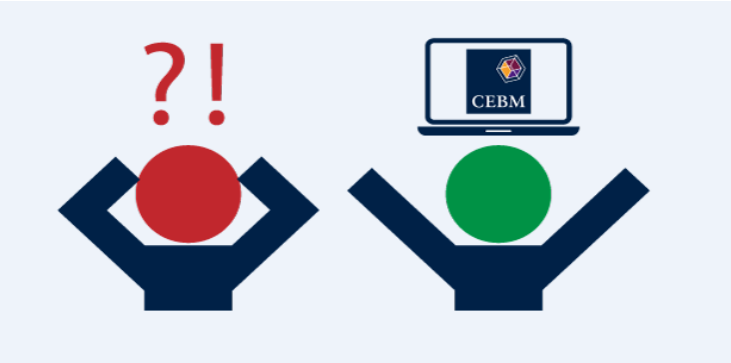
Tip for data extraction for meta-analysis – D6



Wanting a particular reference category in categorical risk data

Kathy Taylor

Previously, I showed a step-by-step guide (post D4) and worked example (post D5) of a trend estimation method for summarising categorical risk (quantile or dose-response) data, using the trend estimation method of Greenland and Longnecker, the STATA glst command and the R dosresmeta command. In my last post I also showed that you could deal with the problem unbounded limits of categories by imputing values derived from the ranges of other categories. In this post I will look at the problem of wanting a particular reference category which may not be the category that's reported. I will present three different examples.

Example 1 – switching the reference category

You may want to change the reference category from that with the lowest exposure to the category with the highest exposure. Looking again at the data from one of the studies in the worked example in my previous post (Table 1), the reference category has the lowest exposure (body mass index).

Table 1. Cumulative incidence data on body mass index and risk of atrial fibrillation

HR	lowerCI	upperCI	category	reference	n	cases	BMI range
1	1	1	1	1	1213	138	< 25
1.08	0.81	1.45	2	0	576	76	25 to 27.9
1.74	1.16	2.56	3	0	208	39	≥ 28

To change the reference category to that with the highest exposure we need to divide all the hazard ratios (HRs) by 1.74 (the HR of the category with the highest exposure), divide all the lower confidence interval limits by 1.16 (the lower confidence limit of the highest exposure category) and divide all the upper confidence interval limits by 2.56 (the upper confidence limit of the highest exposure category). Note that you need to swop the upper and lower limits of the confidence intervals (Table 2) because the transformed lower limit become upper limits.

Table 2. With highest exposure as reference category

HR	upperCI	lowerCI	category	reference	n	cases	BMI range
0.57	0.86	0.39	1	0	1213	138	< 25
0.62	0.70	0.57	2	0	576	76	25 to 27.9
1	1	1	3	1	208	39	≥ 28

Example 2 – separating data and switching the reference category if necessary

Sometimes an inner category is the reference category, as in Table 3, which shows data from a study of weight change and risk of atrial fibrillation. In this case, the reference category divides the categories into weight gain and weight loss. It would not be appropriate to include weight gain and weight loss data in the same meta-analysis, so these data need to be analysed separately, with the reference category featuring in both analyses. Having separated the data, the reference category may be changed, if necessary, as shown in Example 1

Table 3. Cumulative incidence data on weight change and risk of atrial fibrillation

HR	lowerCI	upperCI	category	reference	n	cases	weight change
1.52	1.16	1.99	1	0	543	88	>5% loss
1.01	0.79	1.31	2	0	864	98	0 to 5% loss
1	1	1	3	1	1514	154	0 to 4.9% gain
1.33	1.04	1.7	4	0	956	113	5 to 9% gain
1.61	1.24	2.11	5	0	623	87	≥10% gain

Example 3 – setting the reference category when deriving relative risks from event data

In cases where categorical data are reported with rates, unadjusted estimates of relative risks (RRs) may be estimated, and as part of this process, you can chose the reference category. A study which featured in Perez et al presented rates of the first major vascular event in a trial of simvastatin verses placebo for various baseline categories including those of total cholesterol <5.0, ≥ 5.0 and <6.0, and ≥6.0 mmol/L for categories 1, 2 and 3 respectively. In the intervention group, the event rates for categories 1, 2 and 3 were 360/2030 (18%), 744/3942 (19%) and 929/4297 (22%) respectively. You can estimate RRs from these data by using a generalised linear model function (glm) in STATA and the method of Chêne and Thompson. The data are read into Stata as shown below

	event	level	TC1vs2	TC1vs3	TC3vs2
1	1	1	360	360	929
2	0	1	1670	1670	3368
3	1	0	744	929	744
4	0	0	3198	3368	3198

Looking at the column TC1vs2 (the comparison between category 1 and category 2), the first row gives the number with events (event=1) in category 1. The second row gives the number with no

event (event=0) in category 1. The next two rows give the numbers with events and without events for category 2. The reference category is indicated by level=1.

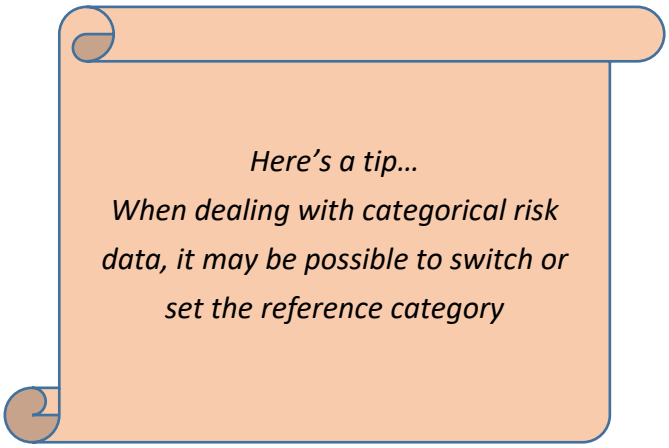
glm event ib1.level [fweight = TC1vs2], fam(bin) link(log) nolog eform
estimates the RR of category 2 compared to category 1 (reference) as 1.06 (0.95 to 1.19).
glm event ib1.level [fweight = TC1vs3], fam(bin) link(log) nolog eform
estimates the RR of category 3 compared to category 1 (reference) as 1.22 (1.09 to 1.36).

In the above commands **event** is the dependent variable and **level** is the independent variable. Frequency weights are applied using **fweight**. The outcome is binary so the family distribution is binomial, shown as **fam(bin)** and the link function between the covariate and outcome is specified as log in **link(log)**, so a log-binomial function is used. **nolog** reduces the output and **eform** exponentiates the output to produce relative risks. Level is specified as **ib1.level** as a factor variable and setting level=1 as the base or reference level.

To estimate the RRs with category 3 as the reference category, you can either do this by hand, as shown in Example 1 (i.e. RR=1/1.219 for category 1 etc), or you can set STATA to do the calculations, as follows:

glm event ib0.level [fweight = TC1vs3], fam(bin) link(log) nolog eform
estimates the RR of category 1 compared to 3 (reference) as 0.82 (0.74 to 0.92)
glm event ib1.level [fweight = TC3vs2], fam(bin) link(log) nolog eform
estimates the RR of category 2 compared to 3 (reference) as 0.87 (0.80 to 0.95)

Both sets of RRs together with the numbers of events and total patients for each category produce cumulative incidence data. Recall that I described different types of categorical data in an earlier post.



My next blog post will focus on situations where categorical risk data are incomplete.

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