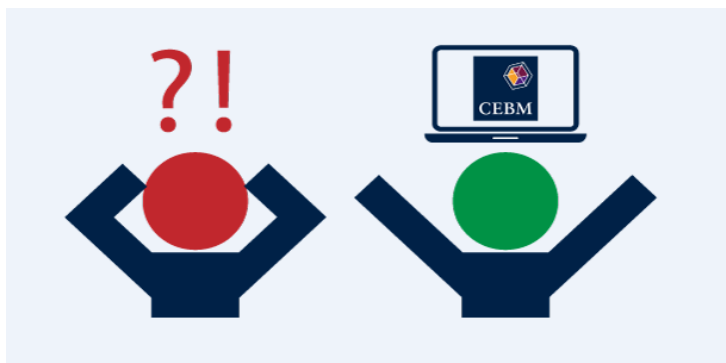**Tip for data extraction for meta-analysis – D5**

A worked example using trend estimation to summarise categorical risk data

Kathy Taylor

Previously, in post D4, I introduced a 5-step guide to summarising categorical risk data for a single study, using the trend estimation method of Greenland and Longnecker, the STATA **glst** command of Orsini et al, or the **dosresmeta** R command of Crippa and Orsini. Here I'll show a worked example based on a study by Grundvold et al (2012) who report hazard ratios (HRs) for the association between body mass index (BMI) and atrial fibrillation for three categories of BMI: BMI <25 kg/m^2 (which they classed as normal weight); BMI 25 to 27.9 (overweight); and BMI ≥28 (obese).

I'll first go through the 5 steps using STATA, and then provide the corresponding R code. I'll end this post by presenting a typical scenario to illustrate the usefulness of the trend estimation method.

1. Applying the trend estimation method in STATA.

The **glst** package needs to be installed in STATA, and installing packages is usually done first. Use the following command to find the link to obtain the **glst** package and follow the installation instructions.

findit glst

Installation only needs to happen once, so delete or comment out this line (start the line with *) once you installed **glst**.

STEP 1: Establish the type of data.

Here's the data in EXCEL

	A	B	C	D	E	F	G	H	I	J
1	BMI range	min	max	adjrr	lb	ub	category	reference	n	cases
2	less than 25		25	1	1	1	1	1	1213	138
3	25 to 27.9	25	27.9	1.08	0.81	1.45	2	0	576	76
4	28 or higher	28		1.74	1.16	2.56	3	0	208	39

These categorical data show an increasing risk of atrial fibrillation associated with increasing BMI. The study reports the number of events of patients experiencing incident atrial fibrillation (cases) and total number of subjects (n) are also given for each category. These are cumulative incidence data.

To import the data stored as an Excel file into STATA:
import excel "L:\Blog data extraction\Other files\trendest.xlsx", ///
sheet("Sheet1") cellrange(A1:J4) firstrow clear

To import the data stored as a comma-spaced values (csv) file into STATA:
import delimited "L:\Blog data extraction\Other files\trendest.csv", clear

STEP 2: Set the average exposure for each category

In this example, the exposure is BMI. I'll estimate the average BMI as the midpoint of each category. I first need to impute sensible values for the unbounded (unreported) limits of the outer categories.

I estimate the ranges of the BMIs of the outer categories to be twice the range of BMI of the inner category. You could run sensitivity analyses to explore the use of other multiples.

To carry out the above in STATA, first calculate the range of the inner category, and store this value in a new variable *range2* by using the saved feature *r(mean)* of the *summarize* command (I'm using the abbreviation, *summ*):

gen range=.
replace range=(max-min) if category==2
summ range if category==2
gen range2=r(mean)
The range of the inner category has value of 2.9 kg/m^2

Then create a variable for the multiplier, *mult,* and use it and the range of the inner category, *range2,* to impute values for the unreported limits of the outer categories:
gen mult=2
*replace min=max - mult*range2 if category==1*
*replace max=min + mult*range2 if category==3*

The average exposures of the categories can then be calculated as the midpoint BMIs:
gen average=(max+min)/2

STEP 3: For each category calculate the change in exposure from that of the reference group

The reference category has a HR of 1. The change in exposure is calculated as the difference in the average exposures from that of the reference category. Again, use the saved feature of ***summarize*** to create the variable, ***average0***, for the average value of the reference category:

summm average if category==1
gen average0=r(mean)
gen change=average - average0

STEP 4: Apply the trend estimation method
Log-transform the hazard ratio and its confidence interval and calculate the standard error for each category:

gen loghr=log(HR)
gen loglb=log(lowerCI)
gen logub=log(upperCI)
gen double se=(logub-loglb)/(2*invnormal(0.975))

We've established that we have cumulative incidence data, so I use the ***ci*** option in ***glst***:
glst loghr change, se(se) cov(n cases) ci

Running this command produces the following output:

```
Generalized least-squares regression            Number of obs   =       2
Goodness-of-fit chi2(1)    =   1.66              Model chi2(1)   =    5.95
Prob > chi2                =  0.1982             Prob > chi2     =  0.0147
```

logrr	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
change	.052018	.0213198	2.44	0.015	.010232	.0938041

STEP 5: Calculate the linear trend
I want to exponentiate (back-log transform) the STATA output to calculate the HR on the continuous scale and rescale to an increase of 5 kg/m^2 of BMI. I showed how to rescale hazard ratios in post D1, but here I'll use the appropriate ***lincom*** command to carry out both the exponentiation and rescaling:

lincom change*5, hr

```
( 1)   5*change = 0
```

logrr	exp(b)	Std. Err.	z	P>|z|	[95% Conf. Interval]
(1)	1.297047	.138264	2.44	0.015	1.052491 1.598428

So the categorical HRs for Grundvold 2012 are converted into a HR on a continuous scale of 1.30 (1.05 to 1.60), which indicates a 30% increased risk of atrial fibrillation associated with a 5 kg/m^2 increase in BMI.

2. Applying the trend estimation method using R (code and output)

 # Install the dosresmeta package
 # This only needs to be done once and then commented out
install.packages("dosresmeta")

 # Load library
 # This needs to be done every time you run this program
library("dosresmeta")

 # STEP 1: Establish the type of data.
 # Load and look at the data
mydata<-read.csv("L:/Blog data extraction/Other files/trendest.csv", header=TRUE, sep=",")
View(mydata)

 # STEP 2: Set the average exposure for each category
 # Calculate and save the range of the 2nd (inner) category
mydata$range<-ifelse(mydata$category==2,mydata$max-mydata$min,NA)
mydata$range2<-mydata$range[2]

 # Set the multiplier
mydata$mult<-2

 # Impute values for the unbounded limits
mydata$min<-
ifelse(mydata$category==1,mydata$max-ydata$mult*mydata$range2,mydata$min)

mydata$max<-
ifelse(mydata$category==3,mydata$min+mydata$mult*mydata$range2,mydata$max)

 # Calculate the midpoints of the categories

mydata$average<-(mydata$min+mydata$max)/2

 # STEP 3: For each category calculate the change in exposure from that of the reference group

 # Extract the average of the reference category (1st category)

mydata$average0<-mydata$average[1]

 # Calculate the change from the reference category

mydata$change<-mydata$average-mydata$average0

 # STEP 4: Apply the trend estimation method of dosresmeta

mydata$logrr<-log(mydata$adjrr)

mod.ci<-dosresmeta(formula=logrr~change,   type="ci",   cases=cases,   n=n,   lb=lb,   ub=ub,
data=mydata)

summary(mod.ci)

```
One-stage fixed-effects meta-analysis
Covariance approximation: Greenland & Longnecker

Chi2 model: X2 = 5.9584 (df = 1), p-value = 0.0146

Fixed-effects coefficients
         Estimate   Std. Error      z   Pr(>|z|)   95%ci.lb   95%ci.ub
change     0.0518       0.0212   2.4410    0.0146     0.0102     0.0935   *
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

1 study 2 values, 1 fixed and 0 random-effects parameters
 logLik      AIC       BIC
 0.8938   0.2125   -1.0944
```


 # STEP 5: Calculate the linear trend

predict(mod.ci, delta=5, exp=TRUE)


```
delta      pred      ci.lb      ci.ub
    5 1.295914 1.052408 1.595761
```


You will notice that the output in R is slightly different to that of STATA, but produces the same results quoted to 2 decimal places with HR 1.30 (1.05 to 1.60) for a 5 kg/m^2 increase in BMI. Slight differences between the output of different computing packages is not unusual. I found that **dosresmeta** and **glst** produced identical results using the incidence rate and case-control sample datasets that are provided with these packages and slight differences with the cumulative incidence sample dataset.

3. Using the trend estimation method to pool data reported in different forms

The trend estimation method is very useful it can be applied to derive continuous estimates from different sets of categorical risk data. This enables the pooling of diverse categorical data with data expressed on the continuous scale. Let me present a scenario to illustrate.

I've already looked at Grundvold et al (2012), who provide categorical risk data for three categories of BMI. Two other studies which report the association of BMI with incident atrial fibrillation include one by Grundvold et al (2015), who report categorical risk data for quintiles of BMI, and another by Berkovitz et al (2015), who give a HR on a continuous scale. They reported that each unit increment of BMI was associated with an increased risk of 4.3% of the development of atrial fibrillation (HR 1.04, 95% CI 1.02 to 1.07).

Applying the trend estimation method to the data from Grundvold et al (2015) estimates a HR of 1.48 (1.32 to 1.68) i.e. a 48% increased risk in atrial fibrillation associated with a 5 kg/m^2 increase in BMI. Scaling up (see post D1) the data from Berkovitch et al (2015) from a 1 kg/m^2 to a 5 kg/m^2 increase in BMI gives a HR of 1.23 (1.10 to 1.40) i.e a 23% increased risk of atrial fibrillation. Data from all three studies have now been converted into a common form (Figure) and are ready to pool in meta-analysis.
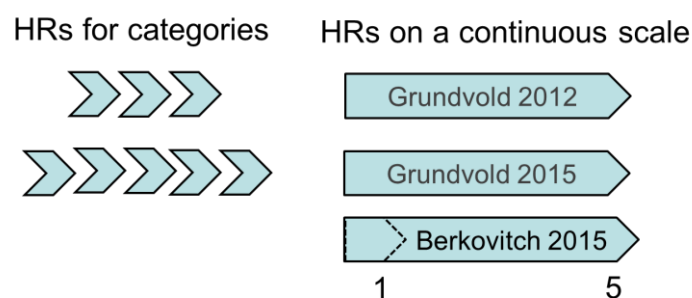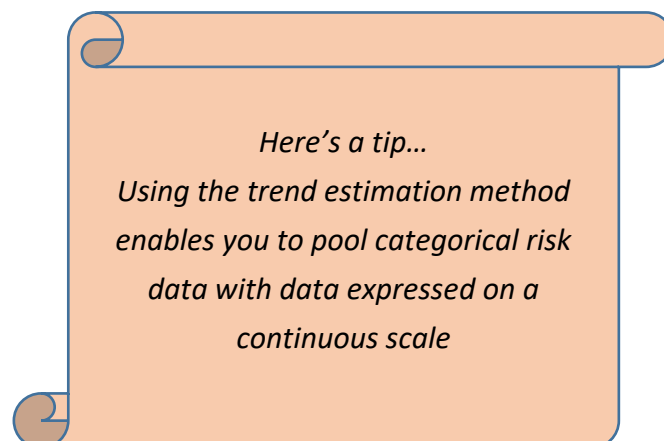

Figure. Converting varied risk data into a common form for meta-analysis

Here's a tip…
Using the trend estimation method
enables you to pool categorical risk
data with data expressed on a
continuous scale

In this post I looked at how you may deal with the problem of unbounded limits in categorical risk (dose response) data. In my next blog post I'll show how to deal with other problems that may arise with categorical risk (dose response) data.

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