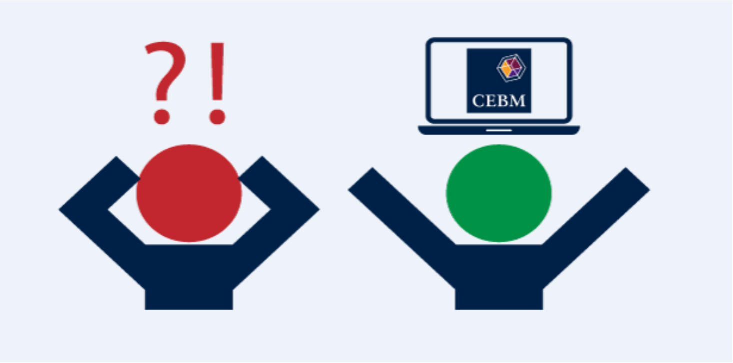
Tip for data extraction for meta-analysis - 7



How can categorical risk data be pooled?

Kathy Taylor

A common problem in meta-analysis of observational studies arises when data are reported for categories of the exposure variable, instead of for a continuous variable. Studies may divide the exposure variable into different numbers of categories, or the same number of categories, but using different thresholds, making it hard to combine them in a meta-analysis.

One way to tackle this is to estimate the trend, across the hazard ratios (odds ratios or relative risks) for the categories, to produce a hazard ratio (odds ratio or relative risk) for a unit change in the exposure variable. Greenland and Longnecker describe a trend estimation method, which Orsini et al have implemented in STATA with the glst command, Li and Spiegelman have implemented in SAS with the metadose macro, and Crippa and Orsini have implemented in R with the dosresmeta command. These packages provide estimates for single and multiple studies and the source publications are very useful as they include the underlying equations and examples.

I will describe the basic process of applying the method to a single study using STATA and R. I will refer to categorical risk data, but note that these data may also be called quantile data or dose-response data.

The basic process involves 5 steps (see below for definitions of technical terms):

1. *Establish the type of data*.

Categorical risk data will typically be given in a table, for each category reporting the hazard ratios (odds ratios or relative risks) and other data which may be either:

- *Incidence-rate data* - number of events (or cases) and number of person-years (e.g. Wolk et al , Table 2, reporting data on long-term intake of dietary fibre and risk of coronary heart disease).
- *Cumulative incidence data* - number of events and total number of people (e.g. Grundvold et al, Table 2, reporting data on body mass index and risk of atrial fibrillation).

- *Case-control data* - number of case subjects and number of control subjects (e.g. Rohan and McMichael, Table II, reporting data on alcohol intake and risk of breast cancer).

2. Set the average exposure for each category.

Wolk et al reports quintiles of fibre intake and provides a median dose of fibre for each quintile: 11.5, 14.3, 16.4, 18.8 and 22.9 g/day. These medians represent the average dose. Grundvold et al report quintiles of body mass index (BMI) and provide a range for each quintile: BMI≤ 26.2, BMI 26.3 to 28.7, BMI 28.8 to 31.3, BMI 31.4 to 34.8, and BMI ≥ 34.9 kg/m^2. The average BMI can be taken as the midpoint for the inner quintiles. The outer quintiles have one limit unbounded, so averages need to be assigned based on actual knowledge or reasonable assumptions about the range of BMIs in the population.

3. Calculate the change in exposure from that of the reference group (STATA only).

The reference group is the group to which the other groups are compared. If you are going to use STATA, calculate the change in exposure from that of the reference group; the change in exposure for the reference category will be zero. If the reported exposure for the reference category is zero, you don't need to worry, as the change in exposures will equal the reported exposures.

4. Apply the trend estimation method

The trend estimation method is based on generalised least squares. This is a technique used for estimating the unknown parameters in a linear regression model where there is some correlation between the residuals.

Below is the basic code to apply the trend estimation method to data from a single study in STATA and the corresponding code in R.

STATA CODE

First install glst

```
findit glst
```

```
glst depvar change, se(varname) cov(n cases) type
```

where

`depvar`, the dependent variable, is the log hazard ratios (log relative risks or log odds ratio) of the outcome variable.

`change`, the independent variable, is the change in exposure (dose) from that of the reference category.

`varname` in `se(varname)` is the estimate of the standard error of `depvar` `n` is the number of person-years for incidence-rate data and the total number of subjects for cumulative incidence and case-control data.

`cases` is the number of events for incidence-rate and cumulative incidence data, and the number of subjects for case-control data.

`type` is the code for the study design – “cc” for case control data, “ir” for incidence rate data and “ci” for cumulative incidence data.

R CODE

First install dosresmeta

```
install.packages("dosresmeta")
library("dosresmeta")
dosresmeta(formula, type, v, se, lb, ub, cases, n, data)
```

where

`formula` is the relationship between the outcome and the exposure (dose)

`type` is the code for the study design – “cc” for case control data, “ir” for incidence rate data and “ci” for cumulative incidence data.

Either use `v`, for the variance of the log hazard ratio (odds ratio or relative risks) or the corresponding standard error in `se` for the standard error, or `lb` and `ub` for the limits of the confidence interval.

`cases` is the number of events for incidence-rate and cumulative incidence data, and the number of subjects for case-control data.

`n` is the number of person-years for incidence-rate data and the total number of subjects for cumulative incidence and case-control data.

`data` is the data file.

5. Calculate the linear trend

Because log-transformed data were analysed it’s necessary to exponentiate (back-transform) the output to calculate the hazard ratio (relative risk or odds ratio) on the continuous scale. You may also want to rescale to a specified change in the exposure variable. Both may be achieved by using the `lincom` command in STATA or the `predict` command in R.

Here’s a tip…
There’s a trend estimation method which summarises categorical (quantile or dose-response) risk data and code is available to implement the method in STATA, SAS and R.

In my next blog post, I'll provide a worked example and present a typical scenario where trend the estimation method is useful.

Technical terms

Correlation	A measure of linear association between two continuous variables. Not to be confused with causation, which indicates that there is a causal relationship between the two variables (one causes the other).
Dependent variable	In regression, it's the variable which is supposed to be explained (or predicted) by changes in the other variable(s). Also known as a response variable or outcome.
Independent variable	In regression, it's the variable that is supposed to explain the dependent variable. Also known as an explanatory, predictor or exposure variable, or risk factor.
Linear regression	Linear regression is used to estimate the linear relationship between a dependent variable and one independent variable (simple linear regression) or more than one independent variables (multiple linear regression). The basic aim is to find a linear relationship that is provides the best summary of the data.
Log-linear regression model	This is a model which is a function which, when logarithmically transformed, becomes a linear combination of the parameters in the model so that it's possible to apply linear regression.
Reference group	A reference group is a group to which other groups are compared. With risk data, the reference group is assigned a hazard ratio (relative risk or odds ratio) of 1. If the relative risk of another group is 1.4, this means that this group has a 40% higher risk than the reference group. If the relative risk of another group is 0.8, it means that the group has a 20% lower risk than the reference group.
Residual	The difference between a data point and the value predicted by a model. In linear regression, the residual is the vertical distance between the data point and the regression line.

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