Dealing with categorical risk data when extracting data for meta-analysis

Kathryn S Taylor, Kamal R Mahtani, Jeffrey K Aronson

Nuffield Department of Primary Care Health Sciences, University of Oxford, OX2 6GG, United Kingdom

Correspondence to: K Taylor kathryn.taylor@phc.ox.ac.uk

Word count: 1498

A common problem in meta-analysis of observational studies arises when the exposure variable is categorical rather than continuous. These data may be referred to as quantile or quintile data (depending on the number of categories) or dose-response data, and in this article, the term 'categorical risk data' will be used. These data may be reported to reflect the increase in cardiovascular risk associated with increasing weight gain, alcohol consumption, or frequency of smoking. Further problems arise when studies divide the exposure variable into different numbers of categories, or the same number of categories, but using different thresholds, when data are missing, or when studies include different reference categories. These problems make it difficult to combine data in meta-analysis, but there are methods that can deal with these problems. Consider a clinical question as an example:

How is body mass index associated with the risk of incident atrial fibrillation?

Consider three studies as examples for which this question was the focus. The first study[1] provided hazard ratios (HRs) for three categories of body mass index (BMI): <25 kg/m^2, 25≤BMI<30, and ≥30 kg/m^2. The second study[2] reported HRs for quintiles. The third study[3] reported that one unit increase of BMI was associated with a 4.3% increased risk of incident atrial fibrillation (HR 1.04, 95% CI 1.02 to 1.07). This is an example of a HR reported on a continuous scale.

To carry out meta-analysis, the data from each study need to be in the same form. The same form would apply if all three studies reported HRs for incident atrial fibrillation on a continuous scale and for the same increase in BMI. For example, it may be desirable to derive a pooled HR for an increase of 5 units of BMI. It is possible to convert these HRs to those desired by estimating the trend across the categorical HRs and rescaling (Figure 1).

Figure 1. Converting varied categorical risk data into a common form for meta-analysis

For a study that reports the HR on a continuous scale, a HR for an increase of x units (HR_x) and its confidence interval can be rescaled to a HR for an increase of y units (HR_y) by the following calculation:[4]

$$HR_y = \left(HR_x \right)^{\frac{y}{x}}$$

For example, the HR reported in the third study rescales from an increase of one unit to an increase of five units (x=1; y=5) from 1.04 (95% CI 1.02 to 1.07) to 1.22 (95% CI 1.10 to 1.40).

The trend across categorical risk data can be estimated by a method based on generalised least squares,[5] which is used to estimate the unknown parameters in a linear regression model when the residuals are correlated. The residuals are the vertical distances between the data points and the regression line (error terms). The trend estimation method has been implemented in Stata with the glst command,[6] in SAS with the metadose macro,[7] and in R with the dosresmeta package.[8] These provide estimates for single studies, and also estimation and meta-analysis of data from multiple studies. To highlight the problems that can arise with implementing this method we shall focus on its use in analysing categorical risk data from a single study.

The method applies to HRs, odds ratios, and relative risks, and to outcome data that may be either incidence-rate data (expressed as the number of events, or cases, and number of person-years), cumulative incidence data (the number of events and number of people), and case-control data (the numbers of case subjects and control subjects). A worked example[9] shows implementation of the method with Stata and R for data from the first example study, which reported cumulative incidences.

Applying the trend estimation method involves several steps:
Step 1 - Establish the type of outcome data and extract these data.
Step 2 - Calculate the average exposure for each category.
Step 3 - Calculate the difference between the average exposure and the exposure in the reference category. For the first and second example studies the lowest category was the reference.
Step 4 - Apply the trend estimation method to the natural logarithm of the HRs and their standard errors, incorporating the outcome data.
Step 5 - Using a single command, the desired HR can be calculated, by exponentiating (back-transforming) the output from Step 4, and rescaling.

By producing a HR for a continuous scale, the method addresses the problems of categorical risk data and variations in the numbers of categories and thresholds. However, missing data can present problems. For example, the average exposure for each category may not be known. In a study of the association between long-term intake of dietary fibre and the risk of coronary heart disease[10] median fibre intake was reported for each category, but in our first and second example studies, the average BMI for each category was not reported. The average exposures may be estimated from the midpoint of the range of the categories, but further complications arise if the outer limits are unbounded, as is the case with the first and second example studies (e.g. BMI <25 kg/m^2).

Unbounded limits of outer categories in categorical risk data commonly pose a problem in meta-analysis of prognostic studies. If the global range of the exposure variable is reported, this will provide the outer bounds of the outer categories, but the global range is not often available. In this case, sensible estimates should be made and the impact of the choice of imputed values may be explored by sensitivity analysis.

If a confidence interval is missing, it may be estimated from a P value].[11] In a case-control study, if the numbers of patients in the case and control groups are not reported, they can be estimated from a reported odds ratio, provided the total number of first events in each group are also reported].[12] Perez *et al*[13] show how simulation can be used to derive hazard ratios when risk ratios are not reported but the numbers of cases and controls for each category and the overall mean and standard deviation of the exposure are available. They also show how a generalised linear model and the method of Chêne and Thompson[14] can be used when categorical risk data are expressed only in terms of event rates. A generalised linear model is a generalisation of ordinary linear regression, in which the usual assumption that the errors are normally distributed is relaxed. Further examples of dealing with missing categorical risk data are given by Bekkering *et al*.[15]

Another problem with categorical risk data is when the reference category in a particular study is different from those in other studies. For example, it may be that the reference category for a particular study is that of the lowest exposure, while the other studies, which may align with your study purpose, have specified their reference category as that of the highest exposure. A reference category from the lowest exposure can be switched to that of the highest exposure using simple division and reordering. This can be illustrated using data from the first example study (Figure 2).[1] To switch categories, all the HRs, lower confidence limits, and higher confidence limits need to be divided by the corresponding values of the category of the highest exposure (i.e. 1.74, 1.16, and 2.56 respectively). Then the columns for the two confidence limits need to be reordered to HR, lower limit, and higher limit.

Figure 2. Changing the reference category

Sometimes an inner category is the reference category, as in Table 1, which shows data from a study of weight change and the risk of atrial fibrillation.[16] 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, featuring the reference category in both analyses, as in a published systematic review.[17]

Table 1. Cumulative incidence data on weight change and the risk of incident 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

It is important that as much evidence as possible is summarised in meta-analyses of observational studies, but this can be hampered by inconsistent and missing data. Some data cannot be pooled (for example, if HRs are reported without a measure of variability). Poor quality of reporting of results of observational studies can lead to exclusion of studies from meta-analyses, and this will undermine the validity of systematic reviews. It is therefore important to minimise this possibility, by making use of the methods available for helping with extracting data for meta-analysis.

Acknowledgements

This research was supported by the National Institute for Health Research Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Contributions

KT and KM conceived the idea of the series of which this is one part. KT wrote the first draft of the manuscript. All authors revised the manuscript and agreed the final version.

Competing interests

Dr Mahtani and Dr Aronson were Associate Editors of BMJ Evidence Medicine at the time of submission.

References

1. Grundvold I, Skretteberg PT, Liestøl K, *et al*. Importance of physical fitness on predictive effect of body mass index and weight gain on incident atrial fibrillation in healthy middle-age men. *Am J Cardiol* 2012;110(3):425-432. doi:10.1016/j.amjcard.2012.03.043

2. Grundvold I, Bodegard J, Nilsson PM, *et al*. Body weight and risk of atrial fibrillation in 7,169 patients with newly diagnosed type 2 diabetes; an observational study. *Cardiovasc Diabetol* 2015;14:5. Published 2015 Jan 15. doi:10.1186/s12933-014-0170-3

3. Berkovitch A, Kivity S, Klempfner R, *et al*. Body mass index and the risk of new-onset atrial fibrillation in middle-aged adults. *Am Heart J* 2016;173:41-48. doi:10.1016/j.ahj.2015.11.016

4. Taylor K. What can you do when prognostic studies report measures of risk on different scales? https://www.cebm.ox.ac.uk/resources/data-extraction-tips-meta-analysis/[accessed 14 Oct 2020]

5. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *Am J Epidemiol* 1992;135(11):1301-1309. doi:10.1093/oxfordjournals.aje.a116237

6. Orsini N, Bellocco R, Greenland S. Generalized least squares for trend estimation of summarized dose–response data. *The Stata Journal* 2006; 6(1):40-57

7. Li R, Spiegelman D.The SAS % METADOSE macro. https://www.hsph.harvard.edu/donna-spiegelman/software/metadose/[accessed 14 Oct 2020]

8. Crippa A, Orsini N. Multivariate Dose-Response Meta-Analysis: The dosresmeta R Package. *Journal of Statistical Software*, Code Snippets 2016; 72(1), 1-15. doi:10.18637/jss.v072.c01

9. Taylor K. A worked example using a trend estimation method to summarise categorical risk data. https://www.cebm.ox.ac.uk/resources/data-extraction-tips-meta-analysis[accessed 14 Oct 2020]

10. Wolk A, Manson JE, Stampfer MJ, *et al*. Long-term intake of dietary fiber and decreased risk of coronary heart disease among women. *JAMA* 1999;281(21):1998-2004. doi:10.1001/jama.281.21.1998

11. Altman DG, Bland JM. How to obtain the P value from a confidence interval. *BMJ* 2011; 343 :d2304

12. Taylor K. What if something is missing from categorical risk data? https://www.cebm.ox.ac.uk/resources/data-extraction-tips-meta-analysis[accessed 14 Oct 2020]

13. Pérez T, McLellan J, Perera R. Extraction of unadjusted estimates of prognostic association for meta-analysis: simulation methods as good alternatives to trend and direct method estimation. *J Clin Epidemiol* 2018;99:153-163. doi:10.1016/j.jclinepi.2017.12.017

14. Chêne G, Thompson SG. Methods for summarizing the risk associations of quantitative variables in epidemiologic studies in a consistent form. *Am J Epidemiol* 1996;144(6):610-621. doi:10.1093/oxfordjournals.aje.a008971

15. Bekkering GE, Harris RJ, Thomas S, *et al*. How much of the data published in observational studies of the association between diet and prostate or bladder cancer is usable for meta-analysis? [published correction appears in *Am J Epidemiol* 2009;170(4):536]. *Am J Epidemiol*. 2008;167(9):1017-1026. doi:10.1093/aje/kwn005

16. Huxley RR, Misialek JR, Agarwal SK, *et al.* Physical activity, obesity, weight change, and risk of atrial fibrillation: the Atherosclerosis Risk in Communities study. *Circ Arrhythm Electrophysiol* 2014;7(4):620-625. doi:10.1161/CIRCEP.113.001244

17. Jones NR, Taylor KS, Taylor CJ, *et al*. Weight change and the risk of incident atrial fibrillation: a systematic review and meta-analysis *Heart* 2019;105:1799-1805.