Cancer Causes and Control (2005) 16:383–388
DOI 10.1007/s10552-004-5025-x

Choice of exposure scores for categorical regression in meta-analysis: a case study of a common problem

Dora Il'yasova[1,4,*], Irva Hertz-Picciotto[1,5], Ulrike Peters[2], Jesse A. Berlin[3] & Charles Poole[1]
[1]*Department Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, North Carolina, USA;* [2]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, Maryland, USA;* [3]*Center for Clinical Epidemiology and Biostatistics and Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, USA;* [4]*Departments of Cancer Biology and Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA;* [5]*Department of Epidemiology and Preventive Medicine, University of California, Davis, USA*

Received 6 October 2004; accepted in revised form 18 October 2004

Key words: colorectal cancer, dose–response relationships, epidemiology, epidemiologic method, meta-analyses, tea.

Abstract

Objective: Reporting categorical relative risk estimates for a series of exposure levels *versus* a common reference category is a widespread practice. In meta-analysis, categorical regression estimates a dose–response trend from such results. This method requires the assignment of a single score to each exposure category. We examined how closely meta-analytical categorical regression approximates the results of analysis based on the individual-level continuous exposure.
Methods: The analysis included five studies on tea intake and outcomes related to colorectal cancer. In addition, we derived categorical mean and median values from published distributions of tea consumption in similar populations to assign scores to the categories of tea intake when possible. We examined whether these derived mean and median values well approximate the individual-level results.
Results: In meta-analytical categorical regression, using the midrange scores approximated the individual-level continuous analyses reasonably well, if the value assigned to the uppermost, open-ended category was at least as high as the lower bound plus the width of the second-highest category. Categorical mean values derived from the published distributions of regular tea (in the US) and green tea (in Japan) well approximated the slope obtained from individual-level analysis.
Conclusion: Publication of both the categorical and the continuous estimates of effect in primary studies, with their standard errors, can enhance the quality of meta-analysis, as well as providing intrinsically valuable information on dose–response.

Abbreviations: CLR – confidence limit ratios; RR – relative risk

Introduction

Meta-analysis offers a quantitative method to investigate sources of variation and to reconcile inconsistent results of published observational studies. The most common approach to summarizing dose–response trend across studies draws on estimates of changes in the natural logarithm of relative risk ln(RR), per unit of exposure derived from individual studies [1]. In ideal circumstances, the original investigator has published the ln(RR) estimated from the continuous exposure data and the meta-analyst need not estimate it at all. In reality, very few publications present estimates of trend with the exposure specified as a continuous variable. Instead, category-specific estimates of relative risk, with confidence intervals,

* Address correspondence to: Dora Il'yasova, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157. Ph.: +336-713-0052; Fax: +336-716-3028. E-mail: dilyasov@wfubmc.edu

and an overall *p*-value for trend often constitute the only information available to the meta-analyst. Meta-analytical categorical regression [2] offers a method to obtain ln(RR) trend estimates from publications in which results are reported in this manner [3]. This method has been used to integrate published results on a wide spectrum of exposures and their health effects, such as environmental tobacco smoke [4], chlorinated drinking water [5], homocysteine [6], alcohol consumption [7], meat intake [8], oral contraceptive use [9], obesity [10], and others.

Most commonly used approaches to meta-analytical categorical regression require the assignment of a single score to each exposure category, and the choice of assignment can affect the summary dose–response estimate. Assigning scores to exposure categories is an important step in a meta-analysis. Not all methods rely on assigning a single score. Shi and Copas [11], for example, present a method that involves specifying a distribution of exposure values within a category. Their method, while appealing in many respects, is not easily implemented, and the basis for the distributional assumptions required may not be clear in particular situations. More commonly a single numerical value, which is thought to reflect 'typical' exposure, is assigned to each category [3]. Sometimes, category means and medians are reported in the published manuscripts or they can be obtained from the investigators. Often, however, category means and medians are unavailable. In these cases, the assigned score is usually the category midrange (*i.e.,* the point midway between the lower and upper boundaries of a category). Alternatively, the known distribution of the same exposure in a similar population might be used to estimate the mean, geometric mean, or median for each category. Other choices, not well advised but occasionally employed, include evenly spaced scores (*e.g.,* consecutive integers), lower category boundaries, and others. Evenly spaced scores were employed in Mantel's once-popular method for trend analysis [12], which estimated the trend in proportions and was therefore unsuitable for case–control data [13].

Berlin *et al.* [3] in an analysis of alcohol consumption and breast cancer, demonstrated that the choice among the mean, median, and midrange could considerably influence the estimated slope in categorical regression. In a recently published meta-analysis on tea and cardiovascular disease, we too noted that the estimated dose–response slopes were sensitive to the method of assigning category scores [14]. We were interested in how closely meta-analytical categorical regression approximates the results of continuous regression based on individual-level data. We examined how this approximation depends on the choice of category scores, and what an appropriate choice of scores would be in a typical example of a dose–response meta-analysis.

Methods

For this analysis, we used published results and data provided to us by the authors of five studies examining the relative risk of outcomes related to colorectal cancer and tea consumption [15–19]. Published results are presented in the Appendix A. We compared the estimates of trends in ln(RR) with increasing tea intake obtained from continuous individual-level regression and from meta-analytical categorical regression. We chose these studies specifically because the authors provided subject-level data that permitted the comparison of interest.

We estimated logit-linear dose–response slopes from categorical results using the covariance-corrected method described by Greenland and Longnecker [2]. The major goal of this procedure is to obtain *comparable* dose–response estimates for the association of interest from each of the studies, which are included in a meta-analysis. This method involves several steps. First, measurements of exposure, in this case tea consumption, are converted to the same measurement scale, in our example, to g/day. We estimated the typical portion size of tea cups and amount of dry tea per portion in each study from information received from the authors, the Tea Councils of the UK and the US, and experts in different countries (Dr. G. Beecher and Dr. T. Donovan in USA, Dr. P.C.H. Hollman in the Netherlands, Dr. Y. Tsubono in Japan). We used these estimates to convert measures of tea intake from cups per day to grams of dry tea per day. The second step is assigning scores to the categories of tea intake presented in the publications, which is the subject of the current analysis. The third step is calculation of beta-coefficient for the slope separately for each study using the logit-linear model and the covariance adjusted method by Greenlans and Longnecker [20].

We present 95% confidence limit ratios (CLR, the ratio of the upper confidence interval limit to the lower confidence interval limit) to measure the precision of the relative risk estimates [20]. The rationale for using CLR is presented in the earlier publication by Poole [20]. The exposure scores examined were the midranges of intake categories, since these are the most commonly available values. Specifically, we explored several approaches to assigning scores to the upper, open-ended categories of tea intake. Let b_i represent the lower bound of the ith interval, where the intervals are indexed by $i = 1,\ldots,n$.

These approaches assigned the *n*th interval score as functions of its lower bound, or of its lower bound and the width of the previous (second-to-highest) interval:$\{1.2b_n\}$; $\{b_n + 0.5\ (b_n - b_{n-1})\}$; and $\{b_n + (b_n - b_{n-1})\}$. As an additional possible approach, the intake category means and medians were assigned based on the published distributions of tea intake in similar populations. For an American study [19], we used the distribution of regular tea intake in the US population from the National Health and Nutrition Examination Survey I, Epidemiologic Follow-up Study (NHEFS) [21] and, for a Japanese study [16], the distribution of green tea intake in among Japanese published by Inoue *et al.* [22].

Results

Regression on continuously measured tea intake at the individual-level tended to provide more conservative (closer to the null) and more precise (narrower 95% CLR) estimates of the relative risk per unit of intake (6 g of dry tea daily) than the meta-analytical categorical regression on category scores (Table 1). The categorical regressions overestimated the steepness (*i.e.*, the absolute value) of the individual-level estimate by a factor of 1.2–2.2 with midrange category scores. The shallower slopes of the individual-level regression might be due to saturation of the effect at very high intake levels, *i.e.*, a greater change in the ln(RR) per unit of intake at lower intake levels of exposure than at high intake levels. This notion is supported by our results for different approaches to assigning scores to the uppermost intake category. The use of higher scores for the uppermost category yielded the closest approximation to the individual-level regression estimates (Table 1). In most studies, the highest assigned value was $\{b_n + (b_n - b_{n-1})\}$, followed by $\{b_n + 0.5\ (b_n - b_{n-1})\}$ and $\{1.2b_n\}$. The study by Kono *et al.* [18] is the only exception: there, $\{1.2b_n\} = \{b_n + (b_n - b_{n-1})\}$. Thus, when scores higher than $1.2b_n$ were assigned to the uppermost category, the trend estimate was closer to the slope based on the individual-level exposure (Table 1). In general, meta-analytical categorical regressions produced less precise slope estimates than the continuous regressions did (Table 1). Sometimes the categorical regression estimates were exceedingly imprecise, as for Baron *et al.* [16] and Su *et al.* [19], using midranges with $1.2b_n$ for the uppermost category.

For the study by Baron *et al.* conducted in the US, [16], we used publicly available data from the NHEFS [21] to calculate the category-specific mean and median of tea intakes. Similarly, we calculated categorical mean and median values for the study by Kono *et al.* [18] using the published distribution of green tea intake in the Japanese population [22]. In this unique situation, in which we were able to obtain individual-level data from some studies, we wanted to see whether using publicly available data could adequately approximate the results we would have obtained using the actual person-level data (Table 2). This approach avoids the use of midranges and functions of the lower bound of the *n*th interval, but relies on the assumption that the distribution of tea consumption in the control group in a given study is appropriately represented by values from the available data from similar populations. In both cases, categorical scores derived from published data showed similar (Kono *et al.*) or better (Baron *et al.*) results in approximating the individual-level estimates, as compared with using midrange-based values. For example, published data produced similar estimates for the study by Kono *et al.*: midrange-based estimates were 0.62, 0.58, 0.62 and those based on published distribution of green tea intake were 0.67 and 0.58. Both results similarly approximate the individual-level risk estimate −0.74.

Discussion

This exploration was limited to logit-linear models, and did not explore nonlinear relationships between exposure and outcome. In typical epidemiologic studies, however, the numbers of exposure categories available are seldom large enough to support the estimation of any but linear trends. Moreover, the trends for which '*p* for trend' values are routinely reported are almost always log-linear or logit-linear in shape. In addition, log-linear and logit-linear trends are estimated for applications of epidemiologic results, for instance in risk assessment for regulatory decisions by the US Food and Drug Administration and the US Environmental Protection Agency [23–25].

Based on an examination of a limited number (five) of sets of dose–response data relating tea consumption to colorectal cancer, we observed a tendency for estimated slopes to be further from the null when categorical regression was used to approximate the slopes that would have been obtained using the underlying continuous exposures. In a similar investigation, Berlin *et al.* [3] compared dose–response slopes estimated using national (National Health Interview Survey) mean, median, and midrange scores, the latter using $\{1.2b_n\}$ for the highest exposure category. In their analysis of

Table 1. Categorical regression of midranges and regression on a continuous variable

First author (ref)	Country	Outcome (gender)	Number of categories	RR (95% CLR) for 6 g/day increment in dry tea consumption			
				Score for the highest category			Regression on a continuous variable
				(1.2)*b_n	$b_n + 0.5(b_n - b_{n-1})$	$b_n + (b_n - b_{n-1})$	
Baron *et al.* [15]	Sweden	Rectal cancer (men & women)	3	0.55 (3.2)	0.56 (3.1)	0.65 (2.1)	0.73 (2.8)
Baron *et al.* [16]	USA	Colorectal polyps (men & women)	3	1.74 (10.5)	1.55 (6.6)	1.37 (4.0)	1.05 (2.9)
Il'yasova *et al.* [17]	Russia	Rectal cancer (men)	3	0.81 (3.7)	0.84 (3.1)	0.87 (2.8)	1.06 (2.5)
		Rectal cancer (women)	3	0.43 (2.9)	0.48 (2.5)	0.51 (2.4)	0.59 (2.0)
Kono *et al.* [18]	Japan	Colon polyps (men)	3	0.62 (4.8)	0.58 (5.8)	0.62 (4.8)	0.74 (2.9)
Su and Arab (19)	USA	Colon cancer (men & women)	3	0.35 (10.2)	0.47 (3.3)	0.62 (2.5)	0.78 (2.0)

Table 2. Categorical regression of means and medians derived from the published distributions of tea consumption

First author (ref)	RR (95% CLR) for 6 g/day increment in dry tea consumption		
	Categorical regression		Regression on a continuous variable
	Mean	Median	
Baron *et al.* (16)[†]	1.20 (2.3)	1.31 (3.6)	1.05 (2.9)
Kono *et al.* (18)[‡]	0.67 (3.8)	0.58 (5.8)	0.74 (2.9)

[†] Based on the distribution of regular tea intake in US population in National Health and Nutrition Examination Survey I, Epidemiologic Follow-up Study [21].

[‡] Based on the published distribution of green tea in the Japanese Population [22].

alcohol consumption and breast cancer incidence, the categorical regressions were performed on summary estimates published by Rohan and McMichael [26]. They reported that the categorically estimated slopes using median and midrange scores were, respectively, two and three times as high as the slope estimated with mean category scores. In this study, the slope estimate from categorical regression with mean scores was most precise. Berlin *et al.*[3] recommended the use of median scores because of their insensitivity to outliers. In our examples, median scores performed worse, especially with regard to approximating the slope estimated by individual-level continuous regression. Again, our goal was to see which methods best approximated the slope that would be estimated by using the untransformed continuous exposure measure. One might argue in favor of using a transformed version of the continuous exposure measure in some situations, but for tea intake, we limited consideration to the untransformed values. In this case study, the distributions of tea consumption were not severely skewed [15–19], which justifies use of non-transformed continuous variables.

Usually, only categorical effect estimates are available from the publications (along with the '*p* for trend,' which is not useful in estimating slopes) and in most cases, midrange scores are the only choice for the meta-analyst who does not contact investigators for supplemental information. Our results indicate that midranges may perform reasonably well, if the value assigned to the uppermost open-ended category is at least as large as the lower bound plus the width of the second highest category. This latter finding is likely a reflection of the types of distributions common for dietary data, and may not apply to more or less skewed distributions, depending on how long the uppermost tail is. Thus, for alcohol consumption, for example, this approach may not be the most appropriate.

Cook *et al.* offered an interesting approach to summarizing dose–response trend across the epidemiologic studies of different design and with different

categorization of the exposure [27]. However, this method does not address data with open-ended categories and therefore, is not applicable to meta-analysis of typical nutritional exposures. Thus, meta-analytical categorical regression [2] remains an attractive and readily accessible approach to summarizing dose–response trend across the studies for continuous exposure. A new extension to this method has been presented by Bagnardi *et al*. [28]. The authors use fractional polynomials and cubic splines to fit smooth dose–response relations in summarizing meta-analytical aggregate data. This method still requires assignment of category scores and we believe it is more than reasonable to assume that the choice of the score for the open ended category would substantially influence the results produced by this method.

Often, meta-analysts need to seek other information from original investigators. In these cases, requests for means and medians could be made as well. Ideally, however, investigators would supplement, or supplant, the routine reporting of '*p* for trend' with the reporting of 'beta for trend' from continuous analysis , along with a confidence interval or estimated standard error. This improvement in reporting practice would gradually obviate the need for systematic reviewers to conduct categorical regression analyses. Meta-analysts would not be the only ones to benefit from this improvement in reporting. If policy is to be influenced by estimates of relative risk for particular levels of exposure, estimates of the increase in risk *per unit* of exposure would seem to us to be most helpful.

The generalizability of these results remains to be confirmed and, in any case, will depend on two phenomena: first, the distribution of exposures within the categories selected, with the lowest and uppermost categories having the most influence; and secondly, the shape of the true dose–response function, *i.e*., the extent to which it differs from linearity in the log or logit of the outcome measure. Some caveats, should, however, be mentioned. We have focused on the agreement between the slopes estimated from categorical regression and those estimated using the underlying continuous exposure measurement. In adopting this focus, we have assumed that the continuous dose–response is a more accurate representation of the true dose–response than the categorical one. This assumption could be incorrect if measurement error were distorting (attenuating) the continuously estimated dose–response function more than it distorts the categorically estimated function [29]. Our assumption in presenting this example was that most slope estimates from studies of epidemiologic associations would not be corrected for measurement error.

Until all the factors influencing a dose–response are understood, publication of both the categorical and the continuous estimates of effect, as well as their standard errors (or confidence intervals), can enhance the quality of meta-analysis. We recommend that such an approach become a standard practice for epidemiologic investigations. Reporting only categorical effect estimates and a '*p* for trend' does not quantify any estimate of a dose–response relationship, cannot be used in a meta-analysis or systematic review, does not provide adequate information to policy makers, and should not remain the routine manner in which dose–response results are reported. The very fact that the boundaries of exposure categories vary makes results inherently incomparable across studies. We believe that weakness alone argues in favor of reporting slopes.

We also recommend that authors routinely report means and medians for categories, along with category boundaries, rather than merely the category boundaries or the completely uninformative (and semantically incorrect) 'quintiles,' 'tertiles,' etc. When meta-analysts must perform categorical regressions because continuous regression results are not available, we recommend using all three score choices – mean, median, and midrange with more than one approach to assigning scores to the highest categories for midrange scores. If meta-analytic results are sensitive to these choices, this sensitivity should be reported. The meta-analytic results that could be sensitive to these choices include: the *p*-value from the overall homogeneity test, all components of a publication bias analysis, stratified or meta-regression analysis of study characteristics, and the summary effect estimate and its confidence interval [30].

Appendix. Published categorical estimates from the epidemiological studies considered in this analysis

First author (ref)	Country	Outcome (gender)	Tea intake categories	RR or OR
Baron *et al*. [15]	Sweden	Rectal cancer (men & women)	0 1 cup/day ≥2 cups/day	1 ref 1.06 (0.74–1.52) 0.56 (0.34–0.90)
Baron *et al*. [16]	USA	Colorectal polyps (men & women)	0 < 1 cup/day ≥1 cups/day	1 ref 1.12 (0.75–1.66) 1.29 (0.75–2.22)

Appendix. (Continued)

First author (ref)	Country	Outcome (gender)	Tea intake categories	RR or OR	
Il'yasova *et al.* [17]	Russia	Rectal cancer	< 80 g/month	Men	Women
			80–160 g/month	1 ref	1 ref
			> 160 g/month	0.61 (0.38–1.01)	0.48 (0.29–0.78)
				0.77 (0.42–1.43)	0.40 (0.23–0.70)
Kono *et al.* [18]	Japan	Colon polyps (men)	< 3 cups/day	1 ref	
			3–4 cups/day	0.81	
			≥5 cups/day	0.69	
Su and Arab [19]	USA	Colon cancer (men & women)	0	1 ref	
			≤1.5 cups/day	0.57 (0.42–0.78)	
			> 1.5 cups/day	0.59 (0.35–1.00)	

References

1. Greenland S (1998) Basic methods for sensitivity analysis and external adjustment. In: Rothman KJ, Greenland S, eds. *Modern Epidemiology*. Philadelphia: Lippincott-Raven Publishers, pp. 343–358.
2. Greenland S, Longnecker MP (1992) Methods for trend estimation from summarized dose–response data, with applications to meta-analysis. *Am J Epidemiol* **135**: 1301–1309.
3. Berlin JA, Longnecker MP, Greenland S (1993) Meta-analysis of epidemiologic dose–response data. *Epidemiology* **4**: 218–228.
4. Brown KG (1999) Lung cancer and environmental tobacco smoke: occupational risk to nonsmokers. *Environ Health Perspect* **107** (**Suppl 6**): 885–890.
5. Villanueva CM, Fernandez F, Malats N, Grimalt JO, Kogevinas M (2003) Meta-analysis of studies on individual consumption of chlorinated drinking water and bladder cancer. *J Epidemiol Community Health* **57**: 166–173.
6. Ford ES, Smith SJ, Stroup DF, Steinberg KK, Mueller PW, Thacker SB (2002) Homocyst(e)ine and cardiovascular disease: a systematic review of the evidence with special emphasis on case–control studies and nested case-control studies. *Int J Epidemiol* **31**: 59–70.
7. Reynolds K, Lewis B, Nolen JD, Kinney GL, Sathya B, He J (2003) Alcohol consumption and risk of stroke: a meta-analysis. *JAMA* **289**: 579–588.
8. Norat T, Lukanova A, Ferrari P, Riboli E (2002) Meat consumption and colorectal cancer risk: dose–response meta-analysis of epidemiological studies. *Int J Cancer* **98**: 241–256.
9. Spitzer WO, Faith JM, MacRae KD (2002) Myocardial infarction and third generation oral ontraceptives: aggregation of recent studies. *Hum Reprod* **17**: 2307–2314.
10. Bergstrom A, Hsieh CC, Lindblad P, Lu CM, Cook NR, Wolk A (2001) Obesity and renal cell cancer – a quantitative review. *Br J Cancer* **85**: 984–990.
11. Shi JQ, Copas JB (2004) Meta-analysis for trend estimation. *Stat Med* **23**: 3–19.
12. Mantel N (1963) Chi-Square Tests with One Degree of Freedom: Extensions of the Mantel-Haenszel Procedure. *J Am Stat Assoc* **53**: 690–700.
13. Rothman KJ (1986) *Modern epidemiology*. Boston: Little, Brown.
14. Peters U, Poole C, Arab L (2001) Does tea affect cardiovascular disease? A meta-analysis. *Am J Epidemiol* **154**: 495–503.
15. Baron JA, Gerhardsson D, Ekbom A (1994) Coffee, tea, tobacco, and cancer of the large bowel. *Cancer Epidemiol Biomarkers Prev* **3**: 565–570.
16. Baron JA, Greenberg ER, Haile R, Mandel J, Sandler RS, Mott L (1997) Coffee and tea and the risk of recurrent colorectal adenomas. *Cancer Epidemiol Biomarkers Prev* **6**: 7–10.
17. Il'yasova D, Arab L, Martinchik A, Weisgerber UM, Sdvizhkov A, Urbanovich L (2003) Black tea consumption and risk of rectal cancer in Moscow population. *Ann Epidemiol* **13**: 405–411.
18. Kono S, Shinchi K, Ikeda N, Yanai F, Imanishi K (1991) Physical activity, dietary habits and adenomatous polyps of the sigmoid colon: a study of self-defense officials in Japan. *J Clin Epidemiol* **44**: 1255–1261.
19. Su LJ, Arab L (2002) Tea consumption and the reduced risk of colon cancer – results from a national prospective cohort study. *Public Health Nutr* **5**: 419–425.
20. Poole C (2001) Low P-values or narrow confidence intervals: which are more durable? *Epidemiology* **12**: 291–294.
21. US Department of Health and Human Services (DHHS). Second National Health and Nutrition Examination Survey I Epidemiologic Followup Study, 1982–1984. NCHS. 2001. NCHS.
22. Inoue M, Tajima K, Hirose K, *et al.* (1997) Epidemiological features of first-visit outpatients in Japan: comparison with general population and variation by sex, age, and season. *J Clin Epidemiol* **50**: 69–77.
23. U.S. Environmental Protection Agency (1986) Guidelines for Carcinogen Risk Assessment. *Fed. Reg.* **51**: 33992–34003.
24. Hertz-Picciotto I (1995) Epidemiology and quantitative risk assessment: a bridge from science to policy. *Am J Public Health* **85**: 484–491.
25. Shore RE (1995) Epidemiologic data in risk assessment – imperfect but valuable. *Am J Public Health* **85**: 474–476.
26. Rohan TE, McMichael AJ (1988) Alcohol consumption and risk of breast cancer. *Int J Cancer* **41**: 695–699.
27. Cook RJ, Brumback BB, Wigg MB, Ryan LM (2001) Synthesis of evidence from epidemiological studies with interval-censored exposure due to grouping. *Biometrics* **57**: 671–680.
28. Bagnardi V, Zambon A, Quatto P, Corrao G (2004) Flexible meta-regression functions for modeling aggregate dose–response data, with an application to alcohol and mortality. *Am J Epidemiol* **159**: 1077–1086.
29. Hertz-Picciotto I, Smith AH (1993) Observations on the dose–response curve for arsenic exposure and lung cancer. *Scand J Work Environ Health* **19**: 217–226.
30. Poole C, Greenland S (1999) Random-effects meta-analyses are not always conservative. *Am J Epidemiol* **150**: 469–475.