Discriminating Among Correlated Aspects of Exposure

Robert P Hirsch[1]
[1]Stat-Aid Consulting, 6100 W 99th Terrace, Overland Park, KS 66207

Abstract

In epidemiologic research, exposures often have several aspects among which we wish to discriminate. For instance, many environmental exposures can be separated into years of exposure, maximum intensity of exposure, cumulative exposure, and age at first exposure, to name a few. One characteristic of these aspects is they are usually correlated. This characteristic makes it difficult to evaluate the independent contributions of the individual aspects. To draw distinctions among these aspects it is necessary to control for the others, as completely as possible, while evaluating one. This often results in conflict between control of individual characteristics and interpretation of their relationship to the risk of disease. A hybrid method that provides both interpretability and discrimination is described. This method is evaluated using a simulation. As an example of its use, the relationship between aspects of cigarette smoking and lung cancer is investigated.

Key Words: case-control study; categorization; etiologic fraction; exposure; lung cancer; polynomial; simulation; smoking

1. Introduction

Once it has been established that a characteristic is a risk factor for a particular disease it is often of interest to determine the aspects of that characteristic that independently contribute to development of the disease. This is often the case when characterizing behavioral exposures (Al Kazzi, et al 2015, Chivese et al 2015, Mohammed et al 2016), for example smoking behavior. It is also commonly encountered when interested in occupational exposures (Attfield et al 2012, Mattioli et al 2012, Schramm et al 2015) or environmental exposures (Lee et al 2015, Turner et al 2014, Vicedo-Cabera et al 2015). In these studies, exposure can often be expressed as duration of exposure, maximum exposure, mean exposure, and cumulative exposure, for example. It is important to be able to distinguish among these aspects, for the independent contributions of the aspects have etiologic and health policy implications (Turner et al 2014). For example, if duration of exposure has the greatest independent contribution to the risk of disease, then interventions that change the duration of exposure are most likely to have the greatest impact on disease occurrence.

There have been attempts to distinguish among quantitative representations of exposure using either continuous independent variables or categories of those continuous values (Turner et al 2010). Each has its advantages and disadvantages. The use of continuous independent variables in linear representations has the advantage of providing the potential to describe dose-response relationships, but this can be realized only if the relationship between the aspect of exposure and the occurrence of disease is linear, at least on some scale. This is often not the case (Greenland 1995). A common remedy is categorizing the continuous variables. This has the advantage of, not only releasing the requirement for linearity, but also makes interpretation more straightforward. Categorization has two disadvantages as well. First, one needs to decide how to define categories. Most often this is done by using quantiles, which are unlikely to have particular biologic correlates (Taylor and Yu 2002). In addition, categories do not account for all the variation in a quantitative representation of exposure (Taylor and Yu 2002). This is an important disadvantage. Since the purpose in interpreting aspects of exposure is to determine the independent contributions of the various aspects, all of the variation for each of the other aspects must be accounted for when examining a particular aspect. If this is not accomplished, aspects that have little or no independent contribution will appear to have a contribution due to the correlation among aspects of exposure (Mohammed et al 2016).

There is another approach that has been suggested for the control of confounding (a related concept). That is the use of polynomial functions (Greenland 1995, Williams 2001, Brenner and Blettner 1997). Polynomial functions have the advantage of allowing complete control, but they have a disadvantage in that they are difficult to interpret from a biologic point of view. This article describes a hybrid method that provides complete control and interpretability at the same time.

2. Proposed Method

I propose a hybrid approach combining categories and polynomial functions. In this method, separate analyses are done for each of the aspects of exposure. In the analysis for a particular aspect of exposure, that aspect is represented by categories (maximizing interpretability), while all of the other aspects are represented by polynomial functions (maximizing control). This analysis only provides information about the particular aspect of exposure. The polynomial functions are not interpreted. They are included only to provide nearly complete control for other aspects of exposure when examining one aspect.

Then, the next aspect is represented by categories while all other aspects (including the first aspect) are represented by polynomial functions. This is repeated until all aspects of exposure have been represented by categories. Since this analysis is designed to examine aspects of exposure and not exposure itself, only exposed persons are included in these analysis (Robertson et al 1994).

3. Simulation

To evaluate the proposed method, a computer simulation created in Excel using visual basic, is used. In this simulation, a population is considered to have a continuous exposure with three aspects: age at initiation of exposure, maximum exposure, and cumulative exposure. Each aspect is assigned four categories defined by quartiles. Built into the simulation is an algorithm that allows only cumulative exposure to influence the probability of developing the disease as the members of the population are followed over time although all of the aspects are highly correlated with each other (Table 1). If the proposed method works, we should see the influence of cumulative exposure without suggestions of influence of age or maximum exposure.

Table 1. Correlations among aspects of exposure in simulation.

	Age at Initiation	Maximum Exposure	Cumulative Exposure
Age at Initiation	1.000	-0.873	-0.990
Maximum Exposure	-0.873	1.000	0.886
Cumulative Exposure	-0.990	0.886	1.000

The data from exposed individuals from this simulation are analyzed using logistic regression analyses (SPSS 24). The results are expressed as the etiologic fraction among the exposed (Klienbaum et al 1982), since this more relevant to evaluation of aspects of exposure to the development of disease. The etiologic fraction among the exposed (*EF_e*) is calculated from odds ratios (*OR*) estimated from the results of logistic regression analysis. For odds ratios from one to positive infinity, it is calculated as follows (Cole and MacMahon 1971):

$$EF_e = \frac{OR-1}{OR}$$

For odds ratios less than one, the negative etiologic fraction (or negative "preventive" fraction) is calculated as follows (Miettinen 1974):

$$EF_e = -\frac{\frac{1}{OR}-1}{\frac{1}{OR}} = -(1-OR)$$

The etiologic fraction is interpreted as the proportion of exposed persons who develop the disease due to that aspect of exposure. Negative etiologic fractions indicate a protective relationship with the aspect.

For each aspect of exposure, results for three models are examined. These include crude analysis (Crude) not controlling for other aspects, categorical analysis (Categorical) controlling for other aspects of exposure by using all of the aspects represented by categories, and hybrid analysis (Polynomial) using polynomial functions to control for the other aspects. Typical results from the simulation are summarized in Table 2 and Figures 1-3.

Table 2. Typical results of simulation. Etiologic fraction among exposed and 95% confidence interval.

		Level of Exposure*					
Aspect	Method	2		3		4	
Age at Initiation	Crude	-0.90	-0.91, -0.90	-0.94	-0.95, 0.93	-0.95	-0.96, -0.94
	Categorical	-0.53	-0.62, -0.42	-0.47	-0.64, -0.23	-0.58	-0.75, -0.28
	Polynomial	0.04	-0.17, 0.24	0.25	-0.14, 0.52	-0.17	-0.55, 0.36
Maximum Exposure	Crude	0.58	0.50, 0.65	0.72	0.66, 0.76	0.82	0.78, 0.85
	Categorical	-0.02	-0.22, 0.19	0.04	-0.18, 0.24	0.23	0.02, 0.39
	Polynomial	0.01	-0.20, 0.22	-0.03	-0.25, 0.20	-0.03	-0.26, 0.21
Cumulative Exposure	Crude	0.03	-0.17, 0.22	0.08	-0.12, 0.26	0.95	0.96, 0.97
	Categorical	-0.18	-0.46, 0.19	-0.23	-0.54, 0.23	0.92	0.86, 0.95
	Polynomial	-0.25	-0.50, 0.11	-0.19	-0.54, 0.30	0.89	0.80, 0.94

*Lowest level of exposure is the index level

Figure 1. Age at initiation of exposure from simulation.

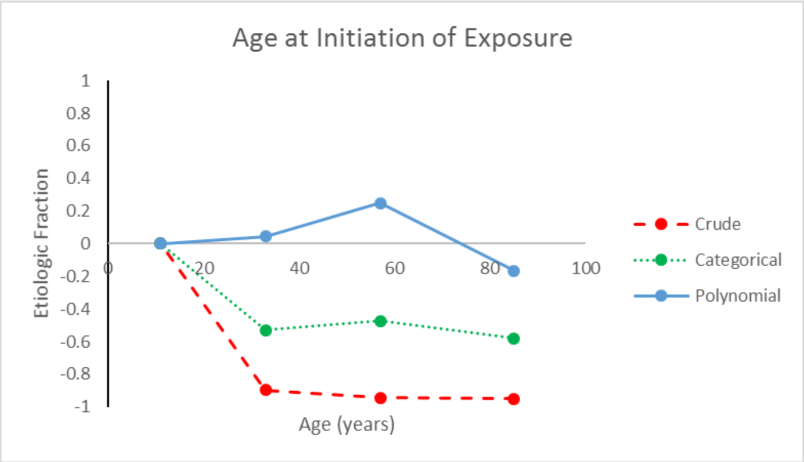


Figure 2. Maximum Exposure from simulation.

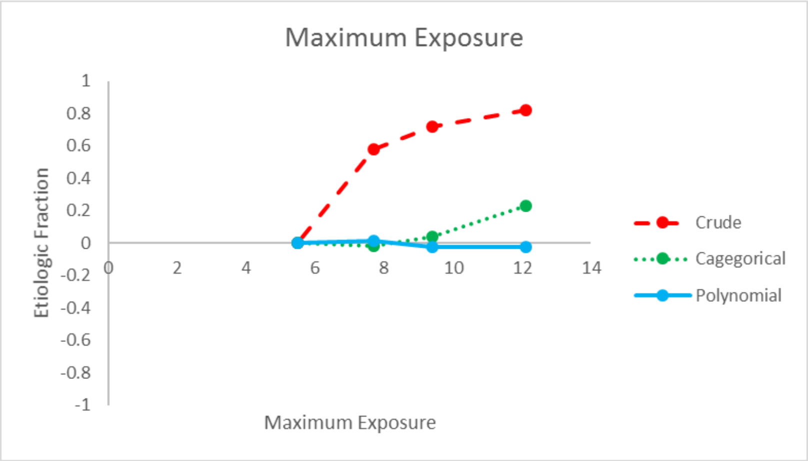
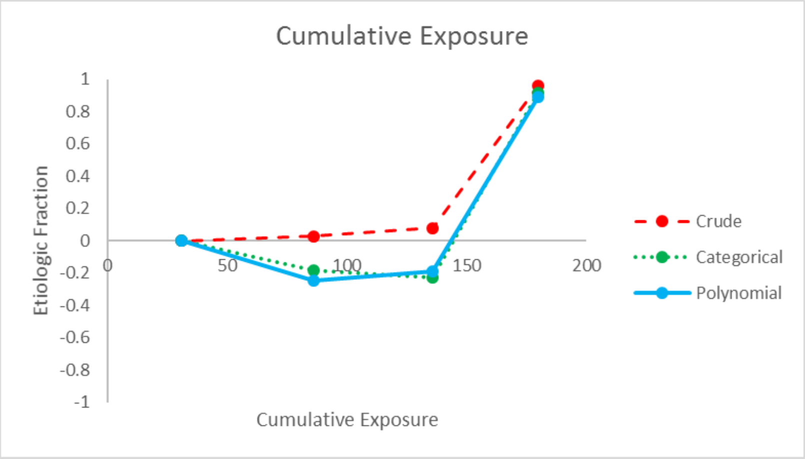


Figure 3. Cumulative Exposure from simulation.



There are two important features of the relationship between method and etiologic fraction. First, the degree of bias is always greatest for the crude analysis and least for the polynomial analysis. This reflects the degree of control for other aspects of exposure provided by each method. Second, the only the hybrid method with polynomial control shows a statistically significant relationship with the cumulative exposure aspect to the exclusion of other aspects. Thus, the proposed method correctly represents the “biologic” relationship built into the simulation.

4. Smoking and Lung Cancer

To demonstrate the use of the proposed method on an actual data set, data were obtained from a large European case-control study of smoking and lung cancer (Phillip Morris International). These data include information from 6,674 smokers. For these data, five aspects of smoking behavior were assessed: age at initiation, duration of smoking, mean annual packs smoked, mean annual tar yield of cigarettes smoked, and maximum annual tar yield of cigarettes smoked. Correlations among those aspects are in Table 3.

Table 3. Correlations among aspects of exposure in case-control study.

	Age at Initiation	Duration of Exposure	Mean Annual Packs	Mean Annual Tar	Maximum Annual Tar
Age at Initiation	1.000	0.459	0.248	-0.086	0.219
Duration of Exposure	0.459	1.000	0.228	0.019	0.443
Mean Annual Packs	0.248	0.228	1.000	0.486	0.575
Mean Annual Tar	-0.086	0.019	0.486	1.000	0.825
Maximum Annual Tar	0.219	0.443	0.575	0.825	1.000

Four categories of each aspect were defined by quartiles. The results of these analyses are summarized in Table 4 and Figures 4-8.

Table 4. Results of analyzing case-control data. Etiologic fraction among exposed and 95% confidence interval.

		Exposure Level*					
Aspect	Method	2		3		4	
Age at Initiation	Crude	-0.21	-0.34, -0.05	-0.22	-0.34, -0.08	-0.39	-0.50, -0.26
	Categorical	-0.07	-0.23, 0.11	0.04	-0.13, 0.20	0.03	-0.17, 0.22
	Polynomial	-0.05	-0.22, 0.14	0.07	-0.10, 0.23	0.17	-0.04, 0.33
Duration	Crude	0.60	0.53, 0.66	0.71	0.65, 0.76	0.75	0.70, 0.79
	Categorical	0.58	0.50, 0.64	0.69	0.63, 0.75	0.76	0.71, 0.80
	Polynomial	0.57	0.49, 0.64	0.68	0.62, 0.74	0.76	0.70, 0.80
Mean Annual Packs	Crude	0.48	0.39, 0.56	0.52	0.43, 0.59	0.61	0.54, 0.68
	Categorical	0.36	0.24, 0.46	0.38	0.25, 0.49	0.48	0.36, 0.59
	Polynomial	0.08	-0.15, 0.27	-0.12	-0.38, 0.20	-0.33	-0.62, 0.17
Mean Annual Tar	Crude	0.28	0.15, 0.40	0.41	0.29, 0.50	0.38	0.27, 0.48
	Categorical	0.12	-0.06, 0.28	0.25	0.06, 0.39	0.24	0.03, 0.41
	Polynomial	0.17	-0.01, 0.32	0.29	0.12, 0.43	0.33	0.10, 0.46
Maximum Annual Tar	Crude	0.41	0.30, 0.50	0.52	0.43, 0.59	0.57	0.49, 0.64
	Categorical	0.27	0.11, 0.40	0.27	0.08, 0.43	0.23	-0.02, 0.42
	Polynomial	0.16	-0.05, 0.35	0.17	-0.08, 0.36	0.11	-0.18, 0.35

*Lowest level of exposure is the index level

Figure 4. Age at initiation of cigarette smoking from case-control study.

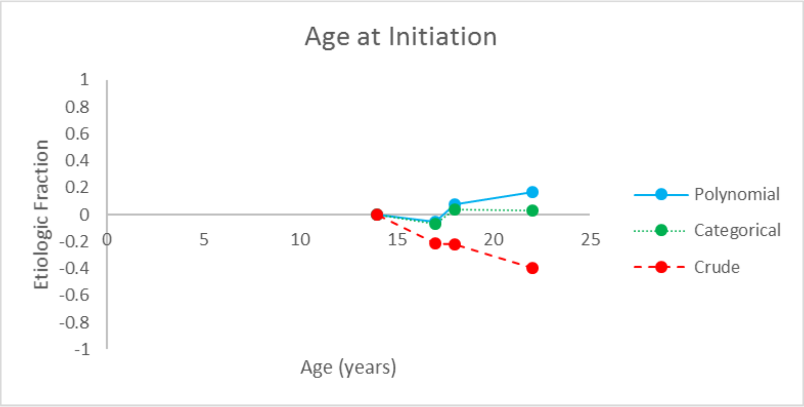


Figure 5. Duration of cigarette smoking from case-control study.

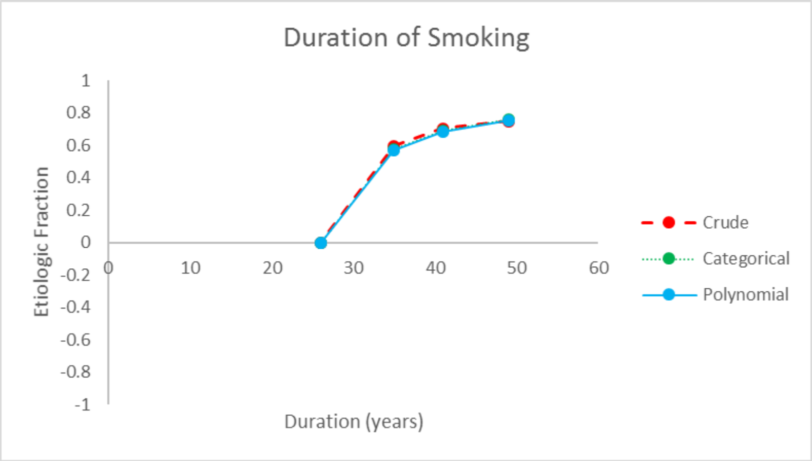


Figure 6. Mean annual packs of cigarettes smoked from case-control study.

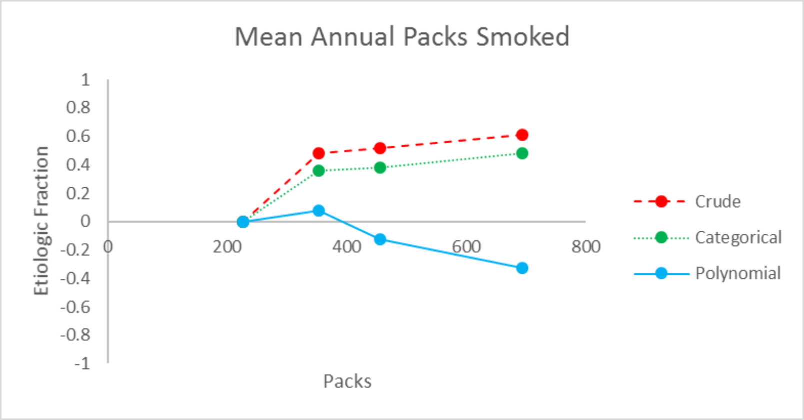


Figure 7. Mean annual tar yield of cigarettes smoked from case-control study.

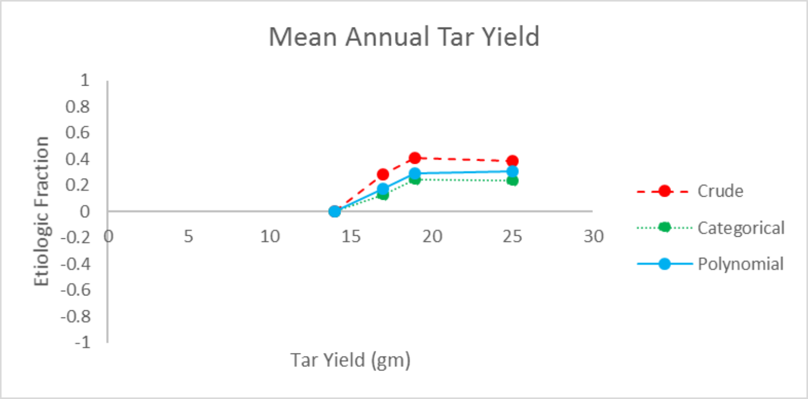
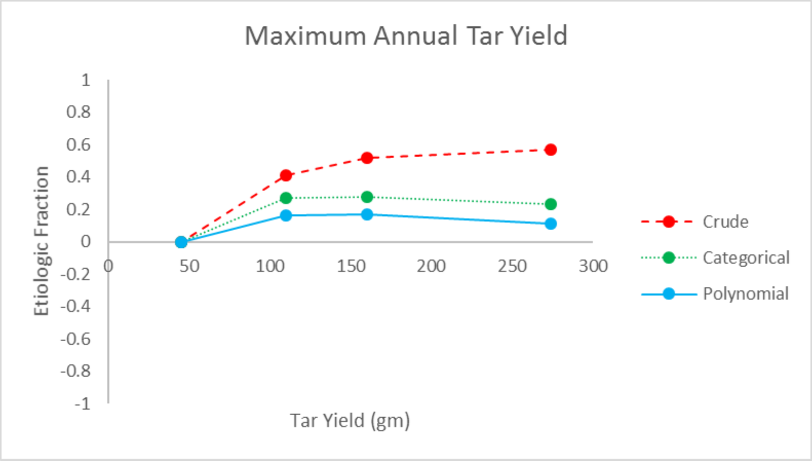


Figure 8. Maximum annual tar yield of cigarettes smoked from case-control study.



Two of the five aspects of cigarette smoking behavior are statistically significant when using the hybrid method with polynomial control. They are the duration of smoking and mean annual tar yield of cigarettes smoked. The other three aspects are not statistically significant for the hybrid method. For the categorization method, we observe that mean annual packs and maximum annual tar yield also are statistically significant. For the crude method, in which there is no control for correlations among the aspect of exposure, all five aspects of exposure are statistically significant.

5. Discussion

For the simulation, we know that the cumulative duration of exposure is the only aspect of exposure that determines the occurrence of disease. Thus, we know that the hybrid method got the right answer. The commonly used categorical method incorrectly indicates that maximum exposure and age at initiation are also associated with occurrence of the disease. These are incorrect conclusions created by the poorly controlled correlations between these aspects and cumulative exposure. This demonstrates the utility of the hybrid approach with polynomial control.

For the case-control data, we do not know the truth of which of the aspects of exposure are independently associated with the occurrence of disease, but we can see a difference in the impressions left by the various methods. All three methods agree that duration of exposure and mean annual tar yield of the cigarettes smoked contribute of the occurrence of disease. The hybrid method limits the independent aspects of exposure to these two. The categorical method includes mean annual packs and maximum annual tar yield as additional, apparently independent aspect of cigarette smoking behavior. This inclusion is likely to be due to incomplete control of the correlation between packs smoked and mean annual tar yield (*r*=0.486) and between maximum annual tar yield and mean annual tar yield (0.825).

The proposed hybrid method using polynomials to control for the correlation among aspects of exposure works well. This suggests it could be used when evaluating exposures with quantitative representations of aspects of exposure.

Acknowledgements

I appreciate access to the case-control data granted by Dr. Myron Weinberg of the Weinberg Group, Washington, DC.

References

Al Kazzi ES, Lau B, Li T, et al. Differences in the prevalence of obesity, smoking and alcohol in the United States nationwide inpatient sample and the behavioral risk factor surveillance system. *Plos ONE* 2015;10(11):e140165.

Attfield MD, Schleiff JH, Lubin AB, et al. The diesel exhaust in miners study: A cohort mortality study with emphasis on lung cancer. *Journal of the National Cancer Institute* 2012;104:869-883.

Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 1997;8:429-434.

Chivese T, Esterhulzen TM, Basson AR. The influence of second-hand cigarette smoke exposure during childhood and active cigarette smoking on Crohn’s disease phenotype defined by Montreal classification scheme in a Western Cape population, South Africa *PLoS ONE* 2015;10(9):e0139597.

Cole P, MacMahon B. Attributable risk percent in case-control studies. *British Journal of the Society of Preventive Medicine* 1971;25:242-244.

Greenland S. Dose-response and trend analysis in epidemiology: Alternatives to categorical analysis. *Epidemiology* 1995;6:356-365.

Klienbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research Principles and Quantitative Methods* New York; John Wiley & Sons, Publishers;1982:160-164.

Lee YK, Ju YS, Lee WJ, et al. Assessment of radiation exposure from cesium-137 contaminated roads for epidemiological studies in Seoul, Korea. *Environmental Health and Toxicology* 2015;30:e2015005.

Mattioli S, Curti S, De Fazio R, et al. Occupational lifting tasks and retinal detachment in non-myopics and myopics: Extended analysis of a case-control study. *Safety and Health at Work* 2012;3:52-57.

Miettinen O. Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology* 99:325-332.

Mohammed MJ, Rakhimov IS, Shitan M, et al. A new mathematical evaluation of smoking problem based of algebraic statistical method. *Saudi Journal of Biologic Science* 2016;23:S11-S15.

Phillip Morris International, CTOR dataset 2006-2009, obtained from Myron Weinberg, the Weinberg Group, Washington, DC.

Robertson C, Boyle P, Hsieh CC, et al. Some statistical consideration in the analysis of case-control studies when the exposure variables are continuous measurements. *Epidemiology* 1994;5:164-170.

Schramm A, Uter W, Brant m, et al. Increased intima-media thickness in rayon workers after long-term exposure to carbon disulfide. *International Archive of Occupational and Environmental Health* (PMID26452498) 2015.

Taylor JMG, Yu M. Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis* 2002;83:248-263.

Turner EL, Dobson JE, Pocock SJ. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiologic Perspective and Innovation* 2010;doi:10.1186/1742-5573-7-9.

Turner MC, Benke G, Bowman JD, et al. Occupational exposure to extremely low frequency magnetic fields and brain tumour risks in the INTEROCC study. *Cancer Epidemiology, Biomarkers, and Prevention* 2014;23(9):1863-1872.

Vicedo-Cabrera AM, Olsson D, Forsberg B. Exposure to seasonal temperatures during the last month of gestation and the risk of preterm birth in Stockholm. *International Journal of Research in Public Health* 2015;12:3962-3978.

Williams JS. Assessing the use of fractional polynomial methods in health services research: a perspective on the categorization epidemic. *Journal of Health Services Research and Policy* 2001;16(3):147-152.