Adjustment for smoking in lung cancer analyses in the EPIC cohort

Boshuizen H.C.[1], Bueno-de-Mesquita H.B.[1], Altenburg H.P., Agudo A., Le Marchand L., Berrino F., Janzon L., Rasmuson T., Vineis P., Lukanova A., Linseisen J., Riboli E., Miller A.
(for the EPIC working group on Lung Cancer)

[1]**National Institute for Public Health and the Environment, Bilthoven, The Netherlands.**

Introduction

As smoking is strongly related to lung cancer, proper adjustment for smoking is essential in assessing its associations with other factors (Whittemore, 1988). Several methods of adjustment are possible. We will compare the following procedures to adjust for smoking:

1. Number of pack-years;
2. The six-category index: never smoker, former cigarette smoker who stopped more than 10 years ago; former cigarette smoker who stopped less than 10 years ago; currently smoking <15 cigarettes/day; currently smoking 15–24 cigarettes/day; currently smoking ≥ 25 cigarettes/day;other (pipe/cigar/insufficient information);
3. The best fitting model (based on a Schwartz' Bayesian Criterion-type criterion) selected from all available variables on smoking status, intensity and duration, including interaction terms.

Methods

We used the EPIC follow-up data as released by IARC on 25 May 2001, excluding Greece (data not yet complete), prevalent lung cancer cases (*n*=188), those completely lost to follow-up (*n*=534), those with follow-up only in the period when cancer-registry data were judged to be incomplete (*n*=4051) and those with unreliable dietary data (ratio of reported dietary intake to estimated caloric requirement in the 2% most extreme values) (*n*=8157), resulting in a cohort of 441 426 persons of whom 608 developed lung cancer. Of this cohort, 3716 persons were lost to follow-up, roughly half of them because of migration. Dietary data were available on 407 131 persons (499 lung cancer cases). Data were analysed with the Cox proportional hazards model, using age as the time variable and stratification on age at baseline (1-year intervals), gender and region. We studied the effect of different adjustment procedures on the relative risk estimates for quartiles of fruit and vegetable consumption. Results presented are for uncalibrated dietary data. Very similar results were obtained when analyses were repeated with crudely calibrated dietary data, using a multiplication factor which centred the mean of the food frequency questionnaire data in each study centre on the mean of the 24-h recall data from this centre. Dietary data were divided into quartiles, with gender-specific quartile boundaries based on the entire cohort. Quartile boundaries for uncalibrated vegetable consumption were 93, 146 and 227 g/day for men and 124, 191 and 286 g/day for women. For uncalibrated fruit consumption, they were 84, 154 and 273 g/day for men and 131, 226 and 338 g/day for women. All analyses with dietary variables were adjusted for reported total energy intake, height and weight.

Results

The data-driven model (approach 3) contained the following baseline smoking variables: current smoker, former smoker, number of cigarettes currently smoked, duration of smoking and a quadratic term for duration, inhaling (0= not, 1= not deeply, 2= deeply), interaction term between genders and number of cigarettes currently smoked. The data-driven approach yielded the best fitting model, while the pack-years approach fitted the worst. Coefficients for quartiles of vegetable and fruit intake differed between adjustment methods with a magnitude up to the coefficient's standard error (Figure 1). Similar results were obtained when excluding the cases diagnosed within the first 2 years of follow-up from the analysis (Figure 2).

Discussion

As expected, the data-driven approach yields a better-fitting model than a *priori* specified models, and the differences between the crude and the adjusted model are also the largest for this model. It should be remembered that the best model described above relates to lung cancer. For other cancers, the relationship with smoking is likely to be different; thus the best adjustment strategy might also differ. Moreover, the issue of residual confounding by smoking will be less substantial for other cancers, as their association with smoking is weaker to begin with.

Our results suggest that outcomes from models using only pack-years as an adjusting variable may be affected by significant residual confounding compared to data-driven adjustment. However, our analyses do not answer the question of whether residual confounding of smoking is still present in the data-driven model. In principle, residual confounding could still occur as a result of miss-specification of the model or because of important unmeasured aspects of smoking. Although the model was selected from a large set of possible models, this set could have been further extended by including other interaction or polynomial terms, or compounded variables of the basic variables. However, given the large number of models already considered, we feel a considerable improvement in fit and adjustment potential is not to be expected. Residual confounding is more likely to occur due to measurement error present in the smoking data. Information on past smoking behaviour was collected by retrospective recall and thus is prone to some misclassification. Similarly, misclassification could be

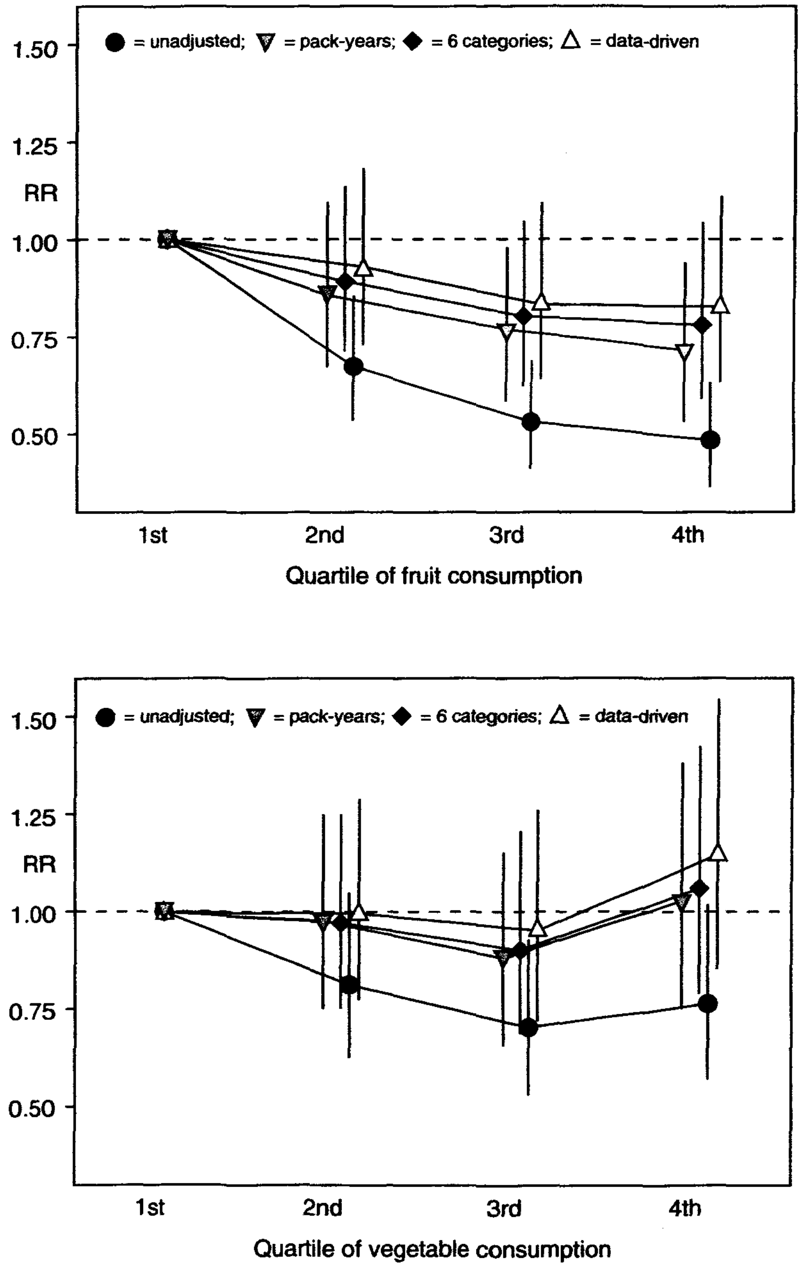

Figure 1
Relative risk of lung cancer from different adjustment procedures by quartile of fruit (above) and vegetable (below) consumption. Reference is the first quartile

present because smoking status might have changed during follow-up. Such bias could work in both directions (away or towards unity), depending on the covariance structure of the measurement errors in smoking and dietary data. Simulation studies based on realistic assumptions might shed some light on the possible magnitude of such a bias.

Reference

Whittemore, A.S. (1988) Effect of cigarette smoking in epidemiological studies of lung cancer. *Stat. Med.*, 7, 223–238

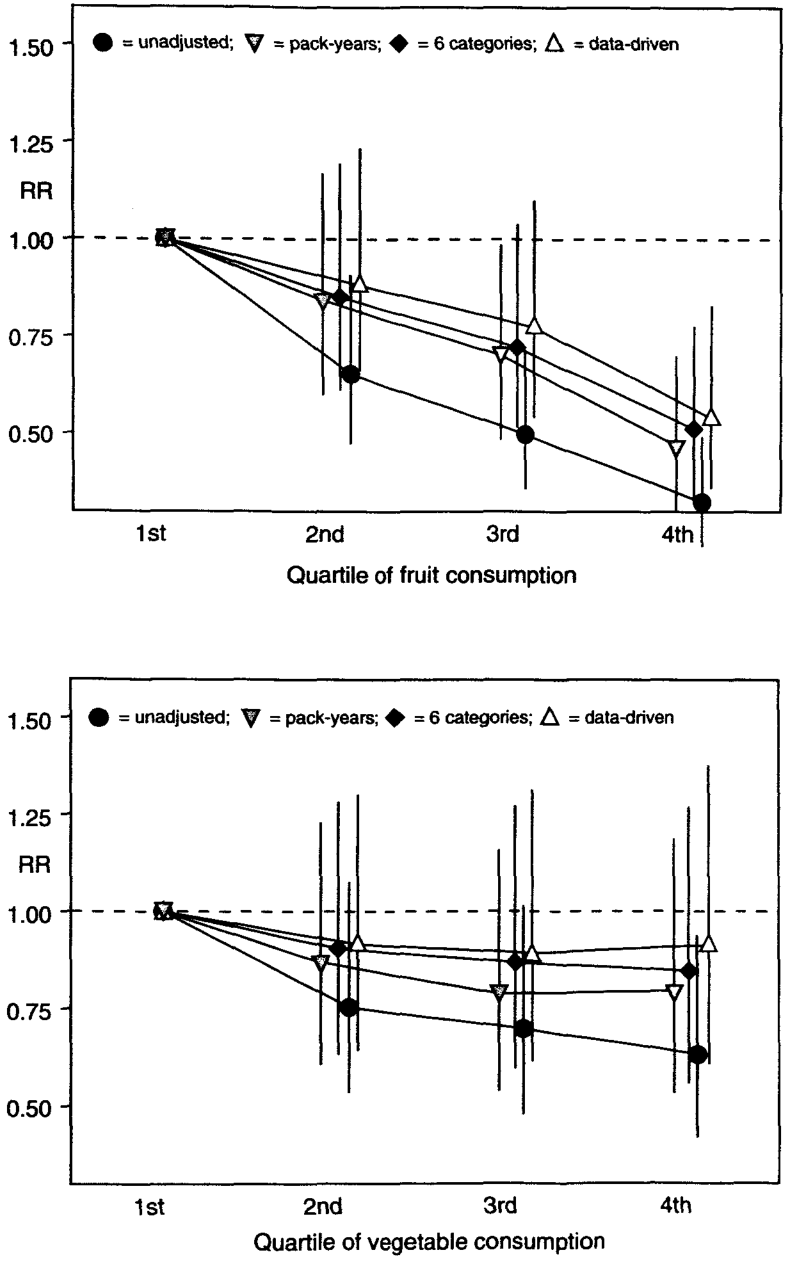

Figure 2
Relative risk of lung cancer from different adjustment procedures by quartile of fruit (above) and vegetable (below) consumption after excluding the first 2 years of follow-up from the analysis. Reference is the first quartile