Supplementary Methods

Regression dilution

Regression dilution is the phenomen in which regression coefficients are biased toward the null by errors or other random noise in the measurement of the independent variables. Regression calibration [1-3] is a method of correcting this bias using a sample in which measurements are made by both the noisy method (denoted Z) and a precise method (X). This allows us to estimate the proportion of the variance in the original measurements Z which is due to noise, and hence apply an appropriate correction factor $Var(Z)/Var(X)$. The technique can also correct for systematic biases due to linear calibration errors, and extends naturally to multivariate regression.

The method is not a panacea, since it makes a strong assumption: that after transformation to the scale used in the regression, the noise in the measurements is independent of the true values X. In addition, the correction factor is necessarily estimated with error, so confidence intervals will typically become wider in the process of correcting bias.

The situation in this study is not the prototypical scenario for regression calibration, but the method can be readily adapted. We assume that the true risk factor for lung cancer is life-long average tobacco use, for which life-long average cotinine level is a proxy. This life-average cotinine is denoted X_c. In the main study sample we have measurements of genotype X_g (assumed to be without error and $Z_g = X_g$) and of cotinine level on a single day, Z_c. Hence the noise in the cotinine level is due to daily fluctuations around the mean value. In an independent calibration sample, we have a series of three repeat measures, taken on different days, of the cotinine level. This allows us to estimate the variance $Var(X_c)$, by comparing the variance of the individual measurements with the within-individual averages $\overline{Z_c}$,

$$Var(X_c) = (3Var(\overline{Z_c}) - Var(Z_c))/2 \qquad (1)$$

In principle we would should also estimate the covariance of the day-to-day variation with the genotype. However in the present case, there was no genetic information available for the sample with repeat cotinine measures. Consequently we assumed they are uncorrelated. This seems the most plausible assumption, however one could imagine a genotypic predisposition for a more variable pattern of smoking: this would invalidate our approach.Under the above independence assumptions, the "corrected" logistic coefficient β_g and β_c for genotype and cotinine level, respectively, can be estimated as suggested by Rosner et al[1;2] through the correction matrix $\langle Z,Z\rangle\langle X,X\rangle^{-1}$ as following:

$$(\beta_g, \beta_c) = \left(\gamma_g - \frac{(1-r)bc}{rac - b^2} \times \gamma_c, \frac{ac - b^2}{rac - b^2} \times \gamma_c \right), \qquad (2)$$

where

$a = Var(X_g) = Var(Z_g)$,

$b = Cov(X_g, X_c) = Cov(Z_g, Z_c)$,

$c = Var(Z_c)$,

$r = Var(X_c) \big/ Var(Z_c)$.

γ_c - crude estimates for cotinine

γ_g - crude estimates for genotype.

Reference List

1. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. Am J Epidemiol 1992;136:1400-1413.
2. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol 1990;132:734-745.
3. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. Am J Clin Nutr 1997;65:1179S-1186S.