STATISTICAL METHODS FOR EXPOSURE RESPONSE ANALYSIS OF HIGHLY CORRELATED EXPOSURES

Lutzen Portengen, *Institute for Risk Assessment Sciences, Utrecht University, The Netherlands*
Roel Vermeulen, *Institute for Risk Assessment Sciences, Utrecht University, The Netherlands*

Background and Aims: Standard multiple regression techniques for exposure response analysis may fail for highly to moderately correlated exposure data due to the strong co-linearity and resulting instability of effect estimates. To identify single polychlorinated biphenyls (PCBs) congeners causally related to Non-Hodgkin Lymphoma in a nested case-control study from among a mixture of 36 different PCBs we performed a simulation study to assess the utility of statistical techniques that were developed to deal with highly dimensional data.
Methods: We simulated 150 different datasets based on the observed standardized multivariate lognormal distribution of 36 PCBs in the JANUS nested case/control study. Of these 50 simulations used the observed means and covariance matrix, 50 used a reduced covariance matrix (off-diagonal entries scaled by 1/3), and 50 used a diagonal covariance matrix to simulate uncorrelated exposures. For every dataset we simulated 6 different exposure-response scenario´s corresponding to either two or four randomly selected PCBs related to the outcome with a weak (OR=1.2), medium (OR=1.5), or strong (OR=2.0) effect. All datasets were analyzed using standard univariate and (stepwise) multiple regression analysis as well as Sparse Partial Least Squares Discriminant Analysis (SPLS-DA), an elastic net for Generalized Linear Models (GLMNET), and a Bayesian hierarchical variable selection model. Models were compared using measures of sensitivity, specificity, and relative bias.
Results: With uncorrelated exposures, univariate analysis outperformed all others in sensitivity and specificity, and had lowest bias. With moderate and strong correlations, naive multiple regression often failed to identify the true causal agents, while more appropriate techniques were often more sensitive and only slightly less specific. No method was uniformly best, but the elastic net appeared to offer a good compromise.
Conclusions: More sophisticated statistical techniques may be needed to identify causal agents in exposure response analysis of highly and moderately correlated exposure data. However, the lack of a formal inference framework for some of these methods may hamper more widespread use.