ELABORATION AND APPLICATION OF A NEW HIERARCHICAL CLASSIFICATION ALGORITHM IN EPIDEMIOLOGICAL RESEARCH

Ekaterina D. Konstantinova, *Institute of Industrial Ecology, Russia*
Anatole N. Varaksin, *Institute of Industrial Ecology, Russia*

Background and Aims: Most noninfectious diseases are caused by a potential influence of a range of risk factors (RF) (100 factors and more) on health. However, only a few those (3 to 5) have a significant influence. The identification of a main RF set is solved by means of multivariate statistical analysis. These methods work well in case of statistically independent RF and when their joint effect is amenable to a clear interpretation. The condition of RF independence is often violated in observation studies where it is impossible to vary experimental conditions.
Methods: In this report, new algorithm to look for main RF set is presented. The algorithm is based on a hierarchical classification method and allows:

- finding small RF sets that produce maximal influence on population health;
- creating decision rules for describing a "generalized image" of the sick and the healthy that would be clear to physicians;
- finding ways to compensate the negative influence of significant RF.

To carry out calculations, an epidemiological database gathered during 2002-2010 for Ekaterinburg preschool children (1300 children) is used. Basic diseases are determined at medical inspection for every child. Levels of atmospheric pollution are measured; data on 50 family-based RFs of health impairment are obtained using parent's questionnaire.
Results: We have identified sets of 3-4 RFs having a maximal negative influence on the prevalence of respiratory, cardiovascular, and musculoskeletal diseases and behavior disorders in children. These factors are: atmospheric pollution, gas-stove in apartment, parent smoking, low-quality drinking water, child's insufficient physical activity, mother's low educational attainment.
Conclusions: The suggested variant of stepwise hierarchical algorithm allows obtaining effective and evident decision rules for classifying children into low and high pathology prevalence categories using 3-4 RF. The algorithm allows one to find factors reducing the adverse impact of environmental contamination on population health.