

# Simulation and Latent Variable Regression Methods for Exposure-Response Analysis in Populations with Unidentified Sensitive Subgroups

Tiffany DeFoe, Mitchell Small, & Elizabeth Casman  
Carnegie Mellon University

Recent research has identified genetic risk factors for some environmental diseases, including cancers and immune system-mediated diseases; however, gene-environment interactions are not well understood for many diseases, and concerns about privacy and ethics have been limiting factors in the collection and use of genetic information in exposure-response modeling and risk assessment. This paper describes an approach to improve exposure-response modeling for environmental diseases with known genetic factors using currently available information on high-risk genotypes and their qualitative effects on risk. Example analyses are presented using epidemiological data on the risk of beryllium sensitization and chronic beryllium disease in a population of beryllium-exposed workers, including a subpopulation known to carry the high-risk genotype HLA-DPB1<sup>Glu69</sup>. The methods may be generalized for use in risk assessments where other factors may cause apparent or actual variation in exposure-response relationships across population subgroups, e.g., when an unidentified portion of the population is at increased risk of disease from sources or routes of exposure for which individual information is not available.

We use simulated epidemiological data to characterize the efficacy of traditional and latent class regression modeling approaches for populations which have subgroups with varying genetic risk factors, but for which individual genetic data are not available (Skrondal and Rabe-Hesketh, 2004). The two approaches are applied to published datasets on beryllium sensitization, including a cross-sectional study of beryllium ceramics workers (Henneberger et al., 2001) and a matched case-control study of beryllium-exposed nuclear weapons workers (Viet et al., 2000). The following example describes the simulation procedure using the reported exposure estimates for individuals in the ceramics workers dataset and assuming a simple genetic risk scenario in which some of the workers are genetically “hypersusceptible” and will become sensitized as a result of exposure to any amount of beryllium.

The exposure values published in the Henneberger et al. (2001) paper are used in the simulation, in order to explore the potential for model sensitivity to unidentified hypersusceptible (HS) individuals within this particular dataset. The two possible genetic types (HS, non-HS) are assigned randomly to the individuals in the dataset with frequencies based on information in the genetic susceptibility literature. We then define a risk function or functions for the non-HS population, such as a smooth function that specifies a probability of sensitization between 0 and 1 for any exposure value, or a distribution of exposure thresholds for non-HS individuals that represent the highest level of exposure each can tolerate without becoming sensitized. When sensitization status is fully determined by exposure level and risk function for both HS and non-HS individuals, we select a number of genotype assignments to generate for the (fixed) set of exposure values. When there is instead a random component to the risk of sensitization, conditioning on exposure and genotype, we additionally select the number of simulations to run for each unique assignment of genotypes to individuals. We also simulate across a range of genotype frequencies and risk functions.

We use datasets generated as in the example above to explore the efficacy of traditional and latent class regression modeling approaches for three types of genetic effects:

“hypersusceptibility”, in which a portion of the population will develop disease from any level of exposure; “effect modification”, in which the risk of disease for each group is described by smooth functions of exposure with identical intercepts; and “immunity”, in which a portion of the population will not develop disease at any level of exposure. Sets of simulated epidemiological data are generated and used to characterize (1) the potential for error in analyses that do not account for genetic subpopulations; (2) the degree of improvement in exposure-response models that may be achieved using traditional regression diagnostics to adjust and re-fit the model; and (3) the performance of latent class regression models designed to estimate the effects of genetic factors on disease risk, using prior estimates of the probability that each individual in the population belongs to each genetic class. Each approach is evaluated in terms of the likelihood that it will correctly detect an exposure-response relationship (or will not detect a nonexistent relationship) and in terms of its accuracy in predicting risk in a variety of scenarios. Using characteristics of the Henneberger et al. ceramics workers study, we found that the existence of a relatively small subpopulation of unidentified hypersusceptible individuals can substantially impair the ability of logistic regression models to detect a relationship between exposure level and risk of sensitization, and the resulting predictions of risk at specific exposure levels may be highly inaccurate. The sensitivity of our regression models to unidentified hypersusceptibles is due in part to the small number of beryllium-sensitized cases and a cluster of individuals with unusually low exposures in the dataset. For simulations in which hypersusceptible individuals were found to substantially affect model results, the use of outlier diagnostics to identify and remove selected influential data points improved the performance and prediction accuracy of the regression models; however, the results of these models must be interpreted with care. Latent class regressions modeling two classes in the population did not improve on traditional logistic regression models in this case, perhaps due to the small size of the dataset.

### *References*

Henneberger, P. K., Cumro, D., Deubner, D., Kent, M., McCawley, M., Kreiss, K; “Beryllium sensitization and disease among long-term and short-term workers in a beryllium ceramics plant”, *Int Arch Occup Environ Health* 74: 167-176, 2001.

Skrondal, A. and Rabe-Hesketh, S.; Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC, Boca Raton, 2004.

Viet, S. M., Torma-Krajewski, J. T., Rogers, J.; “Chronic Beryllium Disease and Beryllium Sensitization at Rocky Flats: A Case-Control Study”, *AIHAJ* 61: 244-254, 2000.