Using the Whole Cohort in the Analysis of Case-Control and Case-Cohort Data

Norman E. Breslow[1], Thomas Lumley[1,]
Christie M. Ballantyne[2], Lloyd E. Chambless[3], and Michal Kulich[4]

[1] Department of Biostatistics, University of Washington, Seattle, WA.
[2] Department of Medicine, Baylor College of Medicine, Houston, TX.
[3] Department of Biostatistics, University of North Carolina, Chapel Hill, NC
[4] Department of Probability and Mathematical Statistics, Charles University, Prague, CZ

Published analyses of data from case-control and case-cohort studies nested in large cohorts often ignore valuable information on cohort members not sampled as controls. The Atherosclerosis Risk in Communities (ARIC) investigators, for example, typically reported data only for the 10-15% of subjects sampled for sub-studies of their cohort of 15,972 participants. The remaining subjects contributed to stratified sampling weights, but not otherwise. Recently improved communication between biostatisticians and survey statisticians has led to better understanding of the two phase designs used in epidemiology and to the development and implementation of more efficient methods of data analysis. Variances of estimated regression coefficients in both parametric (logistic regression) and semi-parametric (Cox regression) models are the sum of two approximately independent terms: (1) the model based variance of coefficients that would be estimated were complete data available for all subjects in the main cohort (phase one sample); and (2) the design based variance of the Horwitz-Thompson estimate, using subjects sampled for the case-control or case-cohort study (phase two), of the sum of main cohort influence function contributions. Adjustment of standard sampling weights by regression calibration or estimation, now implemented in the R survey package, can sometimes dramatically lower the design based phase two variances. When applied with simulated case-cohort data from the National Wilms Tumor Study, where institutional pathology was a strong surrogate for central pathology, calibration and estimation reduced by 24% the standard error of the hazard ratio of relapse associated with "unfavorable histology". In re-analysis of data from an ARIC case-cohort study, no improvement was found for hazard ratios linking lipoprotein-associated phospholipase A_2 (Lp-PLA_2) with risk of coronary heart disease. The standard error of the interaction of Lp-PLA_2 with systolic blood pressure was reduced by 10%, however, and the precision of hazard ratios for covariates used for adjustment improved dramatically. Careful adjustment of sampling weights can protect against possible waste of valuation information.