On the analysis of case-control studies in cluster-correlated data settings

Haneuse and Rivera-Rodriguez

Epidemiology, Jan 2018

* Describes sampling and inference issues for case-control studies with two different sampling strategies – standard case control and case-control with sampling stratified by cluster
  + Situations where case-control study is conducted in a setting where the underlying population exhibits clustering
* As opposed to the standard setting for GEE which assumes you first sample clusters and then sample individuals from those clusters
* Motivated by HIV study in clinics in Malawi
* Analysis: inverse probability weights with GEE
  + Weights:
    - Standard case control
      * Cases: Total # cases in pop / # cases sampled
      * Controls: Total # controls in pop / # controls sampled
    - Case-control within each clinic
      * Cases: Total # cases in clinic k / # cases sampled from clinic k
      * Controls: Total # controls in clinic k / # controls sampled from clinic k
  + Working correlation structure: independent; exchangeable (more efficient)
  + Robust sandwich estimator for inference
* Sampling strategies they tried
  + 4 balanced designs
    - SRS
    - Random sample within each clinic
    - Standard case-control
    - Case-control within each clinic
      * For each, an average of 20, 40, 60, and 80 cases + controls per clinic
  + 2 unbalanced designs stratified by clinic size with 10, 16/20, 40/34 drawn from small, medium and large clinics
* Results
  + Case control with exchangeable correlation structure generally more efficient
  + Tradeoffs in efficiency across covariates
  + Standard case-control generally has more power and efficiency than case-control within clinics
  + If using case-control within clinics, unbalanced designs are more powerful and efficient
* \*Suggests that could use conditional logistic for the designs with sampling stratified by clinic, but this would change the interpretation and could not estimate effects of clinic-invariant characteristics
* \*The authors are working on estimation/inference for generalized linear mixed model for case-control and case-control with cluster-stratified case-control

A new estimation approach for combining epidemiological data from multiple sources

Huang et al.

Journal of the American Statistical Association, March 2014

* Propose two step procedure to combine data from diverse sources, with the aim to fit a parametric model to quantify risk factors affecting probability individual develops disease
  + First step: derive all possible estimating functions based on data from single source or two different sources
  + Second step: combine all available estimating functions efficiently, related to generalized method of moments used in microecon models and longitudinal surveys
  + Avoids use of MCMC and does not assume conditional independence or require precise knowledge on correlation among covariates
* Data sources used:
  + Case control study of pancreatic cancer from incident cases in CT Jan 2005-Aug 2009 and frequency matched controls by gender and age group found through RDD over same time frame
  + Connecticut Tumor Registry Data – only a minority of pancreatic cancer cases in CTR undergo rigorous research study-level validation of primary site (extra cases)
  + Behavioral Risk Factor Surveillance System Data: state based system of health surveys conducted using RDD, uses post-survey weighting techniques to maximize representativeness (extra controls)
    1. Residential locations for these subjects derived based on zip code only
  + Traffic data for CT from dept of transportation
* Rationale: motor vehicle emissions a major source of air pollution but majority of existing studies had small number of incident cancers compared to number existing in registry, want to supplement current case control study with registry and survey data
* \*\*\*Estimating Equation Stuff\*\*\*
  + Develop equations for original case control study where N = cases, M = controls
  + Adapt this to multiple data sources where N\_1 = original cases, M\_1 = original controls, N\_2 = additional CTR data, M\_2 = BRFSS data
  + Two problems: selection bias where N\_2 not a SRS from N\_c = N\_1 union N\_2, adapting methodology from original CC estimating equations to obtain unbiased estimating functions with potentially incomplete covariate data
* Simulation Study
  + Results: generated 1000 realizations N\_1, N\_2, M\_1, M\_2 and applied method to estimate beta, found reduction in SE when any additional data beyond N\_1, M\_1 included
* Discussion
  + Proposed method can yield estimators with smaller variances vs if only original case control data used and can enhance study power
  + Real data analysis revealed that parameter estimates can vary with data sources that are included
  + Authors think it is reasonable to believe that BRFSS data (survey data) can produce less biased results vs the controls collected in the case control study because of more sophisticated sampling designs and statistical tools used to mitigate bias for BRFSS data