Dr Bronwyn Loong
The Australian National University (ANU)

Disclosure Control using Partially Synthetic Data for Large-Scale Health Surveys, with Applications to CanCORS

Joint work with: Alan M. Zaslavsky, Yulei He and David P. Harrington

Statistical agencies have begun to partially synthesize public-use data for major surveys to protect the confidentiality of respondents' identities and sensitive attributes, by replacing high disclosue risk and sensitive variables with multiple imputations. To date, there are few applications of synthetic data techniques to large-scale healthcare survey data. In this talk, we describe partial synthesis of survey data collected by CanCORS, a comprehensive observational study of the experiences, treatments, and outcomes of patients with lung or colorectal cancer in the United States. We discuss the key steps of selecting variables for synthesis, specification of imputation models and measurements of data utility and disclosure risk. We evaluate data utility by replicating published analyses and comparing results using original and synthetic data, and discuss practical issues in preserving inferential conclusions. We found that important subgroup relationships must be included in the synthetic data imputation model, to preserve the data utility of the observed data for a given analysis procedure. We conclude that synthetic CanCORS data are suited best for preliminary data analyses purposes.