

Title: Relating clusters of patients to distal outcomes: a misclassification-correction approach

Yajing Zhu, Steven Kiddle

MRC Biostatistics Unit, University of Cambridge

Latent class analysis (LCA) is widely used to derive categorical variables from multivariate data. One particular interest is to relate these classes to distal outcomes. Commonly used approaches include (1) the one-step approach, where the mixture model and the regression model are estimated simultaneously and (2) naïve step-wise approach (e.g. the modal class approach) where subjects are first assigned to the latent class with the highest posterior probability and then treated as observed predictor in the regression model. However, regression coefficients for the modal class will be biased due to the unintended influence of the distal outcome on class membership (the one-step approach) and the misclassification error (the modal class approach). To address these problems, we treat the derived classes as an imperfect measurement of the true class in the regression for the distal outcome, with measurement error determined by the misclassification probabilities (Vermunt, 2010). This misclassification-correction approach is applied to a study using the UK electronic health records to understand the impact of multimorbidity (co-existence of more than one long term conditions) on service use and mortality.

Keywords: mixture model, 3-step approach, misclassification, cluster analysis