LATENT CLASS REGRESSION ANALYSIS OF COLORECTAL CANCER DATA

W.J. Harrison[1], A. Downing[1], M.S. Gilthorpe[1], D. Forman[1,2], R.M. West[1]

[1]*Centre for Epidemiology & Biostatistics, University of Leeds, UK;* [2]*Northern and Yorkshire Cancer Registry and Information Service, UK.*

Within cancer epidemiology, patients are not assigned randomly to diagnostic centres (hospitals). The *stage* of disease at diagnosis also affects prognosis. Survival and the impact of associated risk factors may vary according to both place of diagnosis (*hospital*) and *stage* at diagnosis. To address these issues we use latent class analysis (LCA). We use a dataset of patients in a large UK regional population diagnosed as having colorectal cancer. We choose to model the outcome as two-year survival (alive/ dead) for ease of comparison with other studies and reports.

A problem arises from the hierarchical nature of such datasets, since patients are clustered by diagnostic centres that are not strictly a random sample. Consequently, standard multilevel modelling may be inappropriate. Another issue is that standard regression models give rise to biased results when model covariates are measured with error or have missing values, and this bias is exacerbated within product interaction terms [1]. Since *stage* is measured with uncertainty and suffers a large degree of missingness, considering it as a covariate and exploring it in interactions with risk factors has the potential to introduce large bias. Bias may also arise where covariates (such as *stage*) lie on the causal path, due to the reversal paradox [2].

By employing LCA at the hospital level we effectively generate a 'semi-parametric' multilevel model. Furthermore, to minimise bias from *stage,* its interactions and the reversal paradox, we use *stage* as a class predictor. The resultant patient classes will have a graduated survival analogous to that observed for different stages of disease. Patient-level latent classes thus become a surrogate for the categories of *stage*. The relationship between mortality and various risk factors was explored (*area deprivation*, *sex*, *age at diagnosis*) across patient classes, introducing an 'interaction' without the risk of exacerbated bias. We may then examine how survival (graduated across patient classes) varies across hospital classes.

Patients diagnosed with colorectal cancer between 1991 and 2004 were identified and LCA models were explored using around 2 patient-classes and 3-hospital classes, comparing log-likelihood statistics (Bayesian Information Criterion) and misclassification rates. *Stage* at diagnosis (38.3% missing in our data, with missing values categorised) was included as a patient-class predictor. The empirically determined latent group structure was informative: the model suggests that there is a typology of patients and a typology of hospitals. By allocation of patients to hospital types that match their profile there might be opportunity to optimise patient care. This analytical strategy has considerable prognostic utility to inform health service providers of disparities within patient care.

References

[1] Greenwood DC, Gilthorpe MS, Cade JE. The impact of imprecisely measured covariates on estimating gene-environment interactions. BMC Med Res Methodol 2006; 6:21.

[2] Tu YK, West R, Ellison GT, Gilthorpe MS. Why evidence for the fetal origins of adult disease might be a statistical artifact: the "reversal paradox" for the relation between birth weight and blood pressure in later life. Am J Epidemiol 2005; 161(1):27-32.