International Journal of Population Data Science

Journal Website: www.ijpds.org

Weighted multiple imputation of ethnicity data that are missing not at random in primary care databases

Pham, Tra My^{1*}, Petersen, Irene¹, Carpenter, James², and Morris, Tim¹

¹University College London

²London School of Hygiene & Tropical Medicine

Background

Ethnicity is an important factor to be considered in health research because of its association with inequality in disease prevalence and the utilisation of healthcare. Ethnicity recording has been incorporated in primary care electronic health records, and hence is available in large UK primary care databases such as The Health Improvement Network (THIN). However, since primary care data are routinely collected for clinical purposes, a large amount of data that are relevant for research including ethnicity is often missing. A popular approach for missing data is multiple imputation (MI). However, the conventional MI method assuming data are missing at random does not give plausible estimates of the ethnicity distribution in THIN compared to the general UK population. This might be due to the fact that ethnicity data in primary care are likely to be missing not at random.

Objectives

I propose a new MI method, termed 'weighted multiple imputation', to deal with data that are missing not at random in categorical variables.

Methods

Weighted MI combines MI and probability weights which are calculated using external data sources. Census summary statistics for ethnicity can be used to form weights in weighted MI such that the correct marginal ethnic breakdown is recovered in THIN. I conducted a simulation study to examine weighted MI when ethnicity data are missing not at random. In this simulation study which resembled a THIN dataset, ethnicity was an independent variable in a survival model alongside other covariates. Weighted MI was compared to the conventional MI and

other traditional missing data methods including complete case analysis and single imputation.

Results

While a small bias was still present in ethnicity coefficient estimates under weighted MI, it was less severe compared to MI assuming missing at random. Complete case analysis and single imputation were inadequate to handle data that are missing not at random in ethnicity.

Conclusions

Although not a total cure, weighted MI represents a pragmatic approach that has potential applications not only in ethnicity but also in other incomplete categorical health indicators in electronic health records.

*Corresponding Author:

Email Address: tra.pham.09@ucl.ac.uk (T. P. Pham)