International Journal of Population Data Science

Journal Website: www.ijpds.org

How do socio-demographic differences in administrative records affect the quality (accuracy) of data linkage?

Randall, S¹, Ferrante, A¹, Boyd, J¹, and Brown, A¹

¹Curtin University

Introduction

Record linkage is inherently uncertain, with all linkages containing some amount of false positive and false negative errors. Previous results have suggested that linkage error may not be evenly distributed throughout the population, with particular subgroups exhibiting higher rates of linkage error. Record linkage is inherently uncertain, with all linkages containing

required into investigating how linkage errors effect research outcomes.

Objectives and Approach

This study investigated the distribution of linkage error using four large-scale Australian administrative datasets; hospital admissions datasets from Western Australia and New South Wales, and emergency presentation datasets from New South Wales and South Australia. Each dataset had been previously de-duplicated to a very high standard, with large scale manual review taking place; these results were used as our truth set.

Each dataset was linked using probabilistic record linkage with results (precision and recall) compared by gender, age, geographic indices of remoteness and socioeconomic status.

Results

Results were highly dataset dependent. Consistent findings were lower linkage quality found for individuals living in remote locations, and lower linkage quality in those in the youngest category (those born after 1980). Some datasets showed lower linkage quality for females, for those in middle age as compared to the elderly, and for those with lower socioeconomic status. The differences in linkage quality found were typically small. Changes in threshold settings had generally no effect on the relationship between sociodemographic characteristics and linkage quality.

Conclusion/Implications

Linkage studies focussing on younger individuals and those in remote areas may have greater uncertainty regarding their results. Targeting efforts by linkage units may be required to ensure even distribution of linkage errors. Further research is