International Journal of Population Data Science

Journal Website: www.ijpds.org

When the Census Comes Marching In: Challenges and Successes in Linking Individual-Level Census Records to the Utah Population Database

Smith, K¹ and Fraser, A²

¹University of Utah

²Huntsman Cancer Institute, University of Utah

Introduction

The availability of historic, individual-level census records in the United States has grown in recent years. With access to identifiers, it is possible to link these records to existing databases. The performance of and strategy for these linking efforts is not well characterized.

Conclusion/Implications

This study found that linking census records to an existing population registry is feasible and with relative success. Using household/genealogy structure of the census is useful when linking to the genealogies in the UPDB. These links allow studies of effects of early life conditions on later life outcomes.

Objectives and Approach

The Utah Population Database (UPDB), launched in 1975, is a population registry comprising comprehensive data from genealogies, medical/vital records, and numerous administrative and demographic records spanning the past two centuries. UPDB initially did not hold individual-level US Census records until now. UPDB has massive volumes of identifiers that we have cleaned and it therefore represents a "gold standard" representation of Utah's population. The objective here is to describe the methods used and the record linking performance applied to census records that we have linked to the UPDB for persons appearing in the 1880, 1900, 1910, 1920, 1930 and 1940 censuses.

Results

We collaborated with FamilyTree, Ancestry, and IPUMS (University of Minnesota) for keying and preparing data from the 1880-1940 censuses. We then linked these records to the UPDB using probabilistic record linking methods and manual review. Linking rates by census year varied by the quality of records and electronic data capture and by specific Census fields for a given census. Data quality was somewhat lower for the 1910 and 1940 censuses and hence they had lower linking rates (66.9% and 70.4, respectively). Household heads enjoyed higher linking rates (72% was the lowest, in 1940). We used household heads to help guide links to offspring and spouses whose linking rates exceeded 75% in general. Non-family members and single men linked at much lower rates (<50%).