International Journal of Population Data Science

Journal Website: www.ijpds.org

Harmonization of data from cohort studies– potential challenges and opportunities

Dahal, KA[1], Patten, S[1], Williamson, T[1], Patel, A[2], Premji, S[1], Tough, S[3], Letourneau, N[1], and Giesbrecht, G[1]

[1]University of Calgary

[2]Alberta Health Services

[3]Paediatrics, Cumming School of Medicine, University of Calgary

Introduction

Pooling data from cohort studies can be used to increase sample size. However, individual datasets may contain variables that measure the same construct differently, posing challenges in the usefulness of combined datasets. Variable harmonization (an effort that provides comparable view of data from different studies) may address this issue.

Objectives and Approach

This study harmonized existing datasets from two prospective pregnancy cohort studies in Alberta Canada (All Our Families (n=3,351) and Alberta Pregnancy Outcome and Nutrition (n=2,187)). Given the comparability of the characteristics of the two cohorts and similarities of the core data elements of interest, data harmonization was justifiable. Data harmonization was performed considering multiple factors, such as complete or partial variable matching regarding question asked/responded, the response coded (value level, value definition, data type), the frequency of measurement, the pregnancy time-period of measurement, and missing values. Multiple imputation was used to address missing data resulting from the data harmonization process.

Results

Several variables such as ethnicity, income, parity, gestational age, anxiety, and depression were harmonized using different procedures. If the question asked/answered and the response recorded was the same in both datasets, no variable manipulation was done. If the response recorded was different, the response was re-categorized/re-organized to optimize comparability of data from both datasets. Missing values were created for each resulting unmatched variables and were replaced using multiple imputation if the same construct was measured in both datasets but using different ways/scales. A scale that was used in both datasets was identified as a reference standard. If the variables were measured in multiple times and/or different time-periods, variables were synchronized using pregnancy trimesters data. Finally, harmonized datasets were then combined/pooled into a single dataset (n=5,588).

Conclusion/Implications

Variable harmonization is an important aspect of conducting research using multiple datasets. It provides an opportunity to increase study power through maximizing sample size, permitting more sophisticated statistical analyses, and to answer novel research questions that could not be addressed using a single study.

http://dx.doi.org/10.23889/ijpds.v3i4.868