A Methodology Template for Constructing Variables and Applying Survey Weights in Complex Survey Design Studies using R

Andrew Jergel, MPH[1]; Scott Gillespie, MS, MPH[1]; Lilian Zapata, MD[2]; Kiesha Fraser Doh, MD[3,4]

[1]Pediatric Biostatistics Core, Department of Pediatrics, Emory University, Atlanta, Georgia, USA.

[2]Emory University School of Medicine, Atlanta, Ga USA

[3] Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, USA.

[4]Children's Healthcare of Atlanta, Atlanta, GA, USA.

Background:

Using The Future of Families and Child Wellbeing Study (FFCWS) database, we aimed to analyze the relationships between Healthcare Utilization (HU), Adverse Childhood Experience (ACE) scores, and Gun Violence Exposure (GVE) in adolescents. The FFCWS follows 4,898 teens and their families and currently spans 6 different years in the focal child's life (Birth-Years15). Each time-point may contain items from parents, caregivers, teachers, focal child, and other sources (e.g., GVE data); moreover, the FFCWS contains national- and city-level survey weights with replicate weights available by year and survey. With complex survey design (CSD) considerations and thousands of variables available, CSD datasets like FFCWS can quickly become daunting and costly to analyze. Sources for proper variable selection/creation, applying the appropriate survey weights, and methods to analyze the final data are few–especially in more niche topics. Outlined here, we hope to provide a R methodology template for using data from large, CSD databases.

Method:

Variables were selected based on their importance to our research topic. We used data from nine surveys from Birth-to-Year15 to construct our variables for HU, ACE, GVE, and Demographics. Some variables were created from data across multiple years/surveys. We only used the focal child's survey weights, as they are our population of interest. Lastly, observations that did not fit the inclusion criteria were removed, in the analytical phase, to ensure the retention of CSD information. This study used the R packages *survey* and *gtsummary*.

Results:

Our final dataset contained variables for ACE scores, HU, GVE, and the focal child's survey weights associated with the least missingness (e.g., removed participants with no GVE data). The use of replicate weights improves the estimation of 95% confidence intervals (CI) and p-values. Results were tabulated using tbl_svysummary() from *gtsummary*, as the function calculates survey-weighted, streamlined tables with customizable summary and inferential statistics. Replicate weight incorporation is not available with tbl_svysummary(); therefore, other functions like svyby() from *survey* may be required.

Discussion:

The methods employed help guide researchers on important statistical and analytical approaches and answer questions like "What survey weights do I use?" that seem simple but can quickly increase analysis time and costs.