Standardizing November 30th (Nov30) Demographic Data with Merged Dataset (MD)

- Resolved discrepancies:
    - PI transcribed to NH in accordance with MD
    - NA transcribed to AE
    - HA is listed as a race in Nov30; classified as ethnicity in MD
    - NH means not Hispanic or Latino in Nov30
    - Nov30 does not have San Francisco; MD does
    - 'Brooklyn' in MD changed to 'New York'
    - 'La Jolla' in MD changed to 'San Diego'
    - 'USC/MASS' omitted from MD
- Added abstracted diagnoses 2 and 3
- Resulting file: bgc_cDiag123_cities_std.csv

Checking SCH_DEP against GROUP from COGS2 dataset

- Main finding: 10 inconsistent points
    - Group suggests SZ, but SCH_DEP has NA
    - Resulting file:
        - cogs2_schdep_inconsistencies.csv
- SCH_DEP encodings reflected in new csv files:
    - bgc_raw.csv
    - bgc_only.csv

Race Hypothesis Testing

- Cities examined: Baltimore and Los Angeles
    - In Baltimore, conditions for Chi-square were not met, so resampling was used
        - Too low expected values for some races ($n < 5$)
        - Result of resampling: significant difference in AA and CA amounts (95% confidence)
    - Chi-square and resampling found significant differences between LA subject demographics vs city demographics for all races
        - (95% confidence)
- Chi-square methodology:
    - Threshold for significance: $p < 0.05$
    - 2010 and 2020 census data were averaged for each race category
    - Averages converted into proportions
    - Proportions applied to sample size of merged dataset (after cleaning) to generate expected values for race
- Resampling methodology:
    - Threshold for significance: $p < 0.05$
        - Bonferroni correction applied
            - $0.05 / 7 = 0.007$ for Baltimore
            - $0.05 / 6 = 0.008$ for Los Angeles
    - Resampled the same number of observations in cleaned dataset from a distribution made from the 2010 and 2020 census average counts; 1000 resamples were taken

    - Tallied the races in each resample
    - Ordered the counts within each race from smallest to largest
    - Constructed a confidence interval out of the lower and upper percentiles as specified by the corrected p-values
        - For Baltimore: 0.3% and 99.7%
        - For Los Angeles: 0.4% and 99.6%
    - If the observed value was contained within the confidence interval, the observed value did not significantly differ from the expected value
- Resulting files:
    - balt_race_resam_hypothesis_test.csv
    - LA_race_resam_hypothesis_test.csv
    - LA_race_chix.jpeg

Relative Risk (RR) for cDiagnosis3 by Race and Gender

- City examined: Los Angeles
- Methodology for finding relative risk:
    - Lowest non-zero incidence rate within each condition was used as reference group
- General findings for race:
    - Caucasians have highest RR for MDD
    - African-Americans have highest RR for SZSAFD, with American Indians a close second
    - Multiracial individuals have highest RR for BAD1 and BAD2
- General findings for gender
    - Females three times more likely as males to have MDD
    - Males twice as likely as females to have SZSAFD
    - Weak female bias for BAD1 and BAD2
- Resulting files:
    - rr_racescdiag3_LA.csv
    - rr_genderscdiag3_LA.csv