**Background**

Due to the relatively low prevalence of schizophrenia, large, multi-center studies are increasingly necessary to advance neuroscience and evaluate novel interventions. However, such studies still experience issues with generalizability. Of the factors that create these issues, race, gender, and ethnicity disparities in schizophrenia research are understudied. Here, we provide a way of quantifying the disparity within a sample and describe what additional sampling is required to achieve racial, gender, and ethnic parity.

**Methods**

We examined race, gender, and ethnicity information for 841 healthy control subjects (HCS) and 1060 subjects with schizophrenia (SZ) between ages 18 and 65 from Los Angeles, New York, Philadelphia, and Seattle within the Consortium on the Genetics of Schizophrenia 2 (COGS2) study. To characterize the diversity within our sample, we drew from the Diversity Index methodology used by the National Equity Atlas and the entropy index calculation described in Massey & Denton (1988). We created 24 categories based on race (White/Caucasian, African-American, American Indian/Alaska Native, Asian, Pacific Islander/Native Hawaiian, and Mixed/Other), ethnicity (Hispanic/Latino or not Hispanic/Latino), and gender (Male/Female). We used the proportions in these categories to generate a Diversity Index by diagnosis for each COGS2 city. These proportions and diversity indices were compared to those generated from American Community Survey (ACS) census 1% sample data on the respective COGS2 cities from 2010-2014, limited to individuals between ages 18 and 65. Data for San Diego was not accessible through ACS, so subjects from San Diego were omitted from analysis.

In addition to diversity characterization, we conducted simulations to more closely examine the demographic closeness between the COGS2 healthy control/schizophrenia samples and the pooled background city population. The pooled background city population was created by combining ACS data, weighting the data from each city by the observed proportions of each city within COGS2. For example, if Los Angeles provided 25% of control subjects in COGS2, then the person-weight (PERWT) variable from the 1% ACS sample data in Los Angeles would be rescaled by 0.25. Our sampling algorithm involves the following steps: 1) Resample a pre-specified percent of original COGS2 sample, 2) Reject the subjects who are oversampled as based on the background city proportions, 3) Add the remaining subjects to the COGS2 sample, 4) Re-calculate the 24-factor demographic proportions, and 5) Stop algorithm if all the proportions are within a pre-specified absolute difference with the proportions from the background city population. The sampling assumption behind our algorithm is that the original COGS sample represents the population of people that enters schizophrenia research studies, which may differ from the city population. Categories that were not present in the original COGS sample but were preset in the ACS sample were omitted. For our simulations, each resample was 50% of the original COGS2 sample size per diagnosis, and the absolute difference threshold was 0.025.

**Results**

[?] Should I do hypothesis testing for the over/under sampled? Some of the differences are very minute.

Using the above sampling parameters for HCS, 95% of the number of resamples for convergence to ACS proportions fell between 12 and 37. For SZ, 95% of the number of resamples for convergence to ACS proportions fell between 32 and 71. 95% of the number of resamples for convergence from SZ to HCS proportions fell between 8 and 29.

Increasing the rescaling factor past 50% of the original sample size had diminishing returns on achieving sooner convergence in both HCS and SZ. Increasing the absolute difference threshold led to sooner convergence in both groups. At an absolute difference of 0.05, HCS achieved immediate convergence to ACS. The SZ group did not achieve immediate convergence even at an absolute difference of 0.10.

**Conclusion**

Our analysis showed that for COGS2, the SZ sample was more divergent from background city proportions than was the HCS sample, and that the SZ sample was more similar to the HCS sample than the background city proportions. Furthermore, we showed that resampling above 50% of the original sample size had marginal benefits to achieving sooner convergence and that increasing the absolute difference threshold decreases the number of resamples needed for convergence.

These findings suggest that schizophrenia samples are more susceptible to divergence from city demographics than are control samples, calling for a more careful sampling approach than broad advertisement and enrollment of subjects. In addition, our findings suggest that there may be reduced benefits to resampling more than 50% of the original sample.

Limitations of our analysis include assuming that the original COGS2 sample accurately reflects the population of participants and assuming that age-matching between ACS and COGS2 was achieved by limiting the ACS population to the same age range as found in COGS2.