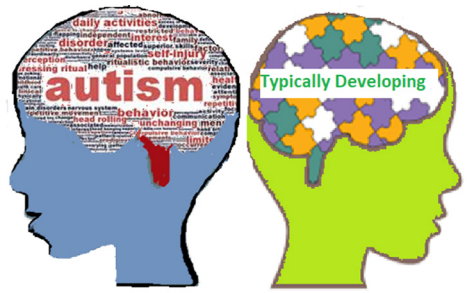
Iterative Matching Algorithms for Observational Data with Small Group Sizes: Application to Autism Spectrum Disorder Study

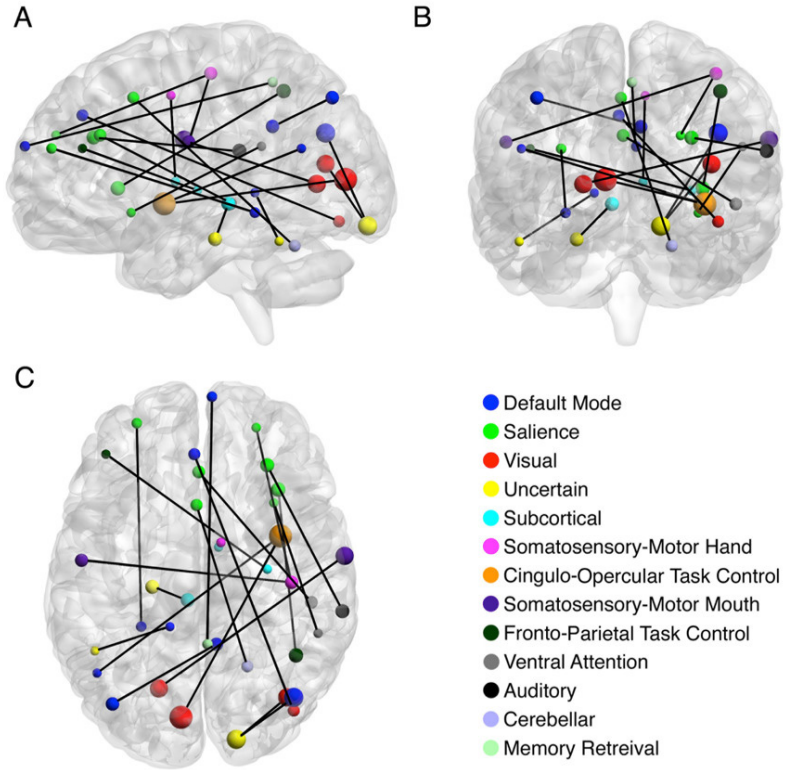
Despite consensus on the neurobiological nature of autism spectrum disorder (ASD), brain biomarkers remain unknown. To pinpoint atypical patterns in brain imaging data, one must first eliminate the undesirable effect of certain background variables before applying any pattern recognition. Matching is a nonparametric method of controlling the influence of confounding variables in order to obtain unbiased inferences in observational studies. Although many matching algorithms can achieve this goal, almost all currently existing methods assume that a large number of control subjects (termed as a control reservoir) are available so that each subject in the treatment or disease group can be matched reasonably well with at least one control subject. Unfortunately we have only a limited number of ASD and TD (typically developing) subjects and the sample sizes in the two groups are rather similar. To this end, we propose two iterative matching algorithms based on propensity scores and proximity matrices from random forest. The performance of these algorithms was evaluated using data from the Brain Imaging Development Lab (BDIL) at SDSU. Results show that these methods can produce ASD and TD samples that are much more balanced in terms of background covariates while retaining most of the subjects in the final matched samples.

Afrooz Jahedi, *Juanjuan Fan, and Ralph-Axel Müller*

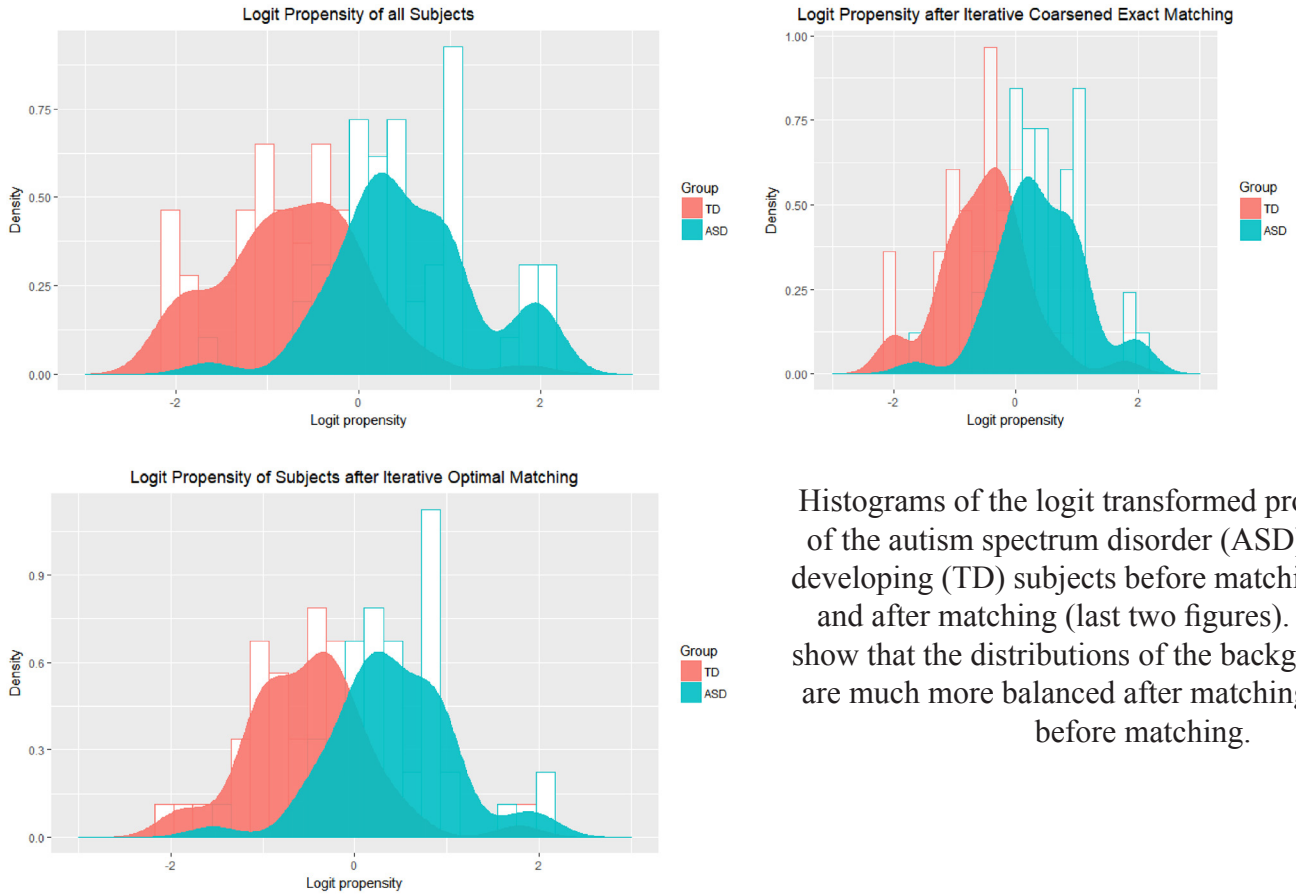
This research is supported by the National Institutes of Health (R01-MH081023; R01-MH101173) grants, and the Computational Science Research Center (CSRC) at San Diego State University



Despite its high prevalence (1 in 45 U.S children), brain biomarkers for ASD remain unknown.



Twenty most informative connections for predicting ASD in (A) sagittal, (B) coronal, and (C) axial views. The size of each node reflects the magnitude of conditional variable importance. Neural networks in the legend are sorted in descending order by frequency of occurrence. The majority of ROIs were selected from the default mode, salience, and visual networks. ASD, autism spectrum disorder; ROI, regions of interest (Jahedi et al., 2017).



Histograms of the logit transformed propensity scores of the autism spectrum disorder (ASD) and typically developing (TD) subjects before matching (first figure) and after matching (last two figures). These figures show that the distributions of the background variables are much more balanced after matching, compared to before matching.