ADHD diagnosis from multiple data sources with batch effects.

Olivetti E, Greiner S, Avesani P.

NeuroInformatics Laboratory, Bruno Kessler Foundation Trento, Italy ; Center for Mind and Brain Sciences, University of Trento Trento, Italy.

Front Syst Neurosci. 2012;6:70. doi: 10.3389/fnsys.2012.00070.
http://www.frontiersin.org/Journal/DownloadFile.ashx?pdf=1&FileId=55434&articleId=31344&Version=1&ContentTypeId=21&FileName=fnsys-06-00070.pdf

The Attention Deficit Hyperactivity Disorder (ADHD) affects the school-age population and has large social costs. The scientific community is still lacking a pathophysiological model of the disorder and there are no objective biomarkers to support the diagnosis. In 2011 the ADHD-200 Consortium provided a rich, heterogeneous neuroimaging dataset aimed at studying neural correlates of ADHD and to promote the development of systems for automated diagnosis. Concurrently a competition was set up with the goal of addressing the wide range of different types of data for the accurate prediction of the presence of ADHD. Phenotypic information, structural magnetic resonance imaging (MRI) scans and resting state fMRI recordings were provided for nearly 1000 typical and non-typical young individuals. Data were collected by eight different research centers in the consortium. This work is not concerned with the main task of the contest, i.e., achieving a high prediction accuracy on the competition dataset, but we rather address the proper handling of such a heterogeneous dataset when performing classification-based analysis. Our interest lies in the clustered structure of the data causing the so-called batch effects which have strong impact when assessing the performance of classifiers built on the ADHD-200 dataset. We propose a method to eliminate the biases introduced by such batch effects. Its application on the ADHD-200 dataset generates such a significant drop in prediction accuracy that most of the conclusions from a standard analysis had to be revised. In addition we propose to adopt the dissimilarity representation to set up effective representation spaces for the heterogeneous ADHD-200 dataset. Moreover we propose to evaluate the quality of predictions through a recently proposed test of independence in order to cope with the unbalancedness of the dataset.