The Phenotype Differences Model:
Identifying Genetic Effects with Incomplete Sibling Data

The identification of causal relationships between specific genes and social, behavioral, and health outcomes is challenging due to confounding factors such as population structure and dynastic genetic effects. Sibling pairs present a useful natural experiment for the identification of causal genetic effects because, conditional on their parents' genes, a child's genes are inherited randomly via recombination. Thus, genetic differences between siblings are ignorably assigned. At present, the *fixed effects* model is the predominant regression specification used to estimate genetic effects using sibling pairs. Such models require four pieces of information: the genotype of both siblings and the phenotype of both siblings. We introduce a new regression specification to compare siblings and estimate direct genetic effects, which we call the *phenotype differences* model. Phenotype differences models require only three pieces of information: the genotype of one sibling and the phenotype of both siblings. We show that, under minor assumptions, the phenotype differences model, like the fixed effects model, provides unbiased and consistent estimates of genetic effects. We also derive the comparative efficiency of phenotype differences and fixed effects in large samples. Phenotype differences provides less precise estimates than does fixed effects, however, the smaller amount of genetic data required by phenotype differences compared to fixed effects will increase sample sizes available for within family genetic analyses, thereby improving the precision and robustness of genetic discoveries. Using sibling pairs in the Wisconsin Longitudinal Study, we implement phenotype differences and show its utility for genome-wide association studies, polygenic score analyses, gene-environment interaction studies, and applications of Mendelian Randomization.