Dr. Thomas DiPrete

Columbia University

"Genetic Instrumental Variable (GIV) Regression: Explaining socioeconomic and health outcomes in non-experimental data"

How does genetics influence socioeconomic and health outcomes?

Social scientists have been investigating the major determinants of socioeconomic and health outcomes prior to the emergence of genetic information. With cutting edge research, genetic information becomes more accessible to scholars; thus, using these factors in modeling common social concerns becomes increasingly widespread. Being able to pinpoint the genetic and environmental causes of life outcomes gives rise to more accurate understandings of social phenomenon. In order to maintain the rigors of methodologies alongside the increasing availability of genetic information, we must examine two challenges.

First, how do we measure accurate estimates of these genetic influences? Second, how do we mitigate the effects of exposures to outcomes from the bias of genetic correlation? These issues present a barrier in analyses because traits are genetically complex with several single neocleotide polymorphisms (SNPs) accounting for any one trait. These relevant SNPs are summarized in a polygenic risk score (PGS), a number based on variation in multiple genetic loci and their associated weights. We simply have little understanding of how genetic markers are linked to the outcomes themselves.

How does attenuation bias in PRS work in studies that incorporate genetic information?

Attenuation bias measures the degree to which the correlation between the independent variable (PGS) and the dependent variable (e.g., educational attainment) is underestimated toward zero. In this case, attenuation bias increases as the sample size for genome-wide association studies (GWAS) discovery decreases. For example, the R-squared value for educational attainment increases from 5% when the sample size for GWAS discovery is 300,000 to 10% when the sample size is 1,000,000

With increases in the number of potentially relevant genetic markers in GWAS, scholars use all available information to produce reliable estimates. With all these advancements, now we could more accurately predict educational attainment with a cheek swab than parental education or income.

Nevertheless, there is great difficulty in obtaining a true PGS because genetic aspects may be correlated with cultural factors like group membership, for instance. As a result, we may draw incorrect and problematic conclusions like identifying a correlation between being East Asian and using chopsticks.

How do we correct attenuation bias?

The current practice is to use all the GWAS sample to estimate the PGS as reliably as possible. An insight of Dr. DiPrete and his colleagues' research is instead to use a measurement error correction approach: splitting the training sample into two independent subsamples, obtain a PGS from both subsamples, and use one as the instrumental variable (IV) and the other as the genetic instrument variable (GIV). This method efficiently obtains coefficients that are consistent with the conventional method

What are the challenges of using genetic instrumental variable regression?

To ensure valid GIV, one must select independent SNPs, and the sample size must be larger than the number of SNPs. However, two difficulties remain: there is the possibility of non-linear effects (epistasis) through interactions of genes, and we generally have finite sample sizes. Simulations are used to explore the behavior of the GIV.

What have the preliminary simulations told us about heritability?

From these simulations, we can surmise that estimated heritability is close to true heritability when GWAS size is as large or larger than the assumed number of SNPs. Specifically, the larger the heritability, the smaller the necessary GWAS sample must be in order to produce accurate estimates. However, in practice, we should expect the effects to be different from the simulation in that some of the effects will be zero, while others will have larger effects than the simulated results as a result of random draws from a normal distribution of parameters.

How do we estimate the causal effect of an exposure?

There are two major, intertwined challenges: environmental endogeneity and genetic endogeneity or pleiotropy, the multiple effects of genes. Because we are not experimenting, it is necessary to use other strategies to address this problem. Specifically, we can use reliable and valid instrumental variables to simulate a natural experiment, but these are difficult to find. Another recourse is to analyze monozygotic (MZ) twins; this solves the problem of pleiotropy, but there is limited availability, small sample sizes, and limited environmental variation. Finally, using fixed effects models with panel data eliminates the main effects of pleiotropy and some environmental endogeneity, but panel data is typically inappropriate or unavailable for the research question.

How do we use Mendelian Randomization (MR) as a strategy for estimating the effect of an exposure?

We know that an offspring's genotype is the product of a random combination of the parents' genotype. If we know which genes affect the exposure, we could use them as GIV to estimate the causal effect of the exposure on an outcome. However, this approach introduces challenges. MR requires us to know the genes with the true effects so that we can distinguish them from the environmental confounders that are correlated with ancestry. With traits that are genetically complex, using individual markers as IVs raises the possibility of weak IVs. Finally, we assume that the genotype of a child is conditionally randomized from the genotype of the parents. Thus, if the parental genotype is not controlled, then those of the child will correlate with everything in the environment that the parental genes correlate with.