

NAME OF PRESENTING AUTHOR: Cornelius A. (Niels) Rietveld

EMAIL ADDRESS OF PRESENTING AUTHOR: nrrietveld@ese.eur.nl

Stop meta-analyzing, start instrumenting: Maximizing the predictive power of polygenic scores

Hans van Kippersluis^{1,2}, Pietro Biroli³, Titus J. Galama^{1,2,4,5}, Stephanie von Hinke^{1,2,6}, S. Fleur W. Meddens¹, Dilnoza Muslimova^{1,2}, Rita Pereira^{1,2}, Cornelius A. Rietveld^{1,2,7}

¹Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, The Netherlands

²Tinbergen Institute, The Netherlands

³Department of Economics, University of Zurich, Zurich, Switzerland

⁴Center for Social and Economic Research, University of Southern California, Los Angeles, United States

⁵School of Business and Economics, VU Amsterdam, Amsterdam, The Netherlands

⁶School of Economics, University of Bristol, Bristol, United Kingdom

⁷Erasmus University Rotterdam Institute for Behavior and Biology, Rotterdam, The Netherlands

KEYWORDS:

Polygenic score, Instrumental Variables, Meta-analysis, UK Biobank

ABSTRACT:

A polygenic score is constructed using results of finite-sample Genome-Wide Association Studies (GWASs), and it is therefore a noisy approximation of the true latent genetic predisposition to a certain trait. The conventional way of boosting the predictive power of polygenic scores is to increase the GWAS sample size by meta-analyzing GWAS results of multiple cohorts. In this paper we challenge this convention. Through simulations, we show that Instrumental Variable (IV) regression using two polygenic scores based on independent GWAS samples outperforms the typical Ordinary Least Squares (OLS) model employing a single meta-analysis based polygenic score in terms of bias, root mean squared error, and statistical power. We verify the empirical validity of these simulations by predicting educational attainment (EA) and height in a sample of siblings from the UK Biobank. We show that IV regression *between*-families approaches the SNP-based heritabilities, while applying IV regression *within*-families provides a lower bound on the direct genetic effect. IV estimation improves the predictive power of polygenic scores by 12% (height) to 22% (EA). Our findings suggest that measurement error is a key explanation for hidden heritability (i.e., the difference between SNP-based and GWAS-based heritability), and that it can be overcome using IV regression. We derive a practical rule of thumb for when IV outperforms OLS: When the correlation between the two polygenic scores used in IV regression is larger than the square root of 10 divided by $N+10$, with N the sample size of the prediction sample.

GRANT SUPPORT:

This research has been conducted using the UK Biobank Resource. The authors also acknowledge funding from NORFACE through the Dynamic of Inequality across the Life Course (DIAL) programme (462-16-100), from the European Research Council (DONNI 851725 and GEPSI 946647), from the National Institute on Aging of the National Institutes of Health (RF1055654 and R56AG058726), and from the Dutch Research Council (016.VIDI.185.044).