An Applied Econometric Assessment of the Quality of Evidence Informing Personalized Medicine

Weili Ding Steven F. Lehrer Jenya Lukinova

Queen's University Queen's University, NYU-SH and NBER NYU Shanghai

March 2019

Motivation

The 10 most-read articles on *Healio Cardiology* in 2018

-
    1. Genetic score identifies young patients at risk for MI
    2. Genetic risk score may reshape primary prevention
    3. Genomic risk score predicts CAD better than conventional factors
    4. Personalized approach to antiplatelet drug selection may improve clinical outcomes
    5. Exercise can decrease genetic risk for CVD
    6. Explore the pros and cons of using data on genetic markers.
    7. SGLT1 variants tied to lower risk for HF, diabetes, obesity, death
    8. Genomic medicine may have great potential in clinical settings
    9. Genetic variant may be effective marker for hypertrophic cardiomyopathy
    10. Polygenic risk score predicts early-onset CAD

Now comes the $$$

- Polygenic risk scoring can help clinicians identify populations who are at risk for diseases such as cancer and heart disease in order to optimize prevention and treatment regimens according to an individualized understanding of risk.
- Drug development typically focused on specific gene–RPE65 and Inherited Retinal Disease
- The sweet spot of affordability, access, and innovation. Health policy trilemma
- In July 2018, Clinical-Trials.gov lists 721 gene therapy trials.
- The Myriad myRisk Hereditary Cancer test uses an extensive number of sophisticated technologies and proprietary algorithms to evaluate 29 clinically significant genes associated with eight hereditary cancer sites including: breast, colon, ovarian, endometrial, pancreatic, prostate and gastric cancers and melanoma.
- These scores are estimated and now available in many data sets

Doing Research that Matters feels Good

- There is substantial excitement about new data sources.
- From an empirical researcher perspective, this can open the blackbox of unobserved heterogeneity.
- Literature still in its infancy and terms are (sadly) used interchangeably.
- Existing lessons from microeconometrics and economics should not be forgotten.
- Lehrer (2015) concludes that researchers should shift their attention away from investigating specific candidate genes to polygenic risk scores,...

Disclaimer

- I was hoping to illustrate some of my concerns (am I right versus am I nuts) with one of these new data sources.
- In progress and lack of numbers reflects even more concerns are emerging
- Lesson: New data is available but documentation is incomplete and getting answers takes a while.
- That said, I think the issues I will highlight will be supported by the data
- Put differently, I am displaying the nearly incredible levels of certitude but my certitude is on the skepticism front.

Let's take the con out of geneconomics applications

- The talk will mostly focus on using polygenic scores versus individual genetic markers (SNPs).
- The punchline is that the answer is truly application dependent.
- A polygenic score captures one's risk on the basis of their genetic make-up. The polygenic risk score is either calculated as
    1. The cumulative weighted sum of the variation in multiple genetic locations, weights obtained from coefficients of a GWAS–all SNPs.
    2. The cumulative weighted sum of the variation in multiple genetic locations, weights obtained from coefficients of a GWAS–only significant SNPs.
    3. Unweighted cumulative sum. A simple allele count.

Now that we defined the score, how it is used (then how is it created)

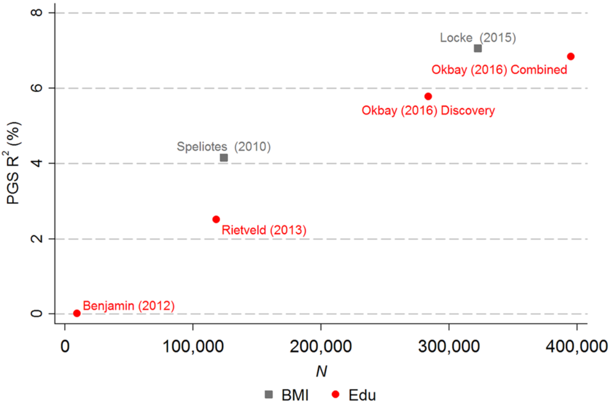
- The polygenic score is a sufficient statistic
- For some economists, they dislike summary statistics particularly in human capital production
- Chetty's bridge and now IO and public economists like sufficient statistics
- Why they like them? Transparency for identification in step 1 and then do your policy simulations later on
- As you will hear transparency in polygenic scores is often absent.

GWAS and gene discovery

- A GWAS is a hypothesis-free scan for associations between a specific outcome and subsets of the millions of genetic variants.

$$EA_i = \sum_{k=1}^{K} \alpha_k SNP + \alpha_{pc} PopStrat + u_i$$

- Assumptions of additive separability and linearity on the genetic effects. What to include is based on linkage disequilibrium.
- Trying to back out heritability, and u_i can then be thought of as environmental factors.
- Large samples needed and many datasets are pooled.
- The α_k are approximate coefficients from a Gibbs sampler that calculate posterior means of effects, conditional on linkage disequilibrium information.
- The consensus emerging in the behavioral genetics literature is that individual markers have very small effects (α_k) on phenotypes of interest to economists.


PGS R^2 (%)
0
2
4
6
8
0
100,000
200,000
300,000
400,000
N
Locke (2015)
Okbay (2016) Combined
Okbay (2016) Discovery
Speliotes (2010)
Rietveld (2013)
Benjamin (2012)
■ BMI
● Edu

Four New from Me Issues Swept Under the Carpet

- GWAS pool data from many studies a la Meta Analysis
  1. In the spirit of Steve Slavin "An Exercise in Mega-Silliness"
  2. The data is confidential, 23 and Me provides one observation only–Ecological Fallacy
  3. Actual versus Implied SNP–Should we treat it the same?
  4. Why are we testing a moment condition and not a functional inequality? For all pitched applications, there is a specific direction in mind.
- My older concerns are coming in a few slides

Small effects–>Application dependent

- Two types of studies where polygenic scores are used.
- Polygenic scores as instrumental variables (returns to education)

$$\begin{aligned} Wage_i &= \beta_x X_i + \beta_{EDU} YearsofEDU_i + \varepsilon_i \qquad (1) \\ YearsofEDU_i &= \alpha_k EAScore_i + X_i \alpha_x + u_i \end{aligned}$$

- Polygenic scores as control variables

$$Wage_i = \beta_x X_i + \beta_{Score} EAScore_i + \varepsilon_i$$

- Studies ignore the earlier stage of GWAS.

$$EA_i = \sum_{k=1}^{K} \alpha_k SNP + \alpha_{pc} PopStrat + u_i$$

Potentially good application

- Instrumental Variables

$$\begin{aligned} Wage_i &= \beta_x X_i + \beta_{EDU} YrsofEDU_i + \varepsilon_i \\ YrsofEDU_i &= \alpha_k Score_i + X_i \alpha_x + u_i \\ EA_i &= \sum_{k=1}^{K} \alpha_k SNP + \alpha_{pc} PopStrat + u_i \end{aligned} \tag{2}$$

versus

$$\begin{aligned} Wage_i &= \beta_x X_i + \beta_{EDU} YrsofEDU_i + \varepsilon_i \\ YrsofEDU_i &= \sum_{g=1}^{G} \alpha_g SNP_g + X_i \alpha_x + u_i \end{aligned}$$

- The main trade-off appears to be the many instrument problem versus interpretation and one should not side-step defending the exclusion restriction assumption; irrespective of how the polygenic score is defined.

Potentially bad application

- Polygenic scores as control variables

$$\begin{aligned} Wage_i &= \beta_x X_i + \beta_{Score} EAScore_i + \varepsilon_i \\ EA_i &= \sum_{k=1}^{K} \alpha_k SNP + \alpha_{pc} PopStrat + u_i \end{aligned} \tag{3}$$

- It does not matter how the polygenic score is calculated, it is a generated regressor. Estimates of the wage equation are i) not consistent, ii) inefficient, and iii) valid inference is not possible with the standard errors.
- Measurement error claims appears second-order at best and disingenuous at worst.
- Why not use two-sample instrumental variables approach and can rely on either GWAS or machine learning strategies for variable selection to explain what is included?

Estimating the reduced form seems safer

- Consider the reduced form of

$$\begin{aligned} Wage_i &= \beta_x X_i + \beta_{Score} EAScore_i + \varepsilon_i \\ EA_i &= \sum_{k=1}^{K} \alpha_k SNP + \alpha_{pc} PopStrat + u_i \end{aligned} \tag{4}$$

- The reduced form is approximately

$$Wage_i = \beta_x X_i + \sum_{l=1}^{L} \alpha_l SNP_i + \alpha_{pc} PopStrat + \varepsilon_i \tag{5}$$

- Clear advantage in it being easier to interpret the effects
- Strategies such as Chernozhukov et al. (2017) can be used to obtain causal effects when there is many covariates–double machine learning but Ding, Lehrer and Xie (2019) point out this estimator flops when there is treatment effect heterogeneity.
- What about discretizing polygenic scores? Measurement errors in discrete indicators cause misclassification bias.

Measurement error expanded or is this the right measure?

- Heritability is generally defined as the proportion of variation in a population that is accounted for by genetic factors, given the importance in the intergenerational transmission of many traits and socioeconomic outcomes.
- Twin studies are viewed as providing upper bounds on heritability. GWAS keep explaining more variation.
- Let's recast a GWAS as

$$\begin{aligned} EA_i &= \sum_{k=1}^{K} \alpha_k SNP + \alpha_{pc} PopStrat + u_i \\ EA_i &= Heritability + v_i \end{aligned} \tag{6}$$

- "Heritability" ignores gene-environment interaction. Tautology and logic comment on polygenic scores as controls.
- Heritability is likely population and time-dependent. Opportunity can reflect the degree to which a genetic or environmental advantage is shaped by choice or circumstance.

Other comments

- The paper discusses applications of gene-environment interactions.
- Many attempts to exploit natural experiments and explore heterogeneity–> gene*environment modifications versus gene*environment responses.
- Are there advantages to looking for gene*environment structural breaks? Environmental stratification in the spirit of Rosenquist et al. (2015).
- Advantages to using theory to add some structure. Biroli (2016) as an example.
- Abusing terminology is also prevalent in the genetics as instruments literature.
- Mendelian randomization versus Mendelian encouragement. Dynastic effects are non-trivial.
- The genetic lottery may hold promise as does adoption studies on the environment side.

Cons

- The majority of evidence so far reflects only simple associations.
- The effect sizes for most genetic factors are very small in magnitude.
- The mechanism underlying how genetic factors operate, either directly or in response to specific environmental stimuli, remains poorly understood.
- Use of genetic data causes concerns for infringement on individual privacy and human rights. The availability of this data may influence decision making and potential discrimination based on one's genotype. Q: Does poorly understood phenomena inhibit macroeconomists?

Pros

- Genetic data provides a useful way to understand individual heterogeneity and often a source of effect heterogeneity.
- By understanding the genetic basis of specific outcomes, policies and treatments could be more efficiently targeted.
- Sheds new insights on the trade-offs made when environments are regulated via socioeconomic policies.
- Proves rich predetermined variation among siblings within the same family to provide a new empirical strategy to identify causal effects.
- My advice for personalized medicine would be to understand the strengths and weaknesses of bagging.

Summing up

- The dangers of automated sophistication with GWAS and polygenic scores appear non-trivial.
- Understanding and making assumptions explicit is crucial.
- Jargon remains a significant barrier to entry.
- Ignoring genetic factors appears unsatisfying and would limit any policy guidance.
- Molecular genetic data does offer the potential to design new effective approaches to improve societal
  outcomes that currently appear intractable with conventional policy options.
- Attention should not be focused upon is whether a specific outcome or trait is primarily a function of genes. Does the available evidence suggests a policy passes a cost-benefit test.
- For policy, society needs to develop a healthy relationship with our genes, one that is neither fanatic nor phobic.