From GWAS to PheWAS: the search for causality in big data

See Articles page e116

Causal investigations in genetics have evolved from agnostic discovery in genome-wide association studies (GWAS) to functional annotation[1] and instrumental variable-informed inference (ie, mendelian randomisation)[2]. In the past decade, big data resources, such as the UK Biobank, have prompted a return to broader discovery through phenome-wide association studies (PheWAS).[3] The work by Elina Hyppönen and colleagues[4] in *The Lancet Digital Health*, joins a small body of studies[5,6] using polygenic risk scores to search for causal effects of an intermediate phenotype such as body-mass index (BMI) on many outcomes, thereby applying mendelian randomisation across the phenome.

The authors used UK Biobank data to construct a BMI genetic risk score based on genetic variants previously identified by the GIANT (Genetic Investigation of ANthropometric Traits) consortium. Using PheWAS followed by mendelian randomisation, Hyppönen and colleagues reproduced the effects of life-long, genetically-influenced changes in BMI on a range of disease outcomes. After Bonferroni correction, PheWAS identified possible associations between BMI genetic risk score and 58 outcomes, and 30 distinct disease associations were supported by follow-up mendelian randomisation analyses. For example, using inverse-variance weighted models, Hyppönen and colleagues found an increase in BMI to be associated with higher odds of endocrine disorders (odds ratio per unit increase in SD of higher BMI 2·72, 95% CI 2·33–3·29, for type 2 diabetes; 2·11, 1·62–2·76, for type 1 diabetes; and 1·46, 1·25–1·70, for hypothyroidism), circulatory diseases (1·96, 1·53–2·51, for phlebitis and thrombophlebitis; 1·89, 1·39–2·57, for cardiomegaly; 1·68, 1·35–2·09, for congestive heart failure; 1·55, 1·37–1·76, for hypertension; 1·31, 1·13–1·52, for ischaemic heart disease; and 1·25, 1·14–1·37, for cardiac dysrhythmias), and inflammatory or dermatological conditions (2·00, 1·72–2·23, for superficial cellulitis and abscess; 3·37, 2·17–5·25, for chronic ulcers of leg and foot; 4·99, 2·54–9·82, for gangrene; and 2·24, 1·53–3·28, for atopy).

The work by Hyppönen and colleagues represents the state-of-the-art in applying a suite of complementary mendelian randomisation methods and adjustments for 40 principal components of polygenic risk scores to address population structure. Unlike a previous study,[5] Hyppönen and colleagues also restricted analyses to clinically-derived or registry-derived outcomes. Crucially, the authors draw attention to some of the more fundamental difficulties of causal discovery when applying this approach. Here, we review a few of these considerations, hoping that they will generate further dialogue. Notably, although challenges exist in using large databases for discovery,[3,7] resilient confounding of polygenic risk scores,[8] and causal interpretations of mendelian randomisation for lifelong traits,[2] have been separately described, the extent of false causal discovery in their joint application is unknown.

As noted by Hyppönen and colleagues[4] and others,[5] false associations among UK Biobank participants might be induced by selection bias. Although corrections such as modelling selection probabilities[7] might be effective for single outcomes, whether similar strategies will be sufficient for PheWAS coupled to mendelian randomisation is not clear. The enrichment of false PheWAS hits have consequences on the number and composition of signals carried forward to mendelian randomisation analyses. For example, past longitudinal investigations[9] have found reduced participation to be associated with BMI, smoking, and mental health polygenic risk scores. It follows that individuals with higher polygenic risk scores for BMI who choose to participate in the study might have lower risk of poor mental health than those who do not participate. This differential participation could explain counter-intuitive associations with reduced neuroticism observed previously[5] or weaken the association with depression observed in this study.[4] Most concerningly, this finding implies that follow-up mendelian randomisation analyses can be enriched with confounded polygenic risk score–outcome associations, as the authors point out, violating the exclusion restriction assumption and invalidating false discovery rate control.

As aptly described by Hyppönen and colleagues, specific assumptions about causal mechanisms are needed to estimate effects with mendelian randomisation (no horizontal pleiotropy, independence of pleiotropic effect sizes, and so on). In single

exposure-outcome mendelian randomisation, our belief that such assumptions are valid can be strengthened by triangulating substantive a priori knowledge, such as experimental findings and functional annotations. In PheWAS coupled to mendelian randomisation, the challenge of justifying independences is multiplied across all target outcomes. This is crucial because the strongest polygenic risk score–outcome associations in the first stage might be the most affected by pleiotropy and, conversely, weaker signals might be excluded even if they are causally valid.

Additionally, a major strength of using genetic variation as so-called causal anchors in standalone PheWAS is that they are fixed at the time of zygote formation and thus, reverse causality is unlikely. However, as the authors point out, this clearly does not extend to BMI–phenotype associations, because pathophysiological development of an outcome can affect observed BMI. Although bidirectional mendelian randomisation can address this issue somewhat, they will only work when statistical power is high because a false negative (not identifying true reverse causation) due to low power will lead to increased confidence in a biased analysis. For conditions where there is a strong preceding suspicion of early onset, such as type 1 diabetes or hypothyroidism, adult BMI will at least partly be determined by disease. In single-outcome studies, one could consider an appropriate latency period after BMI ascertainment and include only incident cases. Automating such a design in PheWAS would be challenging, because latency periods will differ between outcomes.

The use of PheWAS coupled to mendelian randomisation on BMI associations is compelling because adiposity is affected by numerous outcomes during a lifetime and is subject to complex confounding and measurement errors. Crucially, conventional associations do not correspond to well defined causal effects because these will differ greatly depending on how, why, and in whom BMI is changed.[10] Mendelian randomisation presents a potential way forward by providing necessary and sufficient conditions to isolate a particular causal effect—lifelong, genetically-influenced changes in BMI—that might operate similarly, on average, in all humans. However, in coupling mendelian randomisation with discovery (PheWAS), challenges are introduced that might undermine the clarity of any causal investigation. This comes from an inherent tension between structure-free discovery, where statistical inference relies solely on observed distributional characteristics, and causal inference, where validity rests on structural assumptions drawn from previous knowledge. The work by Hyppönen and colleagues represents the state-of-the-art in PheWAS coupled to mendelian randomisation, notably applying strict statistical corrections (ie, pleiotropy identification and multiple-testing corrections) to reduce false discovery. Importantly, their work highlights some broader challenges in balancing discovery and causation that go beyond PheWAS coupled to mendelian randomisation, noting how discovery approaches might amplify non-causal relationships and mask (statistically weaker) causal ones. Although the appropriate balance in PheWAS coupled to mendelian randomisation remains an open question, it is certain that greater consideration for biological function is needed, including—as the authors suggest—a priori negative control outcomes (unlikely to be affected by BMI) and formal use of functional annotations[1] in the development of polygenic risk scores to improve the likelihood of causal discoveries. Ultimately, any causal discovery approach will be successful to the degree that all discoveries are subject to the same careful consideration as single exposure-outcome studies.

Jonathan Y Huang, Jeremy A Labrecque
Singapore Institute for Clinical Sciences, Agency for Science, Technology and Research (A*STAR), 117609 Singapore (JYH); and Department of Epidemiology, Erasmus MC, Rotterdam, Netherlands (JAL)
jonathan_huang@sics.a-star.edu.sg

We declare no competing interests.

1 Iotchkova V, Ritchie GRS, Geihs M, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat Genet* 2019; **51:** 343–53.
2 Labrecque JA, Swanson SA. Interpretation and potential biases of mendelian randomization estimates with time-varying exposures. *Am J Epidemiol* 2019; **188:** 231–38.
3 Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunology* 2014; **141:** 157–65.
4 Hyppönen E, Mulugeta A, Zhou A, Santhanakrishnan VK. A data-driven approach for studying the role of body mass in multiple diseases: a phenome-wide registry-based case-control study in the UK Biobank. *Lancet Digital Health* 2019; **1:** e116–26.
5 Millard LAC, Davies NM, Tilling K, Gaunt TR, Davey Smith G. Searching for the causal effects of body mass index in over 300 000 participants in UK Biobank, using mendelian randomization. *PLoS Genet* 2019; **15:** e1007951.

6 Richardson TG, Harrison S, Hemani G, Davey Smith G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* 2019; **8:** e43657.

7 Swanson JM. The UK Biobank and selection bias. *Lancet* 2012; **380:** 110.

8 Haworth S, Mitchell R, Corbin L, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat Commun* 2019; **10:** 333.

9 Taylor AE, Jones HJ, Sallis H, et al. Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2018; **47:** 1207–16.

10 Chiolero A. Why causality, and not prediction, should guide obesity prevention policy. *Lancet Public Health* 2018; **3:** e461–62.