International Journal of Epidemiology, 2020, 1–3
doi: 10.1093/ije/dyz254
Advance Access Publication Date: 12 December 2019
Editorial

Editorial

Polygenic risk scores and the prediction of common diseases

Mika Ala-Korpela [1,2,3,4,5,6*] **and Michael V Holmes** [7,8,9]

[1]Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, VIC, Australia, [2]Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland, [3]NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland, [4]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK, [5]Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK, [6]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred Hospital, Monash University, Melbourne, VIC, Australia, [7]Medical Research Council Population Health Research Unit, University of Oxford, Oxford, UK, [8]Clinical Trial Service Unit & Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK and [9]National Institute for Health Research Oxford Biomedical Research Centre, Oxford University Hospital, Oxford, UK

*Corresponding author. Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, P.O.Box 5000, FI-90014 Oulu, Finland. E-mail: mika.ala-korpela@oulu.fi

Editorial decision 8 November 2019; Accepted 20 November 2019

There is growing interest in the potential translational applications of omics data. This applies to, e.g. metabolomics, an area in which the *Journal* published a themed issue in 2016 with an accompanying editorial titled 'Metabolic profiling—multitude of technologies with great research potential, but (when) will translation emerge?'[1] Despite two decades of extensive investigations with optimistic statements of potential translational applications, there is no metabolomics-derived biomarker (of either an individual metabolite in isolation or multiple metabolites in combination) that has yet to mature into clinical utility. On appraising the recent activity in polygenic risk scores (PRSs) and disease prediction, we notice parallel themes to the decades-old search for conventional (non-genetic) predictive biomarkers. An overwhelming sense of hype and a rush to translate dominates the field of genetic research of disease prediction using genetic risk scores (GRSs).

A PRS is a combination of single nucleotide polymorphisms (SNPs) that associate with the outcome of interest.[2] There are multiple approaches to constructing PRSs, ranging from inclusion of SNPs surpassing stringent genome-wide significance thresholds (typically called a GRS) to use of millions of SNPs including those that individually only very weakly associate with the phenotype of interest (a PRS). From a statistical standpoint, a GRS or PRS can be considered as a single biomarker similar to an individual (e.g. metabolic) biomarker (or a biomarker score). Thus, we can evaluate the predictive performance of a PRS with the same metrics that have been developed and applied over recent decades. A plethora of literature on the statistical basis of predictive modelling subverts the recent optimism placed in PRSs to predict common complex diseases,[3] the key concepts being that moderate relative risks (achievable by individual SNPs or biomarkers, or their combination into a polygenic or metabolic risk score), struggle to translate into clinically relevant prediction models.[4] These epidemiological principles of disease prediction are robust to 'genetic exceptionalism'.

Several recent high-profile papers have presented interpretations that PRSs convey potential for remarkable

opportunities of improved (clinically relevant) predictions of complex diseases, e.g. coronary heart disease (CHD).[5–7] Parallel themes are well-recognised in biomarker-focused omics research where (as with genome-wide data) technological advances have facilitated the discovery of multiple biomarkers independently associated with disease.[8] However, although such discoveries may provide important aetiological insights into disease, such associations may not necessarily reflect utility in disease prediction.[1,3,4,9]

This is particularly the case for common (polygenic complex) diseases, in other words, for quantitative traits or as put even more succinctly by Plomin *et al.*: 'what we call common disorders are, in fact, the quantitative extremes of continuous distributions of genetic risk'.[10] This continuity of polygenic traits for common complex diseases is superimposed on non-static environmental contributions[11] and stochastic (patho)physiological processes.[12] The oversight in considering these issues, together with unrealistic expectations for 'precision medicine', are likely drivers for the predictive misconceptions.[4,9,13] The ability to reliably categorize individuals into 'healthy' and 'diseased' using biomarkers that are normally distributed under typical physiological settings in the general population—which includes variation in common genetic polymorphisms combined into a PRS, and phenotypic traits, such as low-density lipoprotein (LDL) cholesterol and systolic blood pressure—is likely to remain an unattainable goal.[14] For example, with a 5% false-positive rate, the recently published PRSs by Khera *et al.*[6] and Inouye *et al.*[7] would give a disease detection rate of 15% and 13%. In both these cases, the vast majority (≥85%) of individuals that eventually develop disease would be missed when using such PRSs for disease prediction.[4]

Complex diseases can be considered as the end product of the dynamic interplay between multiple genetic and environmental risk factors. Notably, some of the PRS associations with a disease (or trait) are very likely to be picking up environmental contributions—which may have implications for the temporal performance of a PRS. Unlike genetic variants, environmental risk factors change over the lifespan of individuals and between generations. For example, population characteristics have changed dramatically since the early days of cholesterol and atherosclerosis research. In modern society, individuals have spurious and energy-dense eating patterns, with most individuals living in a non-fasting state. In addition to many general clinical conditions such as obesity, hypertension, insulin resistance and type 2 diabetes, the average population lipid profiles have changed substantially. The metabolic consequences of this relate to the 'contemporary' risk factors of CHD and to a certain extent also to the definition and estimation of PRSs.

An individual's genome is inherited randomly and is not generally modifiable, and these characteristics form the basis for the role of human genetics in elucidating causality through Mendelian randomization. The fact that SNPs are not able to dynamically reflect the extent of disease (or indeed subclinical disease) through reverse causality represents a further hindrance for the use of PRS in disease prediction. If a hypothetical biomarker is generated in response to a disease (i.e. through the process of reverse causality), this may be where such a biomarker might have a role for prediction. Such a biomarker would be different to those that are routinely measured because, unlike LDL cholesterol, the hypothetical biomarker would not be present (or measurable) under normal physiological conditions in disease-free individuals (providing near-perfect discrimination). For example, if the tunica intima of the arterial wall produced a substance in response to subclinical atherosclerosis that 'leaked' into the circulation in such a concentration that it would be detectable before the manifestation of symptomatic disease, but where the same biomarker was not detectable in individuals without disease, this biomarker might be able to discriminate between those that go on to develop disease and those that do not. Whereas a GRS may be used to identify a biomarker arising from reverse causality, the GRS itself in isolation cannot reflect reverse causality.[15]

In contrast to reverse causality, where such a feature may be an advantage of a biomarker for prediction, the causal role of a biomarker is not a requirement for predictive modelling.[15] This is evident from LDL cholesterol, one of the most well-recognized causal biomarkers with well understood molecular pathways and specific drug treatments available; notably, LDL cholesterol is a poor predictor of CHD. However, causality of a biomarker makes all the difference in terms of use in developing population-level interventions for disease prevention.[12,16]

An interesting exception from the predictive perspective are oligogenic medical conditions—that lie between complex and Mendelian diseases—that are likely to be amenable to GRS-based predictions.[3,10] For example, autoimmune diseases may represent one such category where ROC curve values of around 0.9 from a GRS may be feasible.[17,18] Of note, although the high C-statistic does not mean that such a GRS can automatically translate into clinical utility, it is likely a prerequisite for population screening.[3,9] Regarding the potential of genetic prediction, it is notable that studies of monozygotic twins can provide an 'upper limit' of what can be achieved; this information may guide which traits and diseases have sufficient genetic attributes that a GRS could be of potential clinical value.[19] Finally, a GRS captures risk (arising from genetic variants and gene-by-environment interactions) that occurs over a

lifetime, and thus while violating conventional prevention paradoxes that would argue that the focus of preventative strategies should be the entire population rather than just high-risk individuals, the identification of those at high genetic risk may facilitate timely prevention targeted to those who would develop early onset disease: it might therefore be feasible that, e.g., a population-wide treatment with e.g. a polypill given to everyone at say the age of 40 and above might be enhanced with earlier targeted treatment in those at high genetic risk. We note that the availability of genome-wide genotyping facilitated by technological advances and massive reductions in cost are likely to make genotype a readily available trait at the population-level (thus facilitating translational opportunities). Widespread availability of genotyping is likely to occur (at least initially) in high- and middle-income countries, which, together with the predominance of genetic studies being conducted in European populations, may have the net effect of further increasing global health inequalities.

In conclusion, we recognise and celebrate the incontrovertible role that genomics research has, and will continue to provide, in our understanding of the molecular basis of common diseases, in elucidating the mechanisms by which diseases occur and in identifying new therapeutic targets.[20] Nonetheless, the likely inconvenient truth is that for common diseases, no combination of normally-distributed biomarkers, each modestly associated with disease, is likely to lead to clinically-relevant improvements in risk prediction. Geoffrey Rose stated[16] that for common diseases, 'a large number of people at a small risk may give rise to more cases of disease than the small number who are at a high risk', which notably also relates to examining the upper quantiles of a GRS, and concluded that the underlying motivation 'should always be to discover and control the causes of incidence'. This elegant elaboration on sick individuals and sick populations by Rose[16] over 30 years ago was prescient to the contemporary era of big data and genome-wide association studies.

Acknowledgements

We thank Professor George Davey Smith (University of Bristol) for his kind, helpful and insightful comments on an earlier draft of the Editorial. M.A.-K. is supported by a Senior Research Fellowship from the National Health and Medical Research Council (NHMRC) of Australia (APP1158958). He works in a unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_12013/1) and he has a research grant from the Sigrid Juselius Foundation, Finland. M.V.H. is supported by a British Heart Foundation Intermediate Clinical Research Fellowship (FS/18/23/33512). The Baker Institute is supported in part by the Victorian Government's Operational Infrastructure Support Program.

References

1. Ala-Korpela M, Davey Smith G. Metabolic profiling-multitude of technologies with great research potential, but (when) will translation emerge? *Int J Epidemiol* 2016;**45**:1311–18.
2. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* 2016;**17**:392–406.
3. Janssens A, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet* 2008; **17**:R166–73.
4. Wald NJ, Old R. The illusion of polygenic disease risk prediction. *Genet Med* 2019;**21**:1705.
5. Abraham G, Havulinna AS, Bhalala OG *et al*. Genomic prediction of coronary heart disease. *Eur Heart J* 2016;**37**:3267–78.
6. Khera AV, Chaffin M, Aragam KG *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;**50**: 1219–24.
7. Inouye M, Abraham G, Nelson CP *et al*. Genomic risk prediction of coronary artery disease in 480, 000 adults: implications for primary prevention. *J Am Coll Cardiol* 2018;**72**:1883–93.
8. Niiranen TJ, Vasan RS. Epidemiology of cardiovascular disease: recent novel outlooks on risk factors and clinical approaches. *Exp Rev Cardiov Ther* 2016;**14**:855–69.
9. Janssens A, Joyner MJ. Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: is more, better? *Clin Chem* 2019;**65**:609–11.
10. Plomin R, Haworth CMA, Davis O. Common disorders are quantitative traits. *Nat Rev Genet* 2009;**10**:872–8.
11. Saracci R. Epidemiology in wonderland: Big Data and precision medicine. *Eur J Epidemiol* 2018;**33**:245–57.
12. Davey Smith G. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *Int J Epidemiol* 2011;**40**:537–62.
13. Joyner MJ, Paneth N. Promises, promises, and precision medicine. *J Clin Invest* 2019;**129**:946–48.
14. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* 2010;**6**:e1000864.
15. Holmes MV, Davey Smith G. Can Mendelian randomization shift into reverse gear? *Clin Chem* 2019;**65**:363–66.
16. Rose G. Sick individuals and sick populations. *Int J Epidemiol* 1985;**14**:32–38.
17. Abraham G, Tye-Din JA, Bhalala OG, Kowalczyk A, Zobel J, Inouye M. Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genet* 2014;**10**:e1004137.
18. Sharp SA, Rich SS, Wood AR *et al*. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care* 2019; **42**:200–207.
19. Roberts NJ, Vogelstein JT, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE. The predictive capacity of personal genome sequencing. *Sci Transl Med* 2012;**133**:ra58.
20. Visscher PM, Wray NR, Zhang Q *et al*. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet* 2017;**101**:5–22.