Review Article

Cardiovasc Prev Pharmacother 2022;4(1):13-17
eISSN 2671-700X
https://doi.org/10.36011/cpp.2022.4.e7

Cardiovascular Prevention
and Pharmacotherapy

Polygenic risk score: a useful clinical instrument for disease prediction and risk categorization

Jae-Seung Yun

Division of Endocrinology and Metabolism, Department of Internal Medicine, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea

Genetic information is one of the essential components of precision medicine. Over the past decade, substantial progress has been made, such as low-cost, high-throughput genotyping arrays, advances in statistical techniques, and progressively larger discovery datasets, enabling the discovery of alleles contributing to common diseases, such as coronary artery disease and type 2 diabetes. The polygenic risk score (PRS) represents the aggregate contribution of numerous common genetic variants, individually conferring small to moderate effects, and can be used as a marker of genetic risk for major chronic diseases. PRSs can be obtained from early childhood, and only one measurement is needed to determine the score. PRSs can potentially be used for triage of further investigations to confirm disease susceptibility and to optimize individualized preventive strategies for high-risk disease groups. We provide an overview and commentary on important advances in deriving and validating PRSs, as well as the implementation of PRSs for clinically useful purposes.

Keywords: Polygenic risk score; Cardiovascular diseases; Type 2 diabetes mellitus

INTRODUCTION

Precision medicine for common chronic diseases entails the customization of patient care based on genetic, clinical, and environmental information sourced from large populations [1]. Risk prediction models for chronic diseases are useful tools for classifying high-risk groups that require educational or clinical therapeutic interventions. However, traditional risk factors for common chronic diseases do not typically appear early in life, making it difficult to fully identify individuals at high-risk. Genome-level analyses are attractive in that they provide predictive information about the entire trajectory of a disease at an early stage of life; previous studies have reported that around 40% to 50% of phenotypic variance in susceptibility to chronic diseases such as coronary artery disease (CAD) or type 2 diabetes (T2D) can be explained by genetic factors [2,3]. However, over the last decade, genome-wide association studies (GWAS) have demonstrated that each individual genetic variant generally has only a small or modest effect. Thus, from the current perspective, the most common complex diseases that pose public health concerns are highly polygenic in nature, with hundreds or thousands of small-effect genetic variants influencing the development or progression of the disease [4].

This situation has led to the development of models for polygenic risk scoring, which aggregate the effects of multiple single-nucleotide polymorphisms (SNPs) into a single score. More specifically, polygenic risk scores (PRSs)

Received: January 10, 2022; **Revised:** January 13, 2022; **Accepted:** January 20, 2022

Correspondence to Jae-Seung Yun, MD, PhD
Division of Endocrinology and Metabolism, Department of Internal Medicine, St. Vincent's Hospital, College of Medicine, The Catholic University of Korea, 93 Jungbu-daero, Paldal-gu, Suwon 16247, Korea. E-mail: dryun@catholic.ac.kr

represent a method of aggregating an individual's genetic information, weighted by the associations of genetic variants with disease outcomes as identified from GWAS, into simplified scores that capture an individual's susceptibility to disease and can therefore be used for risk prediction [5]. Previously, PRS models were based on a small number of SNPs identified through candidate gene studies and scores calculated simply by summing an individual's risk allele burden for a handful of disease-associated variants, all weighted equally regardless of the underlying strength of association. Recent advances in the organization of large-scale consortia and biobanks have enabled the brisk discovery of genetic variants associated with common complex diseases [6–8]. In addition to these large datasets becoming available, computational and statistical advances have allowed easier derivation, calculation, and validation of PRS models [9]. This has led PRSs to be widely applied in research studies, consequently confirming the genetic contributions of common variants to disease status [3,10]. This review aims to briefly introduce the methods of calculating a PRS, examples of clinical utility for cardiometabolic PRSs, and considerations for the application of PRSs in clinical practice.

THE PROCESS OF CREATING POLYGENIC RISK SCORES

Before discussing the clinical utility of PRS profiling, we briefly introduce the process by which PRSs are developed and evaluated. PRSs are calculated through statistical models, which are constructed via the five-step process outlined below [11].

1) Selection of base and target datasets

2) Single-nucleotide polymorphism selection and weight calculation: First, a base GWAS dataset is needed that provides summary statistics, including beta coefficients and the P-values of genotype and phenotype associations. This dataset is used to identify disease-associated SNPs and their effect sizes. Weighting parameters that affect the calibration and predictive ability of the final model are also determined from the base dataset. In addition, a target dataset consisting of genotypes and phenotypes is required, which should be independent of the base dataset as it is used for calculating individual PRSs and checking the overall performance accuracy.

3) Dataset quality control: To ensure high accuracy and validity when performing a PRS analysis, quality control of both the base and target datasets is very important. Regarding genotype and phenotype data, the quality control process includes addressing missing SNPs and individuals, checking for sex discrepancies, including only SNPs above an appropriate minor allele frequency threshold, and excluding individuals on the basis of relatedness and high or low heterozygosity rates.

4) Shrinkage of GWAS effect size estimates and controlling linkage disequilibrium: Since not all SNPs influence the trait under study, using the unadjusted effect size estimates of all SNPs could lead to poor prediction outcomes with high standard error. To avoid this problem and control for linkage disequilibrium, two methods for shrinking the effect size estimates are broadly used: (1) statistical shrinkage/regularization techniques such as Lasso or ridge regression, or Bayesian approaches that perform shrinkage via prior distribution specification, and (2) a P-value selection threshold with clumping. In some cases, using a threshold below that for genome-wide statistical significance may improve performance, often at the expense of generalizability. PRSs can be readjusted with consideration of effect size biases, including the effect size inflation typical in a base dataset, the presence of multiple linked variants within each disease-associated locus, subphenotypes of interest, and ethnic or demographic factors that may influence generalizability.

5) Generation and validation of PRSs: PRSs can be generated by a mathematical algorithm that combines selected SNPs, assigns weighting based on effect size, and determines the best-performing SNPs in the population of the independent target dataset. Separate populations should be used to confirm the validity of a PRS, and considerations that should be taken into account include the PRS unit, population genetic structure, PRS distribution, and overfitting.

GENETIC EFFECTS OF COMMON VARIANTS ON COMMON COMPLEX DISEASES

The clinical utility of GWAS-based inherited risk estimates has mainly been assessed in terms of discriminative ability as described by the area under the curve, sensitivity, and specificity [5,12,13]. The area under the curve is a population-level metric reflecting the overall probability of discriminating individuals with a given disease from those without [14]. This metric cannot provide information regarding the absolute predictive risk conferred to a single individual or subgroup. In contrast, a PRS can identify a subset of individuals whose absolute risk of disease is significantly higher than that of the average individual in the general population. One of the major utilities of PRSs may be the comprehensive stratification of an overall population through accounting for each individual's or subgroup's respective genetic susceptibility [15].

In previous papers, PRSs have demonstrated considerable performance in predicting target diseases. Khera et al. [5] showed that, in five common diseases, including CAD and T2D, genome-wide PRSs could identify a larger fraction of the population than could carrier status for rare monogenic mutations, while considering a comparable risk level. With respect to CAD in particular, individuals in the top 1% of the PRS distribution had 4.8-fold higher risk than the remaining 99%. In addition, relative to the carrier frequency of familial hypercholesterolemia, which is caused by a rare monogenic mutation, the PRS identified 20-fold more individuals at comparable risk. Those high-risk individuals could benefit from tailored health management strategies, which may include intensive screening or more invasive preventive interventions. Another study suggested that, compared to individuals with average PRSs, those in the top 2.5 percentile experienced the onset of common diseases 4.3 to 6.6 years earlier [13]. They also evaluated the effect of adding the PRS to clinical risk scores on discriminative and reclassification performance and found that the addition of PRS significantly enhanced both the concordance index and net reclassification improvement.

Genetic risk may be modifiable through adherence to a favorable lifestyle or medication. Previous studies have demonstrated the effects of PRSs, lifestyle factors, and their interaction on the risk of chronic metabolic diseases. In particular, Khera et al. [16] suggested that genetic and lifestyle factors are independently associated with CAD risk. Even in individuals at high genetic risk, a favorable lifestyle reduced the relative risk of CAD by nearly half relative to an unfavorable lifestyle. The log additive effects of combined genetic composition and lifestyle behaviors on disease risk have also been reported in atrial fibrillation, stroke, hypertension, and T2D [17].

CLINICAL UTILITY OF POLYGENIC RISK SCORES

Combining PRSs with traditional clinical risk factors can further refine prevention strategies by enabling more sophisticated risk stratification (Fig. 1). Although PRSs are not yet routinely used in clinical practice, they are proposed to have several advantages. First, PRSs can be ascertained

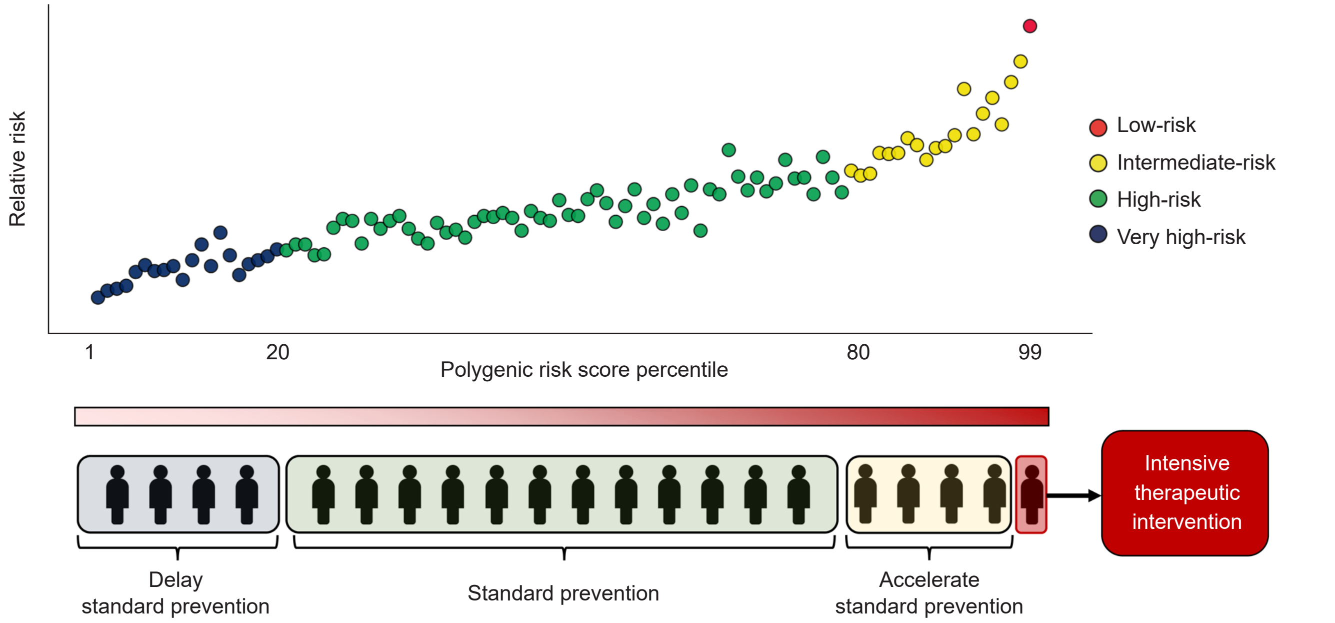

Fig. 1. Risk stratification using polygenic risk scores and the selection of preventive strategies according to risk groups.

from birth, and only one measurement is needed to determine the score. Cardiovascular risk captured by PRSs appears to be largely independent of traditional risk factors, which usually manifest in middle-aged adults [18]. Thus, lifestyles can be optimized from infancy and genetic information can potentially influence motivation for disease prevention from very early in a patient's life. Second, PRSs may be appropriate for screening common metabolic diseases in large community populations, especially in young adults. Several results have demonstrated their utility in decisions for early or late initiation and the interpretation of disease screening (e.g., for breast, colorectal, or prostate cancer) [19–21]. Third, PRSs can affect the disease course or progression as well as disease incidence. We previously demonstrated that individuals with high PRSs for CAD and T2D had significantly higher risk of cardiovascular mortality, the most progressive manifestation of CVD [22]. Fourth, PRSs can factor in target selection for interventions to treat or prevent disease. Individualized treatment is central to precision medicine, and applying PRSs to improve individual patient care would benefit the decision-making process for diagnosis and treatment and thereby guide therapeutic interventions. A high PRS may assist in triaging individuals in a borderline risk group. In the context of cardiovascular disease, there has been considerable debate as to whether PRSs are appropriate as interventions for primary preventive treatments, such as antiplatelet drugs and statins. For example, in individuals with borderline low-density lipoprotein cholesterol levels, a CAD PRS in addition to traditional risk factors may permit a more detailed categorization of risk, thus influencing the decision to initiate statin therapy [23,24].

FUTURE DIRECTIONS

PRSs offer a valuable opportunity to improve the early identification of actionable cardiovascular risk. Nonetheless, despite encouraging findings, the clinical utility of PRS risk estimation remains limited. Most studies have primarily focused on individuals of European ancestry, and generalization to other ancestries might be difficult due to linkage disequilibrium. In addition, evidence relating to PRS application is still lacking, and many models based on a PRS alone still have much lower area under the curves than traditional multivariate models, as expected from the use of a single risk factor due to several limitations. Thus, further research should be conducted, especially more large-scale prospective studies examining the clinical utility of PRSs, and caution remains needed when interpreting PRS results and applying them to the general population. The dissemination of false deterministic beliefs should be rejected. As methodological advances increase the accuracy of PRS determinations and so continue to improve PRS estimates, it is expected that PRSs will become more useful.

ARTICLE INFORMATION

Ethical statement

Not applicable.

Conflicts of interest

The author has no conflicts of interest to declare.

Funding

None.

ORCID

Jae-Seung Yun, https://orcid.org/0000-0001-5949-1826

REFERENCES

1. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med 2015;372:793–5.
2. Udler MS, McCarthy MI, Florez JC, Mahajan A. Genetic risk scores for diabetes diagnosis and precision medicine. Endocr Rev 2019;40:1500–20.
3. Levin MG, Rader DJ. Polygenic risk scores and coronary artery disease: ready for prime time? Circulation 2020;141:637–40.
4. Dudbridge F. Polygenic epidemiology. Genet Epidemiol 2016;40:268–72.
5. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018;50:1219–24.
6. Preuss M, Konig IR, Thompson JR, Erdmann J, Absher D, Assimes TL, et al. Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis CARDIoGRAM) Study: a genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. Circ Cardiovasc Genet 2010;3:475–83.

7. Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet 2011;43:339–44.

8. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12:e1001779.

9. Figtree GA, Vernon ST, Nicholls SJ. Taking the next steps to implement polygenic risk scoring for improved risk stratification and primary prevention of coronary artery disease. Eur J Prev Cardiol 2020 Nov 4 [E-pub]. https://doi.org/10.1093/eurjpc/zwaa030.

10. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med 2020;12:44.

11. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc 2020;15:2759–72.

12. Mosley JD, Gupta DK, Tan J, Yao J, Wells QS, Shaffer CM, et al. Predictive accuracy of a polygenic risk score compared with a clinical risk score for incident coronary heart disease. JAMA 2020;323:627–35.

13. Mars N, Koskela JT, Ripatti P, Kiiskinen TT, Havulinna AS, Lindbohm JV, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med 2020;26:549–57.

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. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet 2018;19:581–90.

16. Khera AV, Emdin CA, Drake I, Natarajan P, Bick AG, Cook NR, et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. N Engl J Med 2016;375:2349–58.

17. Said MA, Verweij N, van der Harst P. Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the UK Biobank Study. JAMA Cardiol 2018;3:693–702.

18. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, 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.

19. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst 2015;107:djv036.

20. Hsu L, Jeon J, Brenner H, Gruber SB, Schoen RE, Berndt SI, et al. A model to determine colorectal cancer risk using common genetic susceptibility loci. Gastroenterology 2015;148:1330–9.

21. Eeles R, Goh C, Castro E, Bancroft E, Guy M, Al Olama AA, et al. The genetic epidemiology of prostate cancer and its clinical implications. Nat Rev Urol 2014;11:18–31.

22. Yun JS, Jung SH, Shivakumar M, Xiao B, Khera AV, Park WY, et al. Polygenic risk, lifestyle, and cardiovascular mortality: a prospective population-based UK Biobank study [Preprint]. Posted 2021 Feb 17. medRxiv 2021.02.15.21251790. https://doi.org/10.1101/2021.02.15.21251790.

23. Kullo IJ, Jouni H, Austin EE, Brown SA, Kruisselbrink TM, Isseh IN, et al. Incorporating a genetic risk score into coronary heart disease risk estimates: effect on low-density lipoprotein cholesterol levels (the MI-GENES Clinical Trial). Circulation 2016;133:1181–8.

24. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet 2010;376:1393–400.