Development and Implementation of Polygenic Risk Score in Vietnamese Population

Nguyen Tran The Hung, Le Duc Hau

Department of Computational Biomedicine, Vingroup Big Data Institute, Hanoi, Vietnam

Correspondence: Le Duc Hau, v.hauld1@vintech.net.vn

Communication: received 18 October 2019, revised 23 December 2019, accepted 29 December 2019

Digital Object Identifier: 10.32913/mic-ict-research.v2019.n2.893

The Editor coordinating the review of this article and deciding to accept it was Prof. Le Hoang Son

Abstract: Recent technological advancements and availability of genetic databases have facilitated the integration of genetic factors into risk prediction models. A Polygenic Risk Score (PRS) combines the effect of many Single Nucleotide Polymorphisms (SNP) into a single score. This score has lately been shown to have a clinically predictive value in various common diseases. Some clinical interpretations of PRS are summarized in this review for coronary artery disease, breast cancer, prostate cancer, diabetes mellitus, and Alzheimer's disease. While these findings gave support to the implementation of PRS in clinical settings, the populations of interest were derived mainly from European ancestry. Therefore, applying these findings to non-European ancestry (Vietnamese in this context) requires many efforts and cautions. This review aims to articulate the evidence supporting the clinical use of PRS, the concepts behind the validity of PRS, approach to implement PRS in Vietnamese population, and cautions in selecting methods and thresholds to develop an appropriate PRS.

Keywords: *Genetic, clinical, single nucleotide polymoyphism (SNP), polygenic risk score (PRS).*

I. Renewed Interest in Polygenic Risk Score

1. Definition

A Polygenic Risk Score (PRS) in the context of genetic studies is a mathematical aggregation of risk effects conferred by many Single Nucleotide Polymorphisms (SNP). Each SNP contributes a small effect to the development of a disease or a complex trait of interest. In the early days of Genome-Wide Association Study (GWAS), researchers expected to find genetic variants that have a large effect on disease risk [1]. While the sample size of GWASs has already surpassed hundreds of thousand of individuals, they failed to capture the genetic variants that can explain the heritability of common diseases, such as breast cancer, prostate cancer, coronary artery disease, diabetes mellitus, Alzheimer disease [2]. Therefore, there is a growing interest in combining all the small SNP effects into a single score that has significant and applicable values [3, 4].

2. PRS Calculation

In its simplest form, the PRS of an individual can be calculated as the sum of all effect sizes of the effective alleles observed in its genotype. The formula to calculate the PRS is given as follows:

$$\mathrm{PRS}_i = \sum_{j=1}^{m} x_{ij} \times \widehat{B}_j,$$

where PRS_i is the risk score for the i^{th} individual, m is the number of SNPs included in the calculation, x_{ij} is the genotype of the i^{th} individual for the j^{th} SNP (can be 0, 1, or 2 depending on the inheritance model), and $\widehat{B}_j$ is the effect size of the j^{th} SNP, usually obtained from GWAS summary statistics [4].

Although the concept of PRS is as old as the finding of genetic materials (DNA), modern technology allows the integration of more genetic variants and more precise effect sizes. Therefore, there are numerous considerations and thresholds related to developing and validating the formula of PRS that are still controversial [5].

3. Advancements in the Field of Genetics

At this junction, there are many developments of technology and findings of new studies that facilitate the development of PRS. The availability of various populations' reference human genomes can be accessed publicly in the 1000 genomes project [6]. Thousands of GWASs comprise of up to millions of samples. These GWAS summary statistics data can be easily accessed through the

"GWAS Catalog" [7], which is an online database with more than 4000 published studies.

New analysis methods for developing the PRS without relying solely on genome-wide significant hits continue to appear, such as Clumping + Thresholding [8], Penalized Regression [9]. The access to large genotype and phenotype data of large longitudinal cohorts becomes easier through online databases such as "dbGAP" [10] and "UK biobank" [11].

II. Clinical Usage of PRS

While making clinical decisions, doctors often have to classify the susceptibility of a patient based on known risk factors. This disease risk classification is very important in providing an appropriate recommendation for the patient. A group of individuals having certain risk factors could have higher relative risk than the general population to guarantee different clinical management. If existing medical intervention can provide more benefits than adverse effects, with reasonable costs, this group of high-risk individuals would receive more benefits from it. Recent studies have suggested that genetic profiling using the PRS can provide some clinical utilities [12].

PRS analysis and its interpretation revolved around some situations: risk prediction performance of PRS independently or in combination with other non-genetic risk factors and estimation of lifetime risk trajectories. Some recent studies have proposed some clinical interpretations of PRS that can modify therapeutic intervention, disease screening and life planning [13]. This review highlights some recent findings of PRS in certain common diseases: coronary artery disease, diabetes mellitus, breast cancer, prostate cancer and Alzheimer's disease.

1. Coronary Artery Disease

Clinical risk scores like the Framingham risk score is a traditional tool in evaluating 10-year coronary artery disease (CAD) risk [14]. This score uses clinical risk factors to score each individual and infer his chance of developing CAD in the next 10 years. Abraham *et al.* have proved that integrating the PRS in traditional clinical risk model can better capture the lifetime risk of CAD in patients [15]. This argument was supported by better C-statistic (measure of goodness-of-fit) in the combined model as compared to the clinical one. More importantly, men in the top quintile of PRS had 10% cumulative CAD risk around 15 years earlier than men in the bottom quintile.

In the primary prevention setting, statin can be used to treat atherosclerosis and reduce the risk of cardiovascular events [16]. The PRS can identify a group of patients having high genetic risk for CAD, who can receive more benefits from primary prevention with statin therapy [17]. Patients having the top quintile of PRS have higher risk of subclinical atherosclerosis and receive greater absolute risk reduction of CAD event from statin therapy (Figure 1).

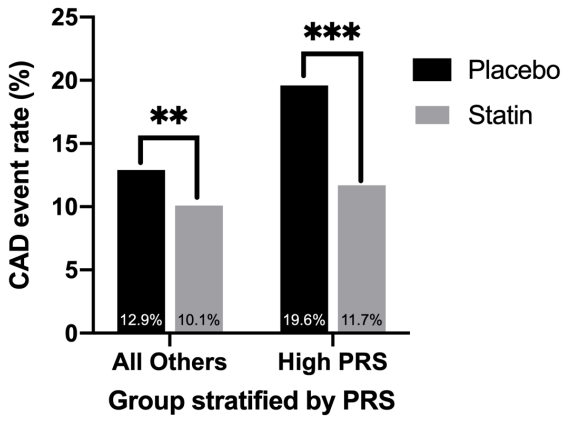

Figure 1. CAD incident by PRS group in statin primary prevention trial. Adapted from a study by Natarajan et al. (2017) [17]. **: chi-square test with p-value < 0.01. ***: chi-square test with p-value < 0.001.

2. Breast Cancer

Breast cancer screening has been recommended for women older than 50 without major risk factors for a long time [18]. The reasoning behind screening for breast cancer in women older than 50 is reducing disease mortality and decreasing false positive diagnosis. A study by Pashayan *et al.* argued that a well-defined risk-stratified screening strategy would improve the quality of life of women and save resources [19]. Based on this risk/benefit threshold, a risk prediction model combining clinical risk factors and the PRS could identify a subgroup of women who had relative risk of developing breast cancer higher than that of 50-year-old women [20]. These high-risk individuals could benefit from earlier screening tests and assertive lifestyle change to reduce certain risk factors. The women in the top quintile of PRS could have the same relative risk as average 50-year-old women around 5-10 years earlier (Figure 2).

3. Prostate Cancer

Current medical guideline suggests that the age to consider prostate cancer screening is 50 years for average-risk men as long as life expectancy is at least 10 years [22]. A study by Seibert *et al.* [23] argued that the PRS was a significant predictor for the age of prostate cancer onset. It was also a relatively inexpensive evaluation of individual's benefits from prostate cancer screening.

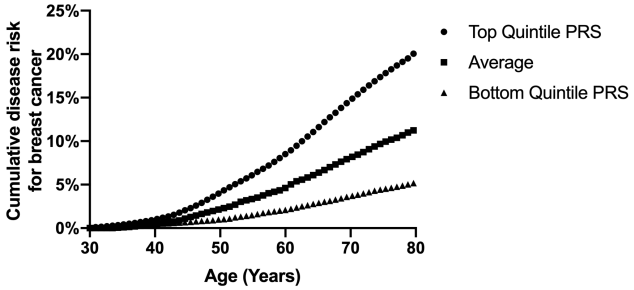

Figure 2. Cumulative breast cancer risk stratified by PRS quintile. Adapted from a study by Maas et al. (2016) [21].

4. Diabetes Mellitus

Early detection of individuals with high risk of type 1 diabetes (T1D) allows better monitoring and prevention of disease progression. Redondo *et al.* evaluated the performance of PRS on T1D patients' relatives without diabetes and with one or more positive autoantibodies [24]. Progression to T1D was best predicted by a combined model of PRS, number of positive auto-antibodies, DPT-1 Risk Score [25], and age. Individuals at high risk of developing T1D can benefit from monitoring and prevention trials.

A prospective cohort study by Lall *et al.* [26] used a PRS that had the strongest association with type 2 diabetes (T2D) in a population-based cohort and evaluated its performance on a prospective individual risk assessment. The hazard for incident T2D was more than 3 times higher in the top quintile of PRS, as compared to others.

5. Alzheimer's Disease

One of the most common causes of dementia is Alzheimer's disease. Desikan *et al.* studied the PRS performance in stratifying Alzheimer's disease risk [27]. This study argued that the PRS could be integrated into screening for individuals with age-specific high genetic risk for Alzheimer's disease. This finding has not been found its use in clinical settings yet, but may prove to be useful for therapeutic trials.

III. Validity of PRS

1. Construct

When the PRS was constructed, it was assumed that SNPs had additive effects on the disease. Because of the large number of SNPs and its unexplained characteristics with the disease of interest, GWAS typically chose the additive model for statistical analysis [28]. However, the biological reality is assuredly more complicated than that. The mode of inheritance of a SNP could be additive, multiplicative, recessive or dominant [29]. Performing only additive model tests could avoid multiple comparisons but overlooked the other inheritance models. Besides, when the gene-gene interaction and gene-environment exposure were taken into account, the model became much more complicated and current statistical methods could not keep up with this complexity [30, 31].

2. Content

In the context of implementation, the PRS was used to predict the genetic disease risk of common diseases. The content of the PRS needed to capture all of the genetic variations with the purpose of reflecting the genetic liability of the disease. However, for many common diseases, genetic variation only accounted for a small portion of the disease phenotype [2].

The diagram in Figure 3 illustrates the contributing factors to T2D development and the way some T2D-risk prediction model were constructed. Conceptually, the SNPs having effect on a complex trait such as T2D consist of SNPs that modified intermediate phenotypes (blood pressure, BMI), which eventually contribute to the risk of T2D. These intermediate phenotypes present themselves as clinical risk factors. T2D is also affected by factors independent with genetics such as age, lifestyle (Figure 3).

The conventional risk prediction model used for T2D in clinical settings only included clinical risk factors [32]. Recent findings in the field of genetics suggested that combining clinical risk factors and the PRS could improve the current risk prediction model and the cost-benefit metrics [33]. The caveat of this approach was that many clinical risk factors are not independent with genetics. The combined prediction model might have the effect of genetic factors and clinical factors of the same pathogenic mechanism counted simultaneously (*e.g.*, the effect of SNPs associated with blood pressure and the effect of clinical high blood pressure were both included in the prediction model in Figure 3). Although the combined model showed improved C-statistic, a single mechanism (blood pressure/BMI) was counted twice: one in the genetic feature and the other in the clinical one. The outcome would be biased toward that mechanism. Another approach was to evaluate the prediction model only with the PRS, stratified by age [26]. This approach highlighted the independent nature of the PRS and visualized the cumulative risk of the disease of the high-risk group compared to the general population.

3. Criterion

Whether the PRS has valid predictive power depends on the specific disease and population. A review by Duncan *et al.* [34] found out that the majority of PRS studies included

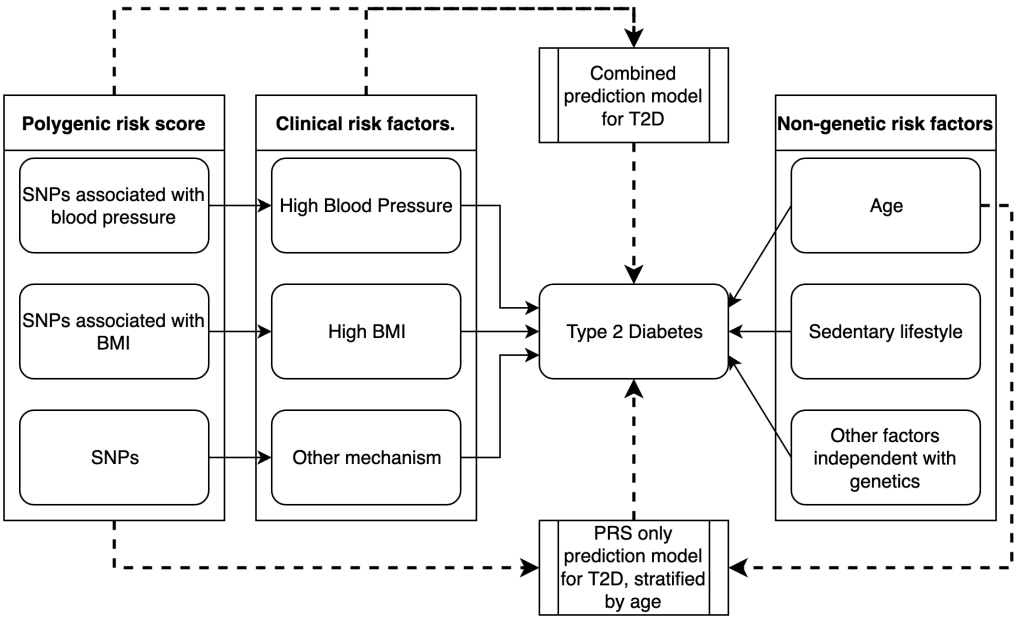

Figure 3. Risk factors of T2D and content of the prediction model for T2D.

European ancestry and East Asian ancestry participants. A PRS derived from the European population had lower performance in the non-European population. As a result, if we wanted to apply PRS utilities in the Vietnamese population, we had to improve methodological choice and threshold to accommodate the difference in linkage disequilibrium and variant frequency between Vietnamese and European/East-Asian. The clinical utility of the PRS needed to be validated in a prospective cohort study [35].

IV. Implementing PRS in Vietnamese Population

1. Method to Read DNA

Genotyping is a method of determining which genetic variants an individual possesses. SNP microarrays allow detection of hundreds of thousands of pre-determined SNPs. In genetic research, SNP arrays are most frequently used for GWASs. A commercial genotyping array included around 500,000 SNPs and cost around 100 USD [37].

Sequencing is the process of determining the DNA sequence, which is the exact order of DNA's bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Whole-genome sequencing and whole-exome sequencing are more expensive but they allow a more precise detection of genetic variations (*e.g.*, structural variants). Depending on the region, a given stretch of sequence may include some DNA that varies between individuals, in addition to the constant region. Thus, sequencing can be used to determine the genotype of an individual for known variants, as well as identify variants that may be unique to that person [38].

The cost of genotyping array is lower and, thus, more suitable for large scale research in the population. However, the commercial genotyping array used in GWAS has a strong ascertainment bias because SNPs are chosen from European individuals [39]. The best way to overcome this problem is to design an SNP-array suitable for the Vietnamese population.

2. Available Databases

The 1000 genome project, which shared reference genome from diverse populations, contained over 88 million variants from 2504 individuals [6]. It provided a broad representation of human genomes in different populations and ethnicities. This database contained 101 Vietnamese individuals (Kinh ethnic from Ho Chi Minh city). The availability of the Vietnamese reference genome is very important for researchers who are interested in conducting genetic research in Vietnam. The more references there are, the better it can represent Vietnamese human genomes. As a result, risk prediction based on genetic factors will be more reliable and accurate. To this end, a Vietnamese genetic database is being built [40]. As data grow larger, the prospect of implementing genetic findings in Vietnamese clinical settings become more imminent.

Of course, this is only the first step in genetic research in Vietnam. In order to catch up with international research and development, Vietnam still has a long way

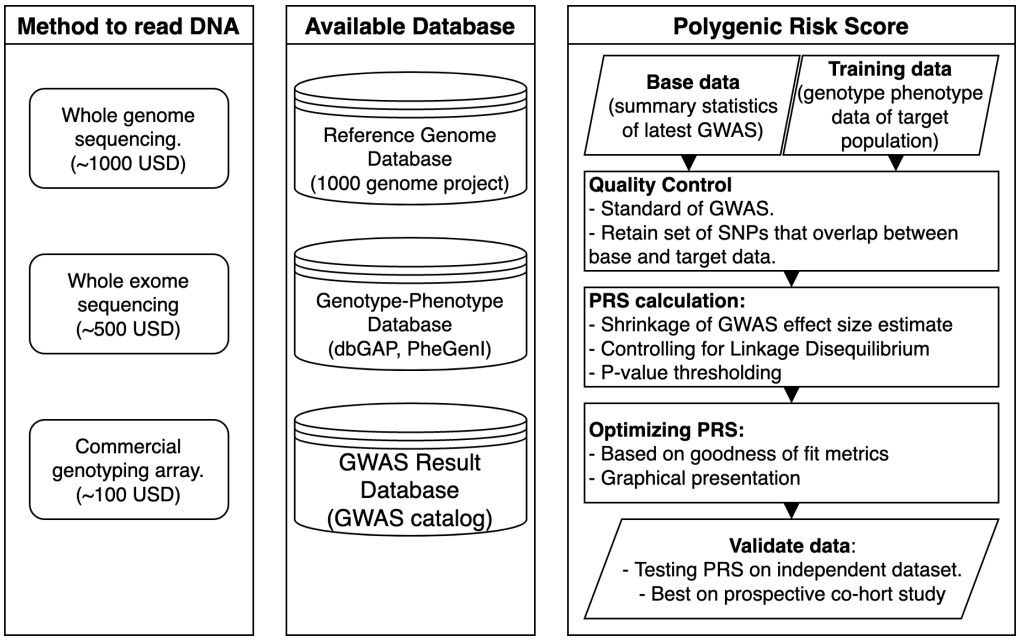

Figure 4. Resources of genetic studies and PRS analysis. Adapted from a tutorial by Choi et al. (2018) [36].

to go. There are some established databases of human genotype-phenotype like dbGaP [41] and PheGenI [42]. These databases allow researchers to share their datasets and to perform large scale and complex analysis on various diseases. A genotype-phenotype database of the Vietnamese population for replicating published genetic findings is sorely needed. The availability of such a database will facilitate the replication of PRS research with high applicability and will optimize the predictive performance in the Vietnamese population. Besides, replication of GWAS with a large Vietnamese sample size will also provide valuable information for genetic research.

V. Polygenic Risk Score Analysis

Polygenic risk score calculation can be characterized by two input data sets: the base data (the summary statistics of the latest GWAS concerning the disease of interest) and the training data (the individual genotype-phenotype from the population of interest) as illustrated in Figure 4.

1. Quality Control of Input Data

Since the base data come from GWAS, input data must be quality-controlled (QC) according to the standard of GWAS. According to a tutorial on conducting GWAS by Marees *et al.* [28], QC steps for a successful GWAS are:

1) Exclude SNPs that are missing in greater than 2% of the subjects;
2) Exclude individuals with genotyping rate greater than 2%;
3) Exclude individuals with sex discrepancies between recorded data and their X chromosome's heterozygosity;
4) Include SNPs with minor allele frequency (MAF) above the threshold (based on the sample size of the study, larger sample size can use lower threshold);
5) Exclude SNPs that deviate from Hardy-Weinberg equilibrium (for binary traits HWE p-value is less than 10^{-10}, for quantitative traits HWE p-value is less than 10^{-6});
6) Exclude individuals with heterozygosity rate deviated more than 3 standard deviations from the population mean;
7) Perform principal component analysis on the training data and using the first 10 eigenvectors as covariates.

All these steps can be done with plink 1.90 software [43].

Because base data and training data come from different sources, some QC steps need to be taken according to Choi *et al.* [36]:

1) Check the integrity of the transferred file with a software like md5sum [44];
2) Ensure that input data have genomic position of the same genome build with LiftOver program [45];
3) Define the effect allele and the reference allele from the base data since some GWAS summaries categorize the allele as risk allele/non-risk allele and major/minor allele;

TABLE I
DIFFERENT METHODS OF COMPUTING PRS. ADAPTED FROM A STUDY BY CHOI ET AL. (2018) [36]

	Clumping + thresholding	**Penalized Regression**	**Bayesian Shrinkage**
Shrinkage of SNPs' Effect Size	P-value threshold	LASSO, penalty parameters, Elastic Net	Fraction of causal SNPs
Handling of Linkage Disequilibrium	Clumping	LD matrix integral to the algorithm	Shrink effect sizes with respect to LD
Software	PRSice [8]	Lassosum [48] bigstatr [9]	LDpred [49]

4) Process the ambiguous SNPs due to unknown chromosome strand (sense/antisense) from different DNA read platform; removing duplicated SNPs;
5) Ensure the independence of base data and training data by removing overlapping samples and closely related individuals (1st/2nd degree relatives);
6) Check chip-heritability from GWAS summary statistics by using LD score regression [46] to avoid reaching misleading conclusion (because the predictive power of the PRS cannot surpass the power and predictive accuracy of the base GWAS [47]).

2. PRS Calculation

After QC has been performed, PRSs are calculated for all individuals in the training data. There are many methods to calculate the polygenic risk score (see Table I), but a good method has to take into account the following factors. The first factor is the effect sizes of SNPs taken from GWAS summary statistics. These effect sizes consist of true association with the disease and some degree of unknown random variations which produce "winner's curse" effect among the most significant SNPs. The second one is the number of SNPs included in the calculation. And the third one is the correlation among SNPs, *i.e.*, linkage disequilibrium. LD difference between base and training data can make the PRS misestimate the risk of disease.

Clumping-and-thresholding is the most common method to compute the PRS. Clumping in this context means selecting the most significant SNP in a block of related SNPs based on LD (usually $r^2 > 0.2$). The p-value threshold is usually genome-wide significant 5×10^{-8} or the threshold that maximizes the AUC is selected. The clumping-and-thresholding method is fast and easy to use. It can capture the independent effect of the SNPs. However, a compelling criticism of this method is the removal of SNPs in LD with arbitrary r^2 value. Clumping-and-thresholding is automated in PRSice software [8]. This is one of the most common tools for computing the PRS. The processes of calculating, evaluating and visualizing the result are automated in the software.

Penalized regression is a method to integrate LD information of the population of interest into GWAS summary statistics. This technique allows for decreasing the variance at the cost of introducing some bias. The selection of tuning parameters λ and s in the elastic net technique results in a model's best fit. This model is complicated to calculate but performs better than the clumping-and-thresholding method in a small sample size [9].

Bayesian shrinkage is a method that infers the mean effect size of SNP based on a reference LD panel and assumption of causal SNP fraction in the genotype. Simulation and empirical data prove that this method outperforms the clumping-and-thresholding method, particularly in a large sample size [49, 50].

Penalized regression and Bayesian shrinkage lead to millions of SNPs included in the PRS calculation. Most of these SNPs have a negligible effect on individual risk score. Their effect sizes are so small and would have been zero if the effect size would be rounded up to 3 or 4 digits after the decimal point. This approach aims to maximize the C-statistics even if the majority of the included SNPs have minimal effect on individual risk score [50]. This approach, however, convolutes the PRS formula beyond the biological and medical interpretation.

3. Optimizing PRS and Result Presentation

The association between the PRS and the phenotype can be quantified with goodness-of-fit metrics such as the estimate of effect size (beta for quantitative phenotype and OR for binary phenotype), p-value of the null hypothesis test of no association, explained variance (R^2) and area under the curve (AUC).

Explained variance is a statistical measure to quantify the discrepancy between the PRS model and the observed phenotype. A higher percentage of explained variance means a better association of the model, which also means that the PRS has better predictive performance. In linear regression analysis for quantitative phenotype, explained variance is the coefficient of determination. In logistic regression analysis for binary phenotype, explained variance is the pseudo-R^2. Various types of pseudo-R^2 metrics are used in epidemiology field. Among them, Nagelkerke R^2 is perhaps the most popular [51].

The receiver operating characteristic curve illustrates the true positive rate (sensitivity) versus false-positive rate (specificity). The AUC represents the accuracy of the test. A test with AUC between 0.9–1 is an excellent test. A test with AUC between 0.5–0.6 is a worthless test [52].

Result presentation of the PRS analysis should emphasize its relevant usage which is a prognosis of disease liability. The main goal of the PRS is the identification of a group of individuals with high-risk of disease based on genetic factors. Therefore, the result should stratify the population into distinct tiers of risk based on the percentile rank cut-off value, usually top 1 percentile, top 10 percentile, top 20 percentile, etc., *e.g.*, as presented in Figure 1. Generally, the medical community has already defined the appropriate level of risk versus benefits that justifies certain medical interventions [53, 54]. When an individual is determined to have high-risk for a disease, we can infer the relative risk of said individual compared to the reference. The higher the relative risk of said individual is, the more justified the medical intervention becomes.

PRS analysis can also be represented based on age (Figure 2). The cumulative risk of disease stratified by the PRS can guide the decision at which age an individual can benefit the most from a screening test [55]. This age-based criterion can spotlight the balance of average risk of breast cancer and the risk of harm due to the false-positive result.

4. Validating PRS Performance

A common concern in PRS analysis is whether the most optimized PRS overfits the training data [56]. As a result, applying said PRS to the general population can lead to inflated results and false conclusions. The best strategy to prevent overfitting of the PRS-based prediction model is to validate its accuracy on an independent data set. In the absence of an independent data set, the training data can be divided into 2 separated data sets, one for optimizing the PRS and the other for performing out-of-sample prediction [57].

VI. Conclusion

The cost of reading DNA is becoming more and more affordable through advancement of genotyping and sequencing technologies. Alongside the development of data storage, new computing methods and abundance of disease databases, the PRS has provided better accuracy to existing models of risk prediction for common diseases. Consequently, individual clinical management (*e.g.*, disease screening and therapeutic intervention) can be personalized based on individual genetic information. This genetic information can be obtained at any point in life with a minimally invasive procedure (*e.g.*, blood draw or saliva sample) and a single genotype data can be analyzed to provide estimations for many diseases simultaneously. Although the medical community still has doubt and hesitation regarding implementation of the PRS, it will continue to improve and have larger impact in the near future.

References

[1] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *New England Journal of Medicine*, vol. 363, no. 2, pp. 166–176, 2010.

[2] T. A. Manolio, F. S. Collins *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[3] N. Chatterjee, B. Wheeler, J. Sampson, P. Hartge, S. J. Chanock, and J.-H. Park, "Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies," *Nature Genetics*, vol. 45, no. 4, pp. 400–405, 2013.

[4] J. N. Cooke Bailey and R. P. Igo Jr, "Genetic Risk Scores," *Current Protocols in Human Genetics*, vol. 91, no. 1, pp. 1.29.1–1.29.9, 2016.

[5] A. C. J. Janssens, "Validity of Polygenic Risk Scores: Are we measuring what we think we are?" *Human Molecular Genetics*, vol. 28, no. R2, pp. R143–R150, 2019.

[6] 1000 Genomes Project Consortium and others, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[7] J. MacArthur, E. Bowler *et al.*, "The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog)," *Nucleic Acids Research*, vol. 45, no. D1, pp. D896–D901, 2017.

[8] J. Euesden, C. M. Lewis, and P. F. O'Reilly, "PRSice: Polygenic risk score software," *Bioinformatics*, vol. 31, no. 9, pp. 1466–1468, 2015.

[9] F. Privé, H. Aschard, and M. G. Blum, "Efficient implementation of penalized regression for genetic risk prediction," *Genetics*, vol. 212, no. 1, pp. 65–74, 2019.

[10] G. Versmée, L. Versmée, M. Dusenne, N. Jalali, and P. Avillach, "dbgap2x: An R package to explore and extract data from the database of Genotypes and Phenotypes (dbGaP)," *Bioinformatics*, vol. 36, no. 4, pp. 1305–1306, 2020.

[11] C. Bycroft, C. Freeman *et al.*, "The UK biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[12] A. Torkamani, N. E. Wineinger, and E. J. Topol, "The personal and clinical utility of polygenic risk scores," *Nature Reviews Genetics*, vol. 19, no. 9, pp. 581–590, 2018.

[13] S. A. Lambert, G. Abraham, and M. Inouye, "Towards clinical utility of polygenic risk scores," *Human Molecular Genetics*, vol. 28, no. R2, pp. R133–R142, 2019.

[14] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[15] G. Abraham, A. S. Havulinna *et al.*, "Genomic prediction of coronary heart disease," *European Heart Journal*, vol. 37, no. 43, pp. 3267–3278, 2016.

[16] R. S. Rosenson and C. C. Tangney, "Antiatherothrombotic properties of statins: Implications for cardiovascular event reduction," *JAMA*, vol. 279, no. 20, pp. 1643–1650, 1998.

[17] P. Natarajan, R. Young *et al.*, "Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting," *Circulation*, vol. 135, no. 22, pp. 2091–2101, 2017.

[18] J. G. Elmore, "Screening for breast cancer: Strategies and recommendations," *Retrieved from the Up to Date website*, 2019. [Online]. Available: http://www.uptodate.com/contents/screening-for-breast-cancer

[19] N. Pashayan, S. Morris, F. J. Gilbert, and P. D. Pharoah, "Cost-effectiveness and benefit-to-harm ratio of risk-stratified screening for breast cancer: A life-table model," *JAMA Oncology*, vol. 4, no. 11, pp. 1504–1510, 2018.

[20] P. Maas, M. Barrdahl *et al.*, "Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States," *JAMA Oncology*, vol. 2, no. 10, pp. 1295–1302, 2016.

[21] A. Lee, N. Mavaddat *et al.*, "BOADICEA: A comprehensive breast cancer risk prediction model incorporating genetic and non-genetic risk factors," *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, vol. 21, no. 8, pp. 1708–1718, 2019.

[22] R. M. Hoffman, "Screening for prostate cancer," 2019. [Online]. Available: www.uptodate.com/contents/screening-for-prostate-cancer

[23] T. M. Seibert, C. C. Fan *et al.*, "Polygenic hazard score to guide screening for aggressive prostate cancer: Development and validation in large scale cohorts," *BMJ*, vol. 360, 2018.

[24] M. J. Redondo, S. Geyer *et al.*, "A type 1 diabetes genetic risk score predicts progression of islet autoimmunity and development of type 1 diabetes in individuals at risk," *Diabetes Care*, vol. 41, no. 9, pp. 1887–1894, 2018.

[25] J. M. Sosenko, J. P. Krischer *et al.*, "A risk score for type 1 diabetes derived from autoantibody-positive participants in the diabetes prevention trial–type 1," *Diabetes Care*, vol. 31, no. 3, pp. 528–533, 2008.

[26] K. Lall, R. Magi, A. Morris, A. Metspalu, and K. Fischer, "Personalized risk prediction for type 2 diabetes: The potential of genetic risk scores," *Genetics in Medicine*, vol. 19, no. 3, pp. 322–329, 2017.

[27] R. S. Desikan, C. C. Fan *et al.*, "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score," *PLoS Medicine*, vol. 14, no. 3, p. e1002258, 2017.

[28] A. T. Marees, H. de Kluiver *et al.*, "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis," *International Journal of Methods in Psychiatric Research*, vol. 27, no. 2, p. e1608, 2018.

[29] P. G. Bagos, "Genetic model selection in genome-wide association studies: Robust methods and the use of meta-analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 12, no. 3, pp. 285–308, 2013.

[30] D. Thomas, "Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies," *Annual Review of Public Health*, vol. 31, no. 1, pp. 21–36, 2010.

[31] J. H. Moore, "Computational analysis of gene-gene interactions using multifactor dimensionality reduction," *Expert Review of Molecular Diagnostics*, vol. 4, no. 6, pp. 795–803, 2004.

[32] P. W. Wilson, J. B. Meigs, L. Sullivan, C. S. Fox, D. M. Nathan, and S. D'Agostino, R. B., "Prediction of incident diabetes mellitus in middle-aged adults: The framingham offspring study," *Archives Internal Medicine*, vol. 167, no. 10, pp. 1068–1074, 2007.

[33] B. J. Keating, "Advances in risk prediction of type 2 diabetes: Integrating genetic scores with Framingham risk models," *Diabetes*, vol. 64, no. 5, pp. 1495–1497, 2015.

[34] L. Duncan, H. Shen *et al.*, "Analysis of polygenic risk score usage and performance in diverse human populations," *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.

[35] M. Khoury, "Is it time to integrate polygenic risk scores into clinical practice? Let's do the science first and follow the evidence wherever it takes us," *Centers for Disease Control and Prevention*, 2019. [Online]. Available: https://blogs.cdc.gov/genomics/2019/06/03/is-it-time/

[36] S. W. Choi, T. S. H. Mak, and P. O'reilly, "A guide to performing Polygenic Risk Score analyses," *BioRxiv*, 2018.

[37] K. Wetterstrand, "The cost of sequencing a human genome," *National Human Genome Research Institute*, 2019. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

[38] NHGRI, "DNA sequencing fact sheet," *National Human Genome Research Institute*, 2015. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Fact-Sheet.

[39] M. Francisco and C. D. Bustamante, "Polygenic risk scores: A biased prediction?" *Genome Medicine*, vol. 10, no. 1, pp. 1–3, 2018.

[40] V. S. Le, K. T. Tran *et al.*, "A Vietnamese human genetic variation database," *Human Mutation*, vol. 40, no. 10, pp. 1664–1675, 2019.

[41] M. D. Mailman, M. Feolo *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nature Genetics*, vol. 39, no. 10, pp. 1181–1186, 2007.

[42] E. M. Ramos, D. Hoffman *et al.*, "Phenotype–Genotype Integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources," *European Journal of Human Genetics*, vol. 22, no. 1, pp. 144–147, 2014.

[43] S. Purcell, B. Neale *et al.*, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[44] U. Drepper, S. Miller, and D. Madore, "md5sum: Verify compact digital fingerprint of a file (GNU GPL version 3 or later)," *Free Software Foundation*, 2010. [Online]. Available: linux.die.net/man/1/md5sum

[45] R. M. Kuhn, D. Haussler, and W. J. Kent, "The UCSC genome browser and associated tools," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 144–161, 2013.

[46] B. K. Bulik-Sullivan, P.-R. Loh *et al.*, "LD score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nature Genetics*, vol. 47, no. 3, p. 291, 2015.

[47] F. Dudbridge, "Power and predictive accuracy of polygenic risk scores," *PLoS Genetics*, vol. 9, no. 3, 2013.

[48] T. S. H. Mak, R. M. Porsch, S. W. Choi, X. Zhou, and P. C. Sham, "Polygenic scores via penalized regression on summary statistics," *Genetic Epidemiology*, vol. 41, no. 6, pp. 469–480, 2017.

[49] B. J. Vilhjálmsson, J. Yang *et al.*, "Modeling linkage disequilibrium increases accuracy of polygenic risk scores," *The American Journal of Human Genetics*, vol. 97, no. 4, pp. 576–592, 2015.

[50] A. V. Khera, M. Chaffin *et al.*, "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations," *Nature Genetics*, vol. 50, no. 9, pp. 1219–1224, 2018.

[51] S. H. Lee, M. E. Goddard, N. R. Wray, and P. M. Visscher, "A better coefficient of determination for genetic profile analysis," *Genetic Epidemiology*, vol. 36, no. 3, pp. 214–224, 2012.

[52] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[53] P. M. Ridker, J. G. MacFadyen *et al.*, "Rosuvastatin for primary prevention among individuals with elevated high-sensitivity C-reactive protein and 5% to 10% and 10% to

20% 10-year risk," *Circulation: Cardiovascular Quality and Outcomes*, vol. 3, no. 5, pp. 447–452, 2010.
[54] P. C. Gøtzsche and O. Olsen, "Is screening for breast cancer with mammography justifiable?" *The Lancet*, vol. 355, no. 9198, pp. 129–134, January 2000.
[55] G. A. Colditz and B. Rosner, "Cumulative risk of breast cancer to age 70 years according to risk factor status: Data from the Nurses' Health Study," *American Journal of Epidemiology*, vol. 152, no. 10, pp. 950–964, 2000.
[56] B. A. Goldstein, L. Yang, E. Salfati, and T. L. Assimes, "Contemporary considerations for constructing a genetic risk score: An empirical approach," *Genetic Epidemiology*, vol. 39, no. 6, pp. 439–445, 2015.
[57] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: A multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

Nguyen Tran The Hung received his doctor of medicine degree from Universities of Medicine and Pharmacy of Ho Chi Minh city (Viet Nam) in 2016. He then got a master degree in biomedical science from China Medical Universities (Taichung, Taiwan) in 2019. His research field is human genetic and diabetes mellitus. He worked briefly as a pediatrician before pursuing his career in academia as a research scientist at Vingroup Big Data Institute from 2019 until now. His thesis on type 2 diabetic nephropathy and the application of polygenic risk score made him believe in the potential impact that genetic research can make in healthcare.

Le Duc Hau obtained his PhD degree in Bioinformatics from University of Ulsan, Republic of Korea in 2012. He is now leading the Department of Computational Biomedicine, Vingroup Big Data Institute, Vietnam. He has been focusing on proposing computational methods for disease- and drug-related problems in personalized medicine, especially on identification of disease-associated biomarkers, prediction of drug targets and response. In parallel, he has been developed bioinformatics tools. So far, he has been published more than fifty papers in well-recognized journals and conferences, nearly a half of those are in ISI-indexed journals. In addition, he has been a member of program committees and reviewer of several international conferences/journals. Moreover, he is a principal investigator and a key member of some national/ministry-level projects. Specially, he is the principal investigator of the biggest genome project in Vietnam (i.e., building databases of genomic variants for Vietnamese population). Finally, he has been collaborating with some well-recognized international research institutes.