A review of methods and software for polygenic risk score analysis

Sara Benoumhani[1], Areej Al-Wabil[1,2], Niddal Imam[3],
Bashayer Alfawaz[1], Amaan Zubairi[1], Dalal Aldossary[1] and
Mariam AlEissa[1,4,5,6,7]

[1] Artificial Intelligence Research Center, Alfaisal University, Riyadh, Saudi Arabia
[2] Software Engineering Department, Alfaisal University, Riyadh, Saudi Arabia
[3] College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia
[4] Molecular Genetics Laboratory, Public Health Authority, Riyadh, Saudi Arabia
[5] College of Medicine, Alfaisal University, Riyadh, Saudi Arabia
[6] King Khaled Eye Specialist Hospital (KKESH) Research Centre, Riyadh, Saudi Arabia
[7] Computational Sciences Department at the Centre for Genomic Medicine (CGM), King Faisal
Specialist Hospital and Research Center, Riyadh, Saudi Arabia

ABSTRACT

Polygenic risk scores (PRSs) are emerging as powerful tools for predicting individual
susceptibility to various diseases and traits based on genetic variants. These scores
integrate information from multiple genetic markers associated with the trait or
disease of interest, offering personalized risk assessment and enhancing disease
management strategies. PRS is an active area of research and is being studied in
various fields, such as disease prediction. This review explores the advancement of
PRS research, focusing on methodological approaches, software tools, and
applications across diverse disciplines. A systematic literature review identified 40
relevant articles classified based on PRS methods and software. Key methods for PRS
computation, including penalized regression and threshold-based approaches,
Bayesian approaches, and machine learning approaches, are discussed, along with
notable software and their features. Applications of PRS in disease prevention are
highlighted. Challenges and future directions, such as increasing diversity in genetic
data, integrating environmental factors, and evaluating clinical implications, are also
discussed to guide future research and implementation efforts.

INTRODUCTION

The ability to predict complicated traits and illnesses, such as cancer from an individual's
genetic variations is important for effective illness prevention (*Wray et al., 2013*;
Chatterjee, Shi & García-Closas, 2016; *Yang et al., 2017*; *Muñoz et al., 2016*; *Wang et al.,
2017*; *Visscher et al., 2017*). Polygenic risk scores (PRSs) are calculated by the effect sizes of
multiple genetic variants known to be associated with the disease or trait of interest.
Researchers are improving risk prediction for common diseases using genetic data. Risk
scores that incorporate both clinical risk indicators and PRSs for a specific illness would
significantly improve the accuracy of lifetime risk prediction and the intuitiveness of

disease risk management (*Slunecka et al., 2021*). Additionally, PRS's research has been expanded to include many diseases using many methods, such as machine learning (ML) today. The PRS has the potential to predict individual disease risks and potentially offer a more effective predictor with improved discrimination properties compared to one based solely on established markers (*Dudbridge, 2013*). Over the last fifteen years, the escalating presence of PRS research groups, the proliferation of peer-reviewed journals, and the surge in conference abstracts all serve as indicators of the rapidly expanding interest in this field. Researchers have been investigating PRS to understand disease risk, predict outcomes, and potentially inform clinical decisions. These studies received a high number of citations in recorded time (*Dudbridge, 2013*; *Lewis & Vassos, 2020*; *Mavaddat et al., 2019*). Moreover, several companies have joined forces with research groups to advance PRS-related technologies, delineating clear roadmaps for their development (*Slunecka et al., 2021*). This remarkable growth in PRS's research is closely tied to an influx of researchers from diverse disciplines, which fostering an interdisciplinary approach that has led to the creation of PRS systems tailored for various target applications. Since 2018, there has been a growing interest in using PRSs to predict the risk of developing multiple diseases; numerous research studies have demonstrated that PRSs are capable of predicting disease status (*Mavaddat et al., 2019*; *Wray et al., 2018*; *Khera et al., 2018*). Researchers have also been exploring ways to improve the accuracy of PRSs by incorporating additional data, such as environmental factors (*Musliner et al., 2019*; *Lewis & Hagenaars, 2019*). PRSs have been used in various applications such as predicting disease risk (*Haas et al., 2018*), patient stratification (*Mavaddat et al., 2019*), investigating treatment response (*Shi et al., 2020*; *Mega et al., 2015*; *Natarajan et al., 2017*) and experimental perturbation informed by genetics (*Dobrindt et al., 2021*; *Hoekstra et al., 2017*). Most prominent PRS techniques, including those integrating functional annotation (*Márquez-Luna et al., 2021*; *Hu et al., 2017*), are based on the classical polygenic disease model. Recently, there has been a marked rise in the volume of studies, investigations, and articles centered on PRS tools. The diversity in research methodologies employed across these studies has yielded a wide spectrum of outcomes, influenced by numerous variables, such as the dataset's methods of calculation of the PRSs.

This study aims to grasp PRS software trends and examine previous studies to equip researchers with knowledge for forthcoming PRS software advancements. In this review, we found that PRS publications span a range of fields such as genetics, epidemiology, computer science, biostatistics, and mathematics. This diversity presents a challenge for comparative analysis due to the wide array of research focuses and methodologies across different journals and scientific areas.

The primary goal of this review is to evaluate the methods, including tools and software. We aim to establish a conceptual structure for the categorization of PRS-related studies, which will aid in the systematic review of PRS research literature. The subsequent sections will detail the proposed framework for categorizing PRS research. Initially, we will define the research approach. Subsequently, we will expound on the suggested categorization

framework for PRS research reviews. The findings are presented, offering insights for forthcoming research and deliberate the trends and challenges in PRS prediction tools. We conclude by summarizing the review's contributions to the body of knowledge in PRS prediction.

METHODS AND MATERIALS

We conducted a systematic review of techniques by the PRISMA guidelines. The subsequent sections detail the methods for article extraction, including the criteria for article selection and the filtering methodologies employed.

### Data sources and procedures for the extraction of articles

Articles concerning PRS can be found dispersed throughout various academic journals that span multiple disciplines. We performed an initial search using online databases such as Web of Science, PubMed, Google Scholar, and Scopus. We used the Publish or Perish software to obtain an extensive bibliography of the academic literature on PRS. This tool collects and analyzes citation data from multiple sources, including Google Scholar, Microsoft Academic Search, PubMed, Scopus, ScienceDirect (Elsevier), ACM Digital Library, Springer Link, IEEE/IEE Library, and Francis. It provides various citation metrics, such as article counts, and total citations (*Harzing, 2010*).

Based on the plot of the number of publications on PRS topics in Fig. 1, there has been a steady growth in related PRS publications since 2013. Therefore, the first search focused on the time frame between 2013 and 2023, utilizing fundamental search parameters, including the phrases and search terms, such as "polygenic risk score" or "polygenic risk scores tool", "predictive polygenic risk score", and "polygenic risk score software". Most of the studies are found in the databases of PubMed, Google Scholar, Semantic Scholar, arXiv, ScienceDirect, and IEEE Xplore. The subsequent section demonstrates the criteria we have chosen.

Selection criteria

Three criteria were established for the inclusion and subsequent analysis of PRS articles. Any articles failing to adhere to these criteria were omitted:

- The review ensured articles that cover methods, or software for generating polygenic risk scores.
- Articles must be relatively current. In this regard, we chose articles that were published between 2013 and 2023. This 10-year period could correspond to the main research period of interest for the PRS topics. Articles are required to be of recent publication. Consequently, we selected articles released within the timeframe of 2013 to 2023. This decade may align with the principal era of research significance for PRS subjects.
- Exclusion of book chapters, meeting abstracts, conference proceedings, workshop descriptions, non-English articles, and master and doctoral dissertations.
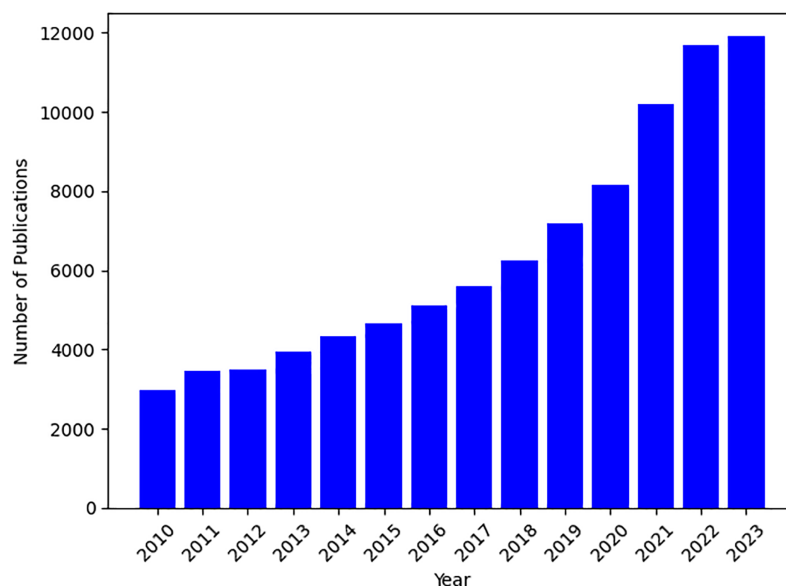
Figure 1 Temporal trends in PRS publications from 2010 to 2023.

Full-size 🖼 DOI: 10.7717/peerj-cs.3039/fig-1
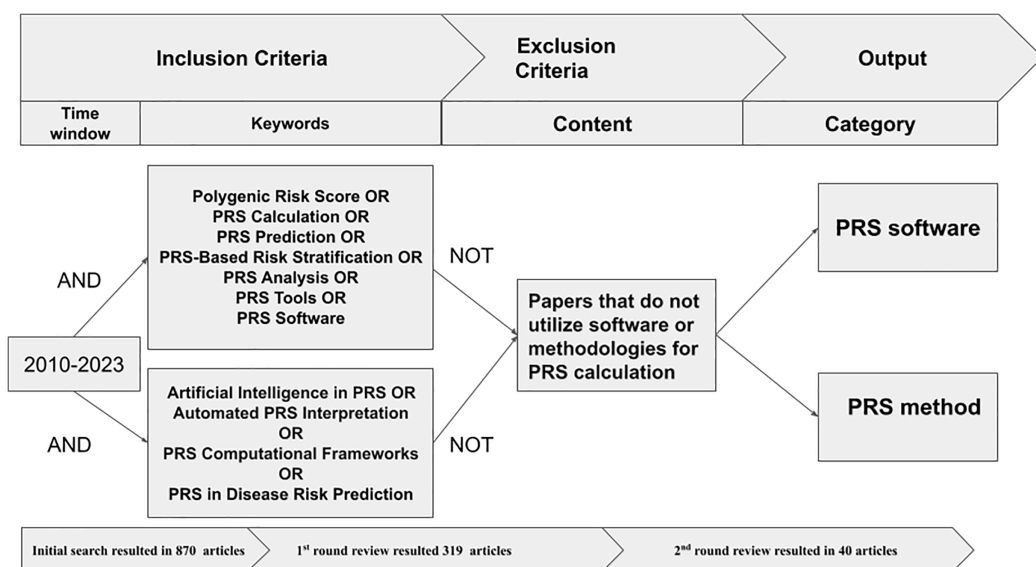
Figure 2 Procedure to extract and filter articles. Full-size 🖼 DOI: 10.7717/peerj-cs.3039/fig-2

Filtering/reviewing process

The goal is to find articles that focus on the software and methods for generating PRS. We manually screened each article in three rounds and classified them.

We initially had 870 articles that matched the criteria. However, some of them were duplicates from different databases. We eliminated 91 duplicate articles and proceeded to the manual screening rounds. We only kept the articles that discussed the methods, tools,

or software for generating PRS. We then sorted them into categories. We conducted the review as following:

- **First round:** We reviewed the titles, abstracts, keywords, and conclusions of each article and discarded those that did not meet the selection criteria. This left us with 319 articles for the next round of review.
- **Second round:** The full texts of the remaining articles were reviewed to ensure they met the criteria, narrowing the selection to 40 articles for the final round. In this round, we conducted an in-depth analysis of each article, focusing on the main theme, and journal rank. Ultimately, we selected and analyzed the most relevant articles. Figure 2 shows the process of filtering and extracting academic articles from the initial search results.

CLASSIFICATION METHOD

We categorized the literature on PRS by their research topics, selecting and filtering 40 articles. These were divided into two main groups: PRS software and PRS methods. The PRS methods were further subdivided into four categories: threshold-based methods, penalized regression methods, Bayesian methods, and machine learning methods. The PRS software was classified into three categories: Command-line, Web Application, and Library. Some methods have dedicated software, which are categorized under both PRS methods and PRS software.

METHODS TO GENERATE PRS

The basic stepwise process for calculating PRS adds up the effects of many genetic variants that are linked to the trait or disease (*Chung, 2021*). Each variation has a weight that shows how much it influences the trait or disease. The formula to calculate the PRS for a person is:

$$\text{PRS}_j = \sum_{i=1}^{N} \beta_i * \text{dosage}_{ij}$$

where N represents the count of SNPs in the score, β_i is the effect size (or beta) of variant i and dosage refers to the number of copies of SNP i present in the genotype of individual j (*Chung, 2021*).

The coefficients are usually derived from a large study that compares the genomes of people with and without the trait or disease (*Collister, Liu & Clifton, 2022*). Only the variants that have a meaningful effect on the trait or disease are selected for the score based on a statistical test (p-value cutoff). The main well-known methods found in the articles are cited in the following sub-sections.

Threshold-based methods
C+T
The Clumping and Thresholding (C+T) method is a widely used approach for calculating PRS. It identifies genome-wide significant variants and groups them based on linkage disequilibrium (LD), excluding those in strong LD with an index variant that has the

lowest p-value in each group, this process helps to identify independent genetic variants associated with a trait (*Wray, Goddard & Visscher, 2007*; *Euesden, Lewis & O'reilly, 2015*). The method operates on the premise that only a few single nucleotide polymorphisms (SNPs) have non-zero effects on the trait. Genetic variants are first clumped based on LD and then filtered based on their p-values to derive polygenic scores (*Kim et al., 2023*; *Mak et al., 2017*).

Breast cancer polygenic risk scores for non-European populations, underrepresented in genetics studies, were developed using the C+T method (*Ho et al., 2022*). It was also used to compare different polygenic profiling methods for Alzheimer's disease risk (*Leonenko et al., 2021*). This method was utilized to assess the race-specific susceptibility of SNPs to AS in the Taiwanese population, as well as to examine the connection between SNPs associated and HLA-B27 with ankylosing spondylitis (AS) susceptibility (*Ko et al., 2022*). A meta-analysis used this method to determine the influence of PRS on the risk of coronary artery disease (*Agbaedeng et al., 2021*). A PRS for autoimmune Addison's disease was constructed and evaluated using the C+T method (*Aranda-Guillé et al., 2023*). Using the UK Biobank dataset, polygenic risk scores for elevated intraocular pressure, a risk factor for glaucoma, were constructed with the C+T method as described in *Gao, Huang & Kim (2019)*.

SCT

The Stacked Clumping and Thresholding (SCT) is an extension of C+T that allows more flexibility in selecting SNPs based on four criteria: p-value threshold, LD window size, LD correlation threshold, and imputation accuracy (*Privé et al., 2019*). SCT generates PRSs for different settings of these criteria and then selects the optimal ones using a penalized regression approach on the validation data (*Privé et al., 2019*).

Penalized regression methods
LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a method used in regression analysis that performs both variable selection and regularization. The goal of LASSO is to achieve the smallest possible sum of squared errors, with the condition that the total absolute value of the coefficients does not exceed a predetermined threshold (*Ribbing et al., 2007*). It is used in machine learning and statistics to select variables in a model by shrinking some of the coefficients to zero. This method helps to prevent overfitting by decreasing the number of variables included in the model. LASSO can be used for PRS development. In this context, LASSO is used as a variable selection technique to select the most important genetic variants for inclusion in the PRSs. Addressing the insufficient representation of non-European communities in genetics studies to develop breast cancer polygenic risk scores utilizing the LASSO method as indicated in *Ho et al. (2022)*.

Lassosum

Similar to LASSO, and it applies a LASSO penalty to nullify the effect sizes of genetic variants. It further prunes genetic variants exhibiting linkage disequilibrium and applies a

threshold to the remaining variants based on their *p*-values (*Mak et al., 2017*). The Lassosum approach was reported in several PRS studies. An assessment was made on the race-specific susceptibility of single nucleotide polymorphisms to ankylosing spondylitis (AS) in the Taiwanese population. The association between human leukocyte antigen (HLA)-B27 and susceptibility SNPs for AS in Taiwan was explored. Polygenic risk scores were used to analyze genetic variations in predicting the development of AS using LassoSum as indicated in *Ko et al. (2022)*. A meta-analysis investigated how PRSs affect the likelihood of developing coronary artery disease by employing the Lassosum method (*Agbaedeng et al., 2021*). A reference-standardized framework assessed the predictive value of several polygenic risk score methodologies, including Lassosum (*Pain et al., 2021*).

SBLUP

The Super Genomic Best Linear Unbiased Prediction (SBLUP) technique adjusts SNP effect magnitudes by utilizing an external LD reference panel. This process transforms the ordinary least squares estimates of SNP into nearly optimal linear unbiased predictions (*Ren et al., 2021*). The SBLUP utilizes a Bayesian framework to calculate the magnitude of genetic variants effects. This method presumes that the distribution of these effects is normal, centering around zero, with their variance inversely proportional to the number of variants in the score. The SBLUP technique provides greater precision in the construction of PRS compared to other methods (*Slunecka et al., 2021*; *Robinson et al., 2017*).

DBSLMM

The Deterministic Bayesian Sparse Linear Mixed Model (DBSLMM) is a technique used to compute polygenic risk scores. This method utilizes a versatile approach to modeling the distribution of effect sizes. This allows for strong and precise predictions over various genetic structures. Additionally, DBSLMM employs a straightforward deterministic search method to produce an estimated analytical solution based solely on summary statistics. Through simulation tests, DBSLMM has demonstrated its ability to provide scalable and precise predictions for a wide array of genuine genetic structures (*Yang & Zhou, 2020*).

Bayesian methods

LDpred

The linkage disequilibrium pred (LDprep) method is widely utilized for the calculation of PRS. It operates by using summary statistics alongside a matrix that measures the correlation among genetic variants. It's a Bayesian approach that accounts for LD among genetic variants, assuming that each variant independently affects the trait. LDpred is a two-step method: it first estimates the LD structure from a reference panel and then uses this structure to adjust Genome-373 Wide Association Study (GWAS) summary statistics for the effects of LD. This method requires defining a tuning parameter (ρ), which is an estimate of the genetic variants assumed to be causal (*Imam, Noguera & Donohue, 2014*; *Vilhjalmsson et al., 2015*).

An evaluation was conducted on the race-specific SNP susceptibility of AS in Taiwanese people, as well as the relationship between HLA-B27 and AS susceptibility SNPs in Taiwan. A PRS technique was also used to examine genetic variations in predicting the

development of AS using LDpred (*Ko et al., 2022*). A meta-analysis investigated how PRSs affect the likelihood of developing coronary artery disease, utilizing the method described by *Agbaedeng et al. (2021)*. Additionally, LDpred was utilized to construct PRS to analyze the contribution of common genetic variations to suicide attempts. The aim was to demonstrate the genetic overlap and correlation between measures of suicide attempts and to explore the genetic associations of suicide attempts with other traits, such as insomnia and psychiatric disorders (*Ruderfer et al., 2020*).

JAMPred

The Joint Analysis of Marginal Summary Statistics Prediction (JAMPred) is a technique for computing polygenic risk scores based on summary data from GWAS and a reference genotyping panel (*Newcombe et al., 2019*). JAMPred considers linkage disequilibrium among genetic variations and uses a Bayesian framework to estimate impact sizes and posterior probability of inclusion for each variant, furthermore, JAMPred employs variable selection and model averaging techniques to enhance the accuracy and stability of polygenic risk scores (*Newcombe et al., 2019*). An example of how PRS was utilized as a predictive tool for identifying high-risk patients with Parkinson's disease. Various methods, including JAMPred as described in *Shan et al. (2021)*.

SBayesR

The SBayesR method is a Bayesian approach, which often used to compute polygenic risk scores, It incorporates a spike-and-slab technique to model the effects of genetic variants on the phenotype of interest (*Pham et al., 2022*). Efforts have been made to elucidate the most effective methodologies for polygenic profiling when screening individuals for Alzheimer's disease risk. Various methods, including the SBayesR approach, are employed utilizing datasets sourced from prominent institutions such as the UK Biobank and National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) (*Leonenko et al., 2021*).

A reference-standardized framework was applied to assess the predictive value of several PRS methods. SBayesR was among the best methods found to assess the predictive value of several PRS as indicated in *Pain et al. (2021)*.

EB-PRS

The emperical Bayes polygenic risk score (EB-PRS) is a method to generate PRS from summary statistics of GWAS and employs a statistical approach known as empirical Bayes to estimate the impact sizes of genetic markers throughout the whole genome, EB-PRS does not need parameter tuning or the use of external data (*Song et al., 2020*). The EB-PRS method has proven to be effective in producing outstanding outcomes independently, without the need for parameter tuning or external datasets. However, research indicates that its performance can be enhanced further when a reference panel is utilized (*Adam et al., 2022*).

BridgePRS

The BridgePRS technique is a Bayesian polygenic risk score strategy that combines the PRS of two populations with differing ancestry. The goal is to address the PRS Portability

Problem by utilizing common genetic effects across ancestries to improve PRS accuracy in non-European communities. In other words, it seeks to increase the accuracy of PRS estimations for people from different and underrepresented heritage groups (*Hoggart et al., 2023*).

Machine learning methods

The computation of PRS often relies on simple linear models which might not fully encompass the intricate interdependencies involved between phenotypes and genotypes (*DeWan, 2018*; *Aschard, 2016*). Therefore, machine learning methods that can account for non-linearities and interactions among genetic variants are of interest for improving the accuracy and interpretability of PRS.

A proposed approach combines an ensemble method for selecting SNPs with Gradient-Boosted Trees (GBT) to account for the non-linear and interactive influences of SNPs on phenotypes. When a PRS is included as a feature within an extreme gradient boosting model, there is a notable enhancement in the explained variance percentage relative to the conventional linear PRS model, as observed across nine complex phenotypes within a diverse ancestral group from the UK Biobank (*Elgart et al., 2022*).

The utilization of machine learning and deep learning techniques was thoroughly explored to compute polygenic risk scores from GWAS data. Random forest (RF) and support vector machines (SVM), and deep learning methods were employed to calculate weight vectors, which play a pivotal role in PRS computation (*Öztornaci et al., 2023*). Additionally, variable importance measurements obtained from the RF method serve as weight vectors. In all these methods, individual risk scores are derived by multiplying each SNP with its corresponding weight vector.

Peng et al. (2024) introduced a DL framework that captures intricate genetic interactions beyond additive effects. In contrast to traditional PRS models, which often assume linear relationships, DeepRisk leverages neural networks to model non-linear associations among single-nucleotide polymorphisms (SNPs). The approach described in *Peng et al. (2024)* has demonstrated superior performance in predicting disease risk, particularly in scenarios involving complex genetic architectures. The study (*Zhou et al., 2023*) introduced a neural network model that captures non-linear interactions among genetic variants, offering enhanced predictive accuracy and deeper insights into disease mechanisms. The PRS-Net, as described in *Li et al. (2024)*, demonstrated that incorporating a lightweight geometric layer into gene-level PRSs yields reproducible and biologically interpretable improvements over both linear and black-box non-linear baselines, particularly for immune-mediated diseases and heterogeneous ancestries.

RESULTS

We extracted a collection of articles on polygenic risk scores (PRS) from online databases. Each article was thoroughly reviewed and classified according to the established categorization method. Figure 3 presents the PRISMA flow diagram outlining the selection process.
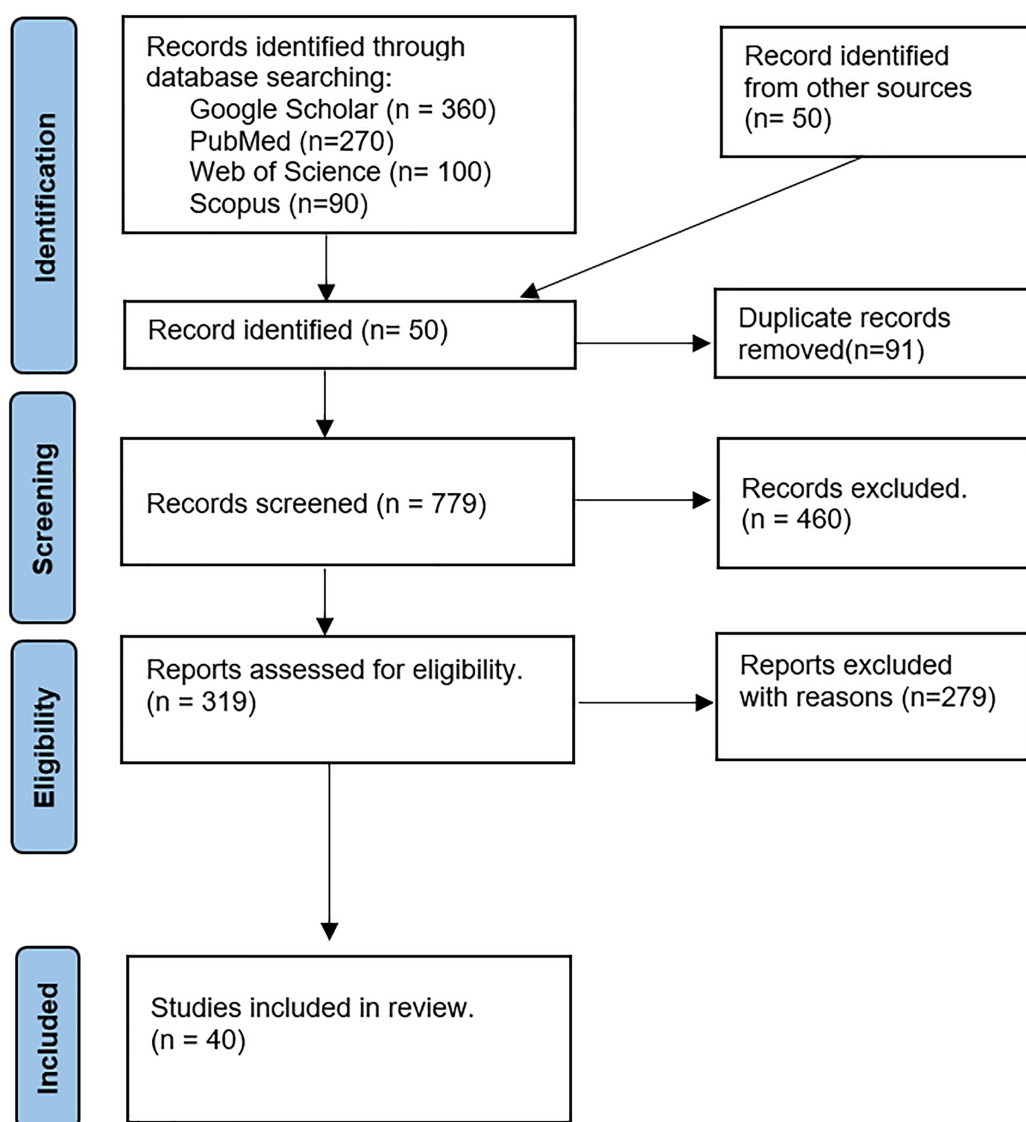
Figure 3 The PRISMA flow diagram-based study selection for the review.
Full-size DOI: 10.7717/peerj-cs.3039/fig-3

Table 1 highlights a range of software tools commonly utilized in PRS analysis, grouped into three main types: command-line tools, web-based applications, and programming libraries. The table also outlines the diverse analytical methods employed in PRS studies.

DISCUSSION

The benefits and limitations of the techniques used
Threshold-based methods

These methods use a predefined threshold to identify variants by utilizing their p-values or the magnitude of their effects derived from GWAS summary data. Research has suggested that threshold-based methods are simple and relatively easy to implement, and widely used

Table 1 Summary of PRS software and methods.

Category	Software/ method	Programming language	Availability	Description	Ref
Command-line	Plink2	C++	Free	GWAS analysis and research in population genetics	*Chang et al. (2015)*
	PRSice	C++, R	Free	Computing, implementing, assessing, and graphically representing PRS results with R	*Euesden, Lewis & O'reilly (2015)*
	PRSice2	C++, R	Free	Automating and simplifying the analysis of PRS on extensive datasets	*Choi & O'Reilly (2019)*
	PRS-CS	Python	Free	Infers posterior SNP effects using continuous shrinkage priors and LD panels	*Ge et al. (2019)*
	PRS-on-Spark (PRSoS)	Python, Spark	Free	Computes PRS handling various inputs and ambiguous SNPs	*Chen et al. (2018)*
	EraSOR	Python	Free	Eliminates bias from overlapping samples in GWAS/PRS data	*Choi et al. (2023)*
	BridgePRS	R, Python	Free	Bridges PRS across populations to address portability issues	*Hoggart et al. (2023)*
	AnnoPred	Python	Free	Predicts disease risk integrating GWAS statistics and annotations	*Hu et al. (2017)*
Web application	Cancer PRSweb	–	Free	Online repository hosting PRS for major cancer traits	*Fritsche et al. (2020)*
	CanRisk	–	Free	Estimates breast/ovarian cancer risks and mutation probabilities	*Carver et al. (2021)*
Library	bigsnpr	R	Free	Calculates PRS using GWAS statistics; supports LDpred2	*Privé, Arbel & Vilhjálmsson (2020)*
	EB-PRS	R	Free	Uses effect size distribution without tuning or external data	*Song et al. (2020)*
	Lassosum	R	Free	Penalized regression on GWAS summary statistics *via* Lasso	*Mak et al. (2017)*
	PolyFun	Python	Free	Fine-mapping and prediction including PolyFun, PolyLoc, PolyPred	*Weissbrod et al. (2022)*
	XPXP	Python	Free	Enhances PRS prediction using cross-population/phenotype analysis	*Xiao et al. (2022)*
	LDpred	Python	Free	Estimates posterior effect sizes using LD and prior models	*Vilhjalmsson et al. (2015)*
Threshold-based	C+T	–	–	Selects SNPs by p-values and LD, sums risk alleles	*Ho et al. (2022)*, *Leonenko et al. (2021)*, *Ko et al. (2022)*, *Agbaedeng et al. (2021)*, *Aranda-Guillé et al. (2023)*, *Gao, Huang & Kim (2019)*
	SCT	–	–	Combines multiple C+T scores with stacking classifier	*Privé et al. (2019)*
Penalized regression	Lasso	R (Lassosum)	–	Selects SNPs and estimates effects *via* Lasso	*Ho et al. (2022)*

(Continued)

Table 1 (continued)					
Category	**Software/ method**	**Programming language**	**Availability**	**Description**	**Ref**
	Lassosum	R	–	Lasso penalty on GWAS statistics; handles overfitting	*Ko et al., (2022), Agbaedeng et al., (2021), Pain et al. (2021)*
	SBLUP	–	–	Bayesian method accounting for LD and SNP interactions	*Robinson et al. (2017), Slunecka et al. (2021)*
	DBSLMM	–	–	Sparse mixed model approximation using heritability tuning	*Yang & Zhou, (2020)*
Bayesian	LDpred	Python	Free	Gibbs sampling with LD to estimate PRS	*Imam, Noguera & Donohue (2014), Vilhjalmsson et al. (2015), Ko et al. (2022), Agbaedeng et al. (2021)*
	Jampred	–	–	Joint analysis across GWAS to improve power and accuracy	*Newcombe et al. (2019), Shan et al. (2021)*
	EB-PRS	R	Free	Uses effect size distribution without external info	*Song et al. (2020)*
	SBayesR	–	–	Bayesian regression with spike-and-slab prior	*Pham et al. (2022), Leonenko et al. (2021), Pain et al. (2021)*
	BridgePRS	R, Python	Free	Combines Bayesian PRS from distinct ancestries	*Hoggart et al. (2023)*
Machine learning	RF	–	–	Uses random forest to weight SNPs for PRS	*Öztornaci et al. (2023)*
	SVM	–	–	Uses support vector machines to improve classification	*Öztornaci et al. (2023)*
	GBT	–	–	Gradient boosting trees + XGBoost for non-linear SNP effects	*Elgart et al. (2022)*

in practice (*Privé et al., 2019*; *Lewis & Vassos, 2020*). They can provide a straight forward interpretation of the PRS as a weighted sum of selected variants. However, research has shown that threshold-based methods may ignore variants with small effects that can collectively contribute to the PRS, they are sensitive to the choice of threshold, which can affect the predictive performance and the number of variants included (*Lewis & Vassos, 2020*).

Penalized regression methods

These methods use a regularization term to penalize the coefficients of the PRS, which can handle high-dimensional data effectively, they have been shown to reduce overfitting by shrinking coefficients, especially for correlated or weakly associated variants (*Privé, Aschard & Blum, 2019*). They also provide variable selection by shrinking less informative variables towards zero, which can improve interpretability and reduce collinearity. However, the choice of penalty parameters can be subjective and depend on the data and the trait (*Pattee & Pan, 2020*). The interpretation of the resulting coefficients may be challenging, especially when using non-linear penalties or complex models, they also assume a linear relationship between predictors and outcomes, which may not hold for some traits or diseases (*Pattee & Pan, 2020*).

Bayesian techniques

These methods use a probabilistic framework to estimate the posterior distribution of the PRS given the prior information and the data. They can incorporate various sources of information such as functional annotations and biological pathways. Also they account for uncertainty and provide credible intervals for the PRS (*Ge et al., 2019*; *Zhou, Qie & Zhao, 2023*). They have the flexibility to incorporate prior knowledge and choose the prior distribution (*Ge et al., 2019*). However, they can be computationally intensive, they may also require expertise in Bayesian statistics for proper implementation and interpretation of results (*Song et al., 2020*).

Machine learning methods

These methods use various algorithms to learn the optimal PRS from the data, such as random forests, support vector machines, neural networks, *etc* (*Öztornaci et al., 2023*). Studies have shown that they can capture complex and nonlinear relationships between variants and outcomes, which linear models may not capture, they optimize predictive performance by using cross-validation, grid search, or other techniques to tune the hyperparameters, and have flexibility in feature engineering, such as using interactions, transformations, or embeddings of variants (*Squires, Weedon & Oram, 2023*; *Mamani, 2020*). However, they may overfit the data if not properly regularized, which can reduce generalizability and robustness. Also, they can be challenging to interpret, particularly in the case of intricate models such as neural networks, which are often perceived as black boxes. They have been shown to have high computational complexity, which can limit their scalability and applicability (*Mamani, 2020*).

The applications of PRS

Polygenic risk scores are a versatile tool for healthcare applications (*Slunecka et al., 2021*). PRS can estimate the probability of an individual having or developing a specific disease or trait, facilitating risk stratification and early intervention that consider both genetic and environmental factors (*Lewis & Green, 2021*; *Corpas & Fatumo, 2023*). These interventions can range from lifestyle changes to preventive surgeries, depending on the condition and the individual's preferences (*Lewis & Green, 2021*). PRS can provide healthcare providers with valuable information for risk assessment, disease management, and preventive care, and help them deliver tailored advice and interventions to their patients (*Chapman, 2023*). Additionally, PRS can inform the design of screening programs and research studies, by dividing populations into different risk groups and adjusting screening criteria and intervention effects accordingly (*Slunecka et al., 2021*). However, the implementation of PRS-based approaches in healthcare requires careful attention to ethical, social, and practical issues, to ensure fair and respectful practices that protect patient autonomy and privacy while maximizing benefits for individuals and society (*Chapman, 2023*; *Aragam & Natarajan, 2020*; *Lewis & Vassos, 2020*).

Challenges and future directions

Increasing diversity and representation of data

PRS methods predominantly rely on data from individuals of European ancestry, leading to limitations in their applicability and generalizability to other ethnic groups. This lack of diversity can result in biased risk predictions and may exacerbate existing health disparities. There is a growing recognition of the importance of including diverse and representative genetic data from different populations. Initiatives are underway to collect and analyze genomic information from historically underrepresented populations, such as the Human Heredity and Health Africa (H3Africa) initiative. Developing methods that account for genetic diversity and mixed ancestry within varied groups is essential for improving the precision and reliability of PRS across all ethnic backgrounds (*Zhang et al., 2023*; *Lam et al., 2019*).

To enhance multi-ancestry prediction models, it is necessary to leverage genetic data from diverse populations to improve PRS performance. For example, the PROSPER method demonstrates improved precision and reliability showing a 70% increase in accuracy for individuals of African ancestry compared to traditional models (*Zhang et al., 2024*). Such approaches help correct biases inherent in European-centric models by accounting for differences in genetic architecture, including linkage disequilibrium patterns and allele frequencies (*Cavazos & Witte, 2021*). Incorporating ancestry-specific data into PRS algorithms is essential for improving their applicability across diverse populations (*Lerga-Jaso et al., 2024*).

A major driver of improved accuracy and generalizability in PRS development is the increasing availability of large-scale biobank databases. Due to the complexity of the human genome, large datasets are critical for identifying associations between genetic variants and complex traits (*Raben et al., 2023*). Biobank resources support the development, validation, and application of PRS by providing extensive training data and, crucially, multi-ancestry samples for cross-population evaluation (*Tsuo et al., 2024*; *Thompson et al., 2022*). Major global initiatives including the UK Biobank, All of Us (AoU), China Kadoorie Biobank, Biobank Japan, deCODE Genetics, the Estonian Biobank, and Lifelines in the Netherlands are helping to address the historical underrepresentation of non-European populations in genetic research (*Ju et al., 2022*). These resources provide the statistical power needed to reduce false positives, identify novel variants, and refine estimates of single nucleotide polymorphism (SNP) effect sizes (*Raben et al., 2023*; *Ju et al., 2022*). With advancements in analytic tools and machine learning algorithms, biobank databases are making PRS construction increasingly accessible to a broader range of researchers (*Sakaue et al., 2020*; *Du et al., 2023*).

Using explainable AI for improving the interpretability of PRS

ML-based methods can be used to learn the optimal PRS from the data. However, the lack of transparency of ML's prediction could lead to a poor generalization on datasets when a model learns to predict on irrelevant features. The explainability of MLs is crucial in healthcare as the consequences of a wrong prediction in diagnostics may cause life-changing decisions for a patient (*Elgart et al., 2022*). Recently, explainable artificial

intelligence (XAI) has been widely used in the literature to overcome the lack of insight of ML-based models in healthcare and medical diagnosis systems (*Zhang, Weng & Lund, 2022*). XAI reveals the decision patterns of ML-based models, which helps medical practitioners understand the logical reasoning for the model's prediction (*Zhang, Weng & Lund, 2022*). XAI is a promising research direction that requires more attention from the PRS research community.

Incorporating environmental and lifestyle factors

PRS methods typically focus solely on genetic components and may not fully capture the multifactorial nature of complex traits and diseases. However, lifestyle and environmental factors—such as exercise, diet, and exposure to toxins—play significant roles in disease risk but are often overlooked in traditional PRS analyses. Integrating these non-genetic variables into PRS models can enhance their predictive power and relevance. For instance, including data on smoking behavior, socioeconomic status, and geographic location can improve risk stratification and facilitate more personalized interventions. Advanced statistical approaches, such as gene-environment interaction modeling, are being developed to better capture the interplay between genetic and environmental influences on health outcomes (*Wang et al., 2021*; *Koch et al., 2023*).

Recent advances demonstrate the potential of integrating such factors. For example, incorporating 109 exposome variables—including tobacco use, education, and others—into cardiovascular disease risk prediction using machine learning increased the area under the curve (AUC) to 0.82 (*Shahbazi & Nowaczyk, 2025*). Additionally, the use of Internet of Things (IoT) devices enables real-time integration of lifestyle and environmental data such as diet and air quality, supporting adaptive health interventions for hyper-personalized medicine (*Tan et al., 2025*). Moreover, social determinants of health (SDoH) have shown significant associations with disease outcomes, particularly in high-risk environments, further highlighting the importance of integrating socio-environmental context into PRS applications (*Guare et al., 2024*).

Evaluating clinical and public health implications

PRSs show great potential for predicting health risks, but their implementation also raises important ethical, societal, and legal concerns. The integration of PRS into clinical practice requires a thoughtful approach to issues such as informed consent, data privacy, protection, and equitable access to genetic services and interventions. It is essential to rigorously evaluate both the effectiveness and broader implications of PRS across diverse populations and clinical settings. Studies assessing the utility of PRS in guiding preventive strategies, screening programs, and therapeutic decisions are vital for shaping evidence-based healthcare policies.

Public understanding and health literacy around genetic risk are equally important. Enhancing awareness can help ensure informed decision making and reduce the risk of misinterpretation or misuse of genetic information in clinical and personal contexts.

Ethical considerations must be addressed, particularly when individuals are assigned high-risk scores, which may lead to psychological distress or stigmatization. In addition,

disparities in access to genetic testing, risk assessment tools, and personalized interventions can exacerbate existing health inequalities (*Andreoli et al., 2024*). Clear communication of genetic findings and proper implementation supported by robust informed consent processes are essential for integrating PRS into personalized medicine and improving shared decision-making between patients and healthcare providers (*King & Bishop, 2017*).

In summary, addressing these challenges and leveraging the opportunities presented by advancements in genomics, data science, and healthcare delivery will be essential for realizing the complete promise of PRS in enhancing health outcomes and minimizing health inequities among different groups (*Koch et al., 2023*; *Simona et al., 2023*; *Lewis & Vassos, 2017*). The criticality of generating precise PRS for various complex traits and illnesses is paramount. PRS offers a gauge of an individual's genetic susceptibility to a complex trait or illness, indicating the probability of manifesting a specific trait or illness grounded on one's genetic makeup. PRS analysis aims to pinpoint individuals at heightened disease risk by analyzing genetic variations alongside clinical factors. Thus, the more precise the PRS, the more effectively we can pinpoint disease risks and devise preventative measures.

CONCLUSION

Polygenic risk scores hold immense promise for predicting disease. The extensive literature on PRS research reflects the growing interest and investment in this field, with significant advancements in methods, tools, and applications. By integrating genetic information with clinical data, PRSs contribute to predicting disease risk and guiding preventive interventions. However, challenges remain in ensuring data diversity, incorporating environmental factors, and addressing ethical considerations. Future research efforts should focus on overcoming these challenges to unlock the full potential of PRSs in improving clinical outcomes and public health interventions.

LIST OF ABBREVIATIONS

PRSs	Polygenic risk scores
PRS	Polygenic risk score
ML	Machine learning
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SNPs	Single Nucleotide Polymorphisms
C+T	Clumping and Thresholding
LD	Linkage Disequilibrium
AS	Ankylosing Spondylitis
HLA	Human Leukocyte Antigen
SCT	Stacked Clumping and Thresholding
Lasso	Least Absolute Shrinkage Selection Operator
SBLUP	Super Genomic Best Linear Unbiased Prediction
DBSLMM	Deterministic Bayesian Sparse Linear Mixed Model
LDpred	Linkage Disequilibrium pred

GWAS	Genome-373 Wide Association Study
JAMPred	Joint Analysis of Marginal Summary Statistics Prediction
EB-PRS	Empirical Bayes-PRS
SVM	Support vector machines
RF	Random Forests
GBT	Gradient-Boosted Trees
CS	Continuous Shrinkage
H3Africa	Human Heredity and Health in Africa
XAI	Explainable AI

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Sara Benoumhani conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Areej Al-Wabil performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Niddal Imam performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Bashayer Alfawaz performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Amaan Zubairi performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Dalal Aldossary performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Mariam AlEissa conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This is a literature review.

REFERENCES

Adam Y, Sadeeq S, Kumuthini J, Ajayi O, Wells G, Solomon R, Ogunlana O, Adetiba E, Iweala E, Brors B, Adebiyi E. 2022. Polygenic risk score in african populations: progress and challenges. *F1000Research* **11**:175 DOI 10.12688/f1000research.76218.2.

Agbaedeng TA, Noubiap JJ, Mato EPM, Chew DP, Figtree GA, Said MA, Van der Harst P. 2021. Polygenic risk score and coronary artery disease: a meta-analysis of 979,286 participant data. *Atherosclerosis* **333**:48–55 DOI 10.1016/j.atherosclerosis.2021.08.020.

Andreoli L, Peeters H, Van Steen K, Dierickx K. 2024. Taking the risk. a systematic review of ethical reasons and moral arguments in the clinical use of polygenic risk scores. *American Journal of Medical Genetics Part A* **194(7)**:e63584 DOI 10.1002/ajmg.a.63584.

Aragam KG, Natarajan P. 2020. Polygenic scores to assess atherosclerotic cardiovascular disease risk: clinical perspectives and basic implications. *Circulation Research* **126(9)**:1159–1177 DOI 10.1161/circresaha.120.315928.

Aranda-Guillé M, Røyrvik EC, Fletcher-Sandersjöö S, Artaza H, Botusan IR, Grytaas MA, Hallgren AE, Breivik L, Pettersson M, Jørgensen AP, Lindstrand A, Vogt E, Norwegian Addison Registry Study Group, The Swedish Addison Registry Study Group, Husebye ES, Kämpe O, Wolff ASB, Bensing S, Johansson S, Eriksson D. 2023. A polygenic risk score to help discriminate primary adrenal insufficiency of different etiologies. *Journal of Internal Medicine* **294(1)**:96–109 DOI 10.1111/joim.13649.

Aschard H. 2016. A perspective on interaction effects in genetic association studies. *Genetic Epidemiology* **40(8)**:678–688 DOI 10.1002/gepi.21989.

Carver T, Hartley S, Lee A, Cunningham AP, Archer S, Babb de Villiers C, Roberts J, Ruston R, Walter FM, Tischkowitz M, Easton DF, Antoniou AC. 2021. Canrisk tool—a web interface for the prediction of breast and ovarian cancer risk and the likelihood of carrying genetic pathogenic variants. *Cancer Epidemiology, Biomarkers & Prevention* **30(3)**:469–473 DOI 10.1158/1055-9965.EPI-20-1319.

Cavazos TB, Witte JS. 2021. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances* **2(1)**:100017 DOI 10.1016/j.xhgg.2020.100017.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* **4(1)**:s13742–015 DOI 10.1186/s13742-015-0047-8.

Chapman CR. 2023. Ethical, legal, and social implications of genetic risk prediction for multifactorial disease: a narrative review identifying concerns about interpretation and use of polygenic scores. *Journal of Community Genetics* **14(5)**:441–452 DOI 10.1007/s12687-022-00625-9.

Chatterjee N, Shi J, García-Closas M. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17(7)**:392–406 DOI 10.1038/nrg.2016.27.

Chen LM, Yao N, Garg E, Zhu Y, Nguyen TTT, Pokhvisneva I, Dass SAH, Forest M, McEwen LM, MacIsaac JL, Kobor MS, Greenwood C, Silveira PP, Meaney MJ, O'Donnell KJ. 2018.

PRS-on-Spark (PRSoS): a novel, efficient and flexible approach for generating polygenic risk scores. *BMC Bioinformatics* **19(1)**:1–9 DOI 10.1101/209833.

Choi SW, Mak TSH, Hoggart CJ, O'Reilly PF. 2023. EraSOR: a software tool to eliminate inflation caused by sample overlap in polygenic score analyses. *GigaScience* **12**:giad043 DOI 10.1093/gigascience/giad043.

Choi SW, O'Reilly PF. 2019. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8(7)**:giz082 DOI 10.1093/gigascience/giz082.

Chung W. 2021. Statistical models and computational tools for predicting complex traits and diseases. *Genomics & Informatics* **19(4)**:e36 DOI 10.5808/gi.21053.

Collister JA, Liu X, Clifton L. 2022. Calculating polygenic risk scores (PRS) in UK biobank: a practical guide for epidemiologists. *Frontiers in Genetics* **13**:818574 DOI 10.3389/fgene.2022.818574.

Corpas M, Fatumo S. 2023. Generalisation of genomic findings and applications of polygenic risk scores. *BMC Medical Genomics* **16(1)**:175 DOI 10.1186/s12920-023-01615-7.

DeWan AT. 2018. Gene-gene and gene-environment interactions. *Genetic Epidemiology: Methods and Protocols* **1793(11)**:89–110 DOI 10.1007/978-1-4939-7868-7_7.

Dobrindt K, Zhang H, Das D, Abdollahi S, Prorok T, Ghosh S, Weintraub S, Genovese G, Powell SK, Lund A, Akbarian S, Eggan K, McCarroll S, Duan J, Avramopoulos D, Brennand KJ. 2021. Publicly available hipsc lines with extreme polygenic risk scores for modeling schizophrenia. *Complex Psychiatry* **6(3–4)**:68–82 DOI 10.1159/000512716.

Du R-Q, Zhao D-D, Kang K, Wang F, Xu R-X, Chi C-L, Kong L-Y, Liang B. 2023. A review of pre-implantation genetic testing technologies and applications. *Reproductive and Developmental Medicine* **7(1)**:20–31 DOI 10.1097/rd9.0000000000000049.

Dudbridge F. 2013. Power and predictive accuracy of polygenic risk scores. *PLOS Genetics* **9(3)**: e1003348 DOI 10.1371/journal.pgen.1003348.

Elgart M, Lyons G, Romero-Brufau S, Kurniansyah N, Brody JA, Guo X, Lin HJ, Raffield L, Gao Y, Chen H, de Vries P, Lloyd-Jones DM, Lange LA, Peloso GM, Fornage M, Rotter JI, Rich SS, Morrison AC, Psaty BM, Levy D, Redline S, de Vries P, Sofer T. 2022. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications Biology* **5(1)**:856 DOI 10.1038/s42003-022-03812-z.

Euesden J, Lewis CM, O'reilly PF. 2015. Prsice: polygenic risk score software. *Bioinformatics* **31(9)**:1466–1468 DOI 10.1093/bioinformatics/btu848.

Fritsche LG, Patil S, Beesley LJ, VandeHaar P, Salvatore M, Ma Y, Peng RB, Taliun D, Zhou X, Mukherjee B. 2020. Cancer prsweb: an online repository with polygenic risk scores for major cancer traits and their evaluation in two independent biobanks. *The American Journal of Human Genetics* **107(5)**:815–836 DOI 10.1016/j.ajhg.2020.08.025.

Gao XR, Huang H, Kim H. 2019. Polygenic risk score is associated with intraocular pressure and improves glaucoma prediction in the UK biobank cohort. *Translational Vision Science & Technology* **8(2)**:10 DOI 10.1167/tvst.8.2.10.

Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW. 2019. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications* **10(1)**:1776 DOI 10.1038/s41467-019-09718-5.

Guare LA, Das J, Caruth L, Setia-Verma S. 2024. Social determinants of health and lifestyle risk factors modulate genetic susceptibility for women's health outcomes. In: *Biocomputing 2025: Proceedings of the Pacific Symposium.* World Scientific, 296–313.

Haas ME, Aragam KG, Emdin CA, Bick AG, Hemani G, Smith GD, Kathiresan S. 2018. Genetic association of albuminuria with cardiometabolic disease and blood pressure. *The American Journal of Human Genetics* **103**(4):461–473 DOI 10.1016/j.ajhg.2018.08.004.

Harzing A-W. 2010. *The publish or perish book*. Australia: Tarma Software Research Pty Limited Melbourne.

Ho W-K, Tai M-C, Dennis J, Shu X, Li J, Ho PJ, Millwood IY, Lin K, Jee Y-H, Lee S-H, Mavaddat N, Bolla MK, Wang Q, Michailidou K, Long J, Wijaya EA, Hassan T, Rahmat K, Tan VKM, Tan BKT, Tan SM, Tan EY, Lim SH, Gao Y-T, Zheng Y, Kang D, Choi J-Y, Han W, Lee H-B, Kubo M, Okada Y, Namba S, Park SK, Kim S-W, Shen C-Y, Wu P-E, Park B, Muir KR, Lophatananon A, Wu AH, Tseng C-C, Matsuo K, Ito H, Kwong A, Chan TL, John EM, Kurian AW, Iwasaki M, Yamaji T, Kweon S-S, Aronson KJ, Murphy RA, Koh WP, Khor CC, Yuan JM, Dorajoo R, Walters RG, Chen Z, Li L, Lv J, Jung KJ, Kraft P, Pharoah PD, Dunning AM, Simard J, Shu XO, Yip CH, Taib NAM, Antoniou AC, Zheng W, Hartman M, Easton DF, Teo SH. 2022. Polygenic risk scores for prediction of breast cancer risk in Asian populations. *Genetics in Medicine* **24**:586–600 DOI 10.1016/j.gim.2021.11.008.

Hoekstra SD, Stringer S, Heine VM, Posthuma D. 2017. Genetically-informed patient selection for ipsc studies of complex diseases may aid in reducing cellular heterogeneity. *Frontiers in Cellular Neuroscience* **11**:164 DOI 10.3389/fncel.2017.00164.

Hoggart CJ, Choi SW, Garcia-Gonzalez J, Souaiaia T, Preuss M, O'Reilly P. 2023. Bridgeprs: a powerful trans-ancestry polygenic risk score method. BioRxiv DOI 10.1101/2023.02.17.528938.

Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, Xu X, Zhao H. 2017. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLOS Computational Biology* **13**(6):e1005589 DOI 10.1371/journal.pcbi.1005589.

Imam S, Noguera DR, Donohue TJ. 2014. Global analysis of photosynthesis transcriptional regulatory networks. *PLOS Genetics* **10**(12):e1004837 DOI 10.1371/journal.pgen.1004837.

Ju D, Hui D, Hammond DA, Wonkam A, Tishkoff SA. 2022. Importance of including non-european populations in large human genetic studies to enhance precision medicine. *Annual Review of Biomedical Data Science* **5**(1):321–339 DOI 10.1146/annurev-biodatasci-122220-112550.

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50**(9):1219–1224 DOI 10.1038/s41588-018-0183-z.

Kim DJ, Kang JH, Kim J-W, Cheon MJ, Kim Sb, Lee YK, Lee B-C. 2023. Evaluation of optimal methods and ancestries for calculating polygenic risk scores in east Asian population. *Scientific Reports* **13**(1):19195 DOI 10.1038/s41598-023-45859-w.

King N, Bishop C. 2017. New treatments for serious conditions: ethical implications. *Gene Therapy* **24**(9):534–538 DOI 10.1038/gt.2017.32.

Ko CL, Lin WZ, Lee M-T, Chang Y-T, Lin H-C, Wu Y-S, Lin J-F, Pan K-T, Chang Y-C, Lee K-H, Lee Y-L, Hsieh T-T, Huang J-H, Wang C-H, Yang S-S, Chen H-C, Chu C-M. 2022. Genome-wide association study reveals ethnicity-specific SNPs associated with ankylosing spondylitis in the Taiwanese population. *Journal of Translational Medicine* **20**:589 DOI 10.1186/s12967-022-03701-3.

Koch S, Schmidtke J, Krawczak M, Caliebe A. 2023. Clinical utility of polygenic risk scores: a critical 2023 appraisal. *Journal of Community Genetics* **14**(5):1–17 DOI 10.1007/s12687-023-00645-z.

Lam M, Chen C-Y, Li Z, Martin AR, Bryois J, Ma X, Gaspar H, Ikeda M, Benyamin B, Brown BC, Liu R, Zhou W, Guan L, Kamatani Y, Kim S-W, Kubo M, Kusumawardhani A, Liu C-M, Ma H, Periyasamy S, Takahashi A, Xu Z, Yu H, Zhu F, Chen WJ, Faraone S, Glatt SJ, He L, Hyman SE, Hwu H-G, McCarroll SA, Neale BM, Sklar P, Wildenauer DB, Yu X, Zhang D, Mowry BJ, Lee J, Holmans P, Xu S, Sullivan PF, Ripke S, O'Donovan MC, Daly MJ, Qin S, Sham P, Iwata N, Hong KS, Schwab SG, Yue W, Tsuang M, Liu J, Ma X, Kahn RS, Shi Y, Huang H. 2019. Comparative genetic architectures of schizophrenia in east Asian and European populations. *Nature Genetics* **51(12)**:1670–1678 DOI 10.1101/445874.

Leonenko G, Baker E, Stevenson-Hoare J, Sierksma A, Fiers M, Williams J, de Strooper B, Escott-Price V. 2021. Identifying individuals with high risk of alzheimer's disease using polygenic risk scores. *Nature Communications* **12(1)**:4506 DOI 10.21203/rs.3.rs-137252/v1.

Lerga-Jaso J, Terpolovsky A, Novković B, Osama A, Manson C, Bohn S, De Marino A, Kunitomi M, Yazdi PG. 2024. Optimization of multi-ancestry polygenic risk score disease prediction models. medRxiv DOI 10.1101/2024.04.17.24305723.

Lewis CM, Hagenaars SP. 2019. Progressing polygenic medicine in psychiatry through electronic health records. *JAMA Psychiatry* **76(5)**:470–472 DOI 10.1001/jamapsychiatry.2018.3975.

Lewis CM, Vassos E. 2017. Prospects for using risk scores in polygenic medicine. *Genome Medicine* **9**:96 DOI 10.1186/s13073-017-0489-y.

Lewis CM, Vassos E. 2020. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* **12**:44 DOI 10.1186/s13073-020-00742-5.

Li H, Zeng J, Snyder MP, Zhang S. 2024. PRS-Net: interpretable polygenic risk scores via geometric learning. In: *International Conference on Research in Computational Molecular Biology*. Cham: Springer, 377–380.

Lewis ACF, Green RC. 2021. Polygenic risk scores in the clinic: new perspectives needed on familiar ethical issues. *Genome Medicine* **13**:14 DOI 10.1186/s13073-021-00829-7.

Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. 2017. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology* **41(6)**:469–480 DOI 10.1101/058214.

Mamani NM. 2020. Machine learning techniques and polygenic risk score application to prediction genetic diseases. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* **9(1)**:5 DOI 10.14201/adcaij202091514.

Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen T-H, Wang Q, Bolla MK, Yang X, Adank MA, Ahearn T, Aittomäki K, Allen J, Andrulis IL, Anton-Culver H, Antonenkova NN, Arndt V, Aronson KJ, Auer PL, Auvinen P, Barrdahl M, Beane Freeman LE, Beckmann MW, Behrens S, Benitez J, Bermisheva M, Bernstein L, Blomqvist C, Bogdanova NV, Bojesen SE, Bonanni B, Børresen-Dale A-L, Brauch H, Bremer M, Brenner H, Brentnall A, Brock IW, Brooks-Wilson A, Brucker SY, Brüning T, Burwinkel B, Campa D, Carter BD, Castelao JE, Chanock SJ, Chlebowski R, Christiansen H, Clarke CL, Collée JM, Cordina-Duverger E, Cornelissen S, Couch FJ, Cox A, Cross SS, Czene K, Daly MB, Devilee P, Dörk T, dos Santos-Silva I, Durcan L, Dwek M, Eccles DM, Ekici AB, Eliassen AH, Ellberg C, Engel C, Eriksson M, Evans DG, Fasching PA, Figueroa J, Fletcher O, Flyger H, Försti A, Fritschi L, Gabrielson M, Gago-Dominguez M, Gapstur SM, García-Sáenz JA, Gaudet MM, Georgoulias V, Giles GG, Gilyazova IR, Glendon G, Goldberg MS, Goldgar DE, González-Neira A, Grenaker Alnæs GI, Grip M, Gronwald J, Grundy A, Guénel P, Haeberle L, Hahnen E, Haiman CA, Håkansson N, Hamann U, Hankinson SE, Harkness EF, Hart SN, He W, Hein A, Heyworth J, Hillemanns P, Hollestelle A, Hooning MJ, Hoover RN, Hopper JL, Howell A, Huang G, Humphreys K, Hunter DJ, Jakimovska M, Jakubowska A, Janni W,

Benoumhani et al. (2025), *PeerJ Comput. Sci.*, DOI 10.7717/peerj-cs.3039

21/27

John EM, Johnson N, Jones ME, Jukkola-Vuorinen A, Jung A, Kaaks R, Kaczmarek K, Kataja V, Keeman R, Kerin MJ, Khusnutdinova E, Kiiski JI, Knight JA, Ko Y-D, Kosma V-M, Koutros S, Kristensen VN, Krüger U, Kühl T, Lambrechts D, Le Marchand L, Lejbkowicz F, Lilyquist J, Lindblom A, Lindström S, Lissowska J, Lo W-Y, Loibl S, Long J, Lubiński J, Lux MP, MacInnis RJ, Maishman T, Makalic E, Maleva Kostovska I, Manoukian S, Margolin S, Martens JW, Martinez ME, Mavroudis D, McLean C, Meindl A, Menon U, Middha P, Miller N, Moreno F, Mulligan AM, Mulot C, Muñoz-Garzon VM, Neuhausen SL, Nevanlinna H, Neven P, Newman WG, Nielsen SF, Nordestgaard BG, Norman A, Offit K, Olson JE, Olsson H, Orr N, Pankratz VS, Park-Simon TW, Perez JI, Pérez-Barrios C, Peterlongo P, Peto J, Pinchev M, Plaseska-Karanfilska D, Polley EC, Prentice R, Presneau N, Prokofyeva D, Purrington K, Pylkäs K, Rack B, Radice P, Rau-Murthy R, Rennert R, Rennert HS, Rhenius V, Robson M, Romero A, Ruddy KJ, Ruebner M, et al. 2019. Polygenic risk scores for prediction of breast cancer and breast cancer subtype. *The American Journal of Human Genetics* **104(4)**:21–34 DOI 10.1111/joim.13649.

Mega JL, Stitziel NO, Smith JG, Chasman DI, Caulfield MJ, Devlin JJ, Nordio F, Hyde CL, Cannon CP, Sacks FM, Poulter NR, Sever PS, Ridker PM, Braunwald E, Melander O, Kathiresan S, Sabatine MS. 2015. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet* **385(9984)**:2264–2271 DOI 10.1016/s0140-6736(14)61730-x.

Musliner KL, Mortensen PB, McGrath JJ, Suppli NP, Hougaard DM, Bybjerg-Grauholm J, Bækvad-Hansen M, Andreassen O, Pedersen CB, Pedersen MG, Mors O, Nordentoft M, Børglum AD, Werge T, Agerbo E, for the Bipolar Disorder Working Group of the Psychiatric Genomics Consortium. 2019. Association of polygenic liabilities for major depression, bipolar disorder, and schizophrenia with risk for depression in the danish population. *JAMA Psychiatry* **76(5)**:516–525 DOI 10.1001/jamapsychiatry.2018.4166.

Muñoz M, Pong-Wong R, Canela-Xandri O, Rawlik K, Haley CS, Tenesa A. 2016. Evaluating the contribution of genetics and familial shared environment to common disease using the UK biobank. *Nature Genetics* **48(9)**:980–983 DOI 10.1038/ng.3618.

Márquez-Luna C, Gazal S, Loh P-R, Kim SS, Furlotte N, Auton A, Price AL. 2021. Incorporating functional priors improves polygenic prediction accuracy in UK biobank and 23andMe data sets. *Nature Communications* **12(1)**:6052 DOI 10.1101/375337.

Natarajan P, Young R, Stitziel NO, Padmanabhan S, Baber U, Mehran R, Sartori S, Fuster V, Reilly DF, Butterworth A, Rader DJ, Ford I, Sattar N, Kathiresan S. 2017. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135(22)**:2091–2101 DOI 10.1161/circulationaha.116.024436.

Newcombe PJ, Nelson CP, Samani NJ, Dudbridge F. 2019. A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genetic Epidemiology* **43(7)**:730–741 DOI 10.1002/gepi.22245.

Öztornaci RO, Coşgun E, Çolak C, Taşdelen B. 2023. Prediction of polygenic risk score by machine learning and deep learning methods in genome-wide association studies. BioRxiv 2022–12 DOI 10.1101/2022.12.30.522280.

Pain O, Glanville KP, Hagenaars SP, Selzam S, Fürtjes AE, Gaspar HA, Coleman JRI, Rimfeld K, Breen G, Plomin R, Folkersen L, Lewis CM. 2021. Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLOS Genetics* **17(5)**:e1009021 DOI 10.1101/2020.07.28.224782.

Pattee J, Pan W. 2020. Penalized regression and model selection methods for polygenic scores on summary statistics. *PLOS Computational Biology* **16(10)**:e1008271 DOI 10.1371/journal.pcbi.1008271.

Peng J, Bao Z, Li J, Han R, Wang Y, Han L, Peng J, Wang T, Hao J, Wei Z, Shang X. 2024. DeepRisk: a deep learning approach for genome-wide assessment of common disease risk. *Fundamental Research* **4(4)**:752–760 DOI 10.1016/j.fmre.2024.02.015.

Pham D, Truong B, Tran K, Ni G, Nguyen D, Tran TT, Tran MH, Nguyen Thuy D, Vo NS, Nguyen Q. 2022. Assessing polygenic risk score models for applications in populations with under-represented genomics data: an example of Vietnam. *Briefings in Bioinformatics* **23(6)**: bbac459 DOI 10.1093/bib/bbac459.

Privé F, Arbel J, Vilhjálmsson BJ. 2020. LDpred2: better, faster, stronger. *Bioinformatics* **36(22–23)**:5424–5431 DOI 10.1093/bioinformatics/btaa1029.

Privé F, Aschard H, Blum MG. 2019. Efficient implementation of penalized regression for genetic risk prediction. *Genetics* **212(1)**:65–74 DOI 10.1534/genetics.119.302019.

Privé F, Vilhjálmsson BJ, Aschard H, Blum MG. 2019. Making the most of clumping and thresholding for polygenic scores. *The American Journal of Human Genetics* **105(6)**:1213–1221 DOI 10.1101/653204.

Raben TG, Lello L, Widen E, Hsu SD. 2023. Biobank-scale methods and projections for sparse polygenic prediction from machine learning. *Scientific Reports* **13**:11662 DOI 10.1038/s41598-023-37580-5.

Ren D, An L, Li B, Qiao L, Liu W. 2021. Efficient weighting methods for genomic best linear-unbiased prediction (blup) adapted to the genetic architectures of quantitative traits. *Heredity* **126(2)**:320–334 DOI 10.1038/s41437-020-00372-y.

Ribbing J, Nyberg J, Caster O, Jonsson EN. 2007. The LASSO—a novel method for predictive covariate model building in nonlinear mixed effects models. *Journal of Pharmacokinetics and Pharmacodynamics* **34(4)**:485–517 DOI 10.1007/s10928-007-9057-1.

Robinson MR, Kleinman A, Graff M, Vinkhuyzen AAE, Couper D, Miller MB, Peyrot WJ, Abdellaoui A, Zietsch BP, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Alizadeh BZ, Boezen HM, Franke L, Harst Pvd, Navis G, Rots M, Snieder H, Swertz M, Wolffenbuttel BHR, Wijmenga C, Abecasis GR, Absher D, Alavere H, Albrecht E, Allen HL, Almgren P, Amin N, Amouyel P, Anderson D, Arnold AM, Arveiler D, Aspelund T, Asselbergs FW, Assimes TL, Atalay M, Attwood AP, Atwood LD, Bakker SJL, Balkau B, Balmforth AJ, Barlassina C, Barroso I, Basart H, Bauer S, Beckmann JS, Beilby JP, Bennett AJ, Ben-Shlomo Y, Bergman RN, Bergmann S, Berndt SI, Biffar R, Di Boehm BO, Boehnke M, Boeing H, Boerwinkle E, Bolton JL, Bonnefond A, Bonnycastle LL, Boomsma DI, Borecki IB, Bornstein SR, Bouatia-Naji N, Boucher G, Bragg-Gresham JL, Brambilla P, Bruinenberg M, Buchanan TA, Buechler C, Cadby G, Campbell H, Caulfield MJ, Cavalcanti-Proença C, Cesana G, Chanock SJ, Chasman DI, Chen Y-DI, Chines PS, Clegg DJ, Coin L, Collins FS, Connell JM, Cookson W, Cooper MN, Croteau-Chonka DC, Cupples LA, Cusi D, Day FR, Day INM, Dedoussis GV, Dei M, Deloukas P, Dermitzakis ET, Dimas AS, Dimitriou M, Dixon AL, Dörr M, van Duijn CM, Ebrahim S, Edkins S, Eiriksdottir G, Eisinger K, Eklund N, Elliott P, Erbel R, Erdmann J, Erdos MR, Eriksson JG, Esko T, Estrada K, Evans DM, de Faire U, Fall T, Farrall M, Feitosa MF, Ferrario MM, Ferreira T, Ferrieʻres J, Fischer K, Fisher E, Fowkes G, Fox CS, Franke L, Franks PW, Fraser RM, Frau F, Frayling T, Freimer NB, Froguel P, Fu M, Gaget S, Ganna A, Gejman PV, Gentilini D, Geus EJC, Gieger C, Gigante B, Gjesing AP, Glazer NL, Goddard ME, Goel A, Grallert H, Gräßler J, Grönberg H,

Groop LC, Groves CJ, Gudnason V, Guiducci C, Gustafsson S, Gyllensten U, Hall AS, Hall P, Hallmans G, Hamsten A, Hansen T, Haritunians T, van der Harris TB, Hartikainen A-L, Hassanali N, Hattersley AT, Havulinna AS, Hayward C, Heard-Costa NL, Heath AC, Hebebrand J, Heid IM, Herzig K-H, Hicks AA, Hingorani A, Hinney A, Hirschhorn JN, Hofman A, Holmes CC, Homuth G, Hottenga J-J, Hovingh KG, Hu FB, Hu Y-J, Huffman JE, Hui J, Huikuri H, Humphries SE, Hung J, Hunt SE, Hunter D, Hveem K, Hyppönen E, Igl W, Illig T, Ingelsson E, Iribarren C, Isomaa B, Jackson AU, Jacobs KB, James AL, Jansson J-O, Jarick I, Jarvelin M-R, et al. 2017. Genetic evidence of assortative mating in humans. *Nature Human Behaviour* **1(1)**:0016 DOI 10.1038/s41562-016-0016.

Ruderfer DM, Walsh CG, Aguirre MW, Tanigawa Y, Ribeiro JD, Franklin JC, Rivas MA. 2020. Significant shared heritability underlies suicide attempt and clinically predicted probability of attempting suicide. *Molecular Psychiatry* **25(10)**:2422–2430 DOI 10.1038/s41380-018-0326-8.

Sakaue S, Kanai M, Karjalainen J, Akiyama M, Kurki M, Matoba N, Takahashi A, Hirata M, Kubo M, Matsuda K, Murakami Y, Daly MJ, Kamatani Y, Okada Y. 2020. Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. *Nature Medicine* **26(4)**:542–548 DOI 10.1101/856351.

Shahbazi Z, Nowaczyk S. 2025. Towards personalized cardiometabolic risk prediction: a fusion of exposome and AI. *Heliyon* **11(1)**:e40859 DOI 10.1016/j.heliyon.2024.e40859.

Shan N, Xie Y, Song S, Jiang W, Wang Z, Hou L. 2021. A novel transcriptional risk score for risk prediction of complex human diseases. *Genetic Epidemiology* **45(8)**:811–820 DOI 10.1002/gepi.22424.

Shi J, Potash JB, Knowles JA, Weissman MM, Coryell W, Scheftner WA, Lawson WB, DePaulo JR, Gejman PV, Sanders AR, Johnson JK, Adams P, Chaudhury S, Jancic D, Evgrafov O, Zvinyatskovskiy A, Ertman N, Gladis M, Neimanas K, Goodell M, Hale N, Ney N, Verma R, Mirel D, Holmans P, Levinson DF. 2020. Antidepressant response in major depressive disorder: a genome-wide association study. MedRxiv2020–12 DOI 10.1101/2020.12.11.20245035.

Simona A, Song W, Bates DW, Samer CF. 2023. Polygenic risk scores in pharmacogenomics: opportunities and challenges—a mini review. *Frontiers in Genetics* **14**:1217049 DOI 10.3389/fgene.2023.1217049.

Slunecka JL, van der Zee MD, Beck JJ, Johnson BN, Finnicum CT, Pool R, Hottenga J-J, de Geus EJ, Ehli EA. 2021. Implementation and implications for polygenic risk scores in healthcare. *Human Genomics* **15**:46 DOI 10.1186/s40246-021-00339-y.

Song S, Jiang W, Hou L, Zhao H. 2020. Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLOS Computational Biology* **16(2)**:e1007565 DOI 10.1371/journal.pcbi.1007565.

Squires S, Weedon MN, Oram RA. 2023. Exploring the application of deep learning methods for polygenic risk score estimation. medRxiv2023–12 DOI 10.1101/2023.12.14.23299972.

Tan MJT, Kasireddy HR, Satriya AB, Abdul Karim H, AlDahoul N. 2025. Health is beyond genetics: on the integration of lifestyle and environment in real-time for hyper-personalized medicine. *Frontiers in Public Health* **12**:1522673 DOI 10.3389/fpubh.2024.1522673.

Thompson DJ, Wells D, Selzam S, Peneva I, Moore R, Sharp K, Tarran WA, Beard EJ, Riveros-Mckay F, Giner-Delgado C, Palmer D, Seth P, Harrison J, Futema M, McVean G, Plagnol V, Donnelly P, Weale ME. 2022. UK biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. MedRxiv2022–06 DOI 10.1101/2022.06.16.22276246.

Tsuo K, Shi Z, Ge T, Mandla R, Hou K, Ding Y, Pasaniuc B, Wang Y, Martin AR. 2024. All of us diversity and scale improve polygenic prediction contextually with greatest improvements for under-represented populations. BioRxiv DOI 10.1101/2024.08.06.606846.

Vilhjalmsson B, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, Genovese G, Loh P-R, Bhatia G, Do R, Hayeck T, Won H-H, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Kathiresan S, Pato M, Pato C, Tamimi R, Stahl E, Zaitlen N, Pasaniuc B, Schierup M, De Jager P, Patsopoulos N, McCarroll SA, Daly M, Purcell S, Chasman D, Neale B, Goddard M, Visscher PM, Kraft P, Patterson NJ, Price AL. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics* **97(4)**:576–592 DOI 10.1101/015859.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics* **101(1)**:5–22 DOI 10.1016/j.ajhg.2017.06.005.

Wang K, Gaitsch H, Poon H, Cox NJ, Rzhetsky A. 2017. Classification of common human diseases derived from shared genetic and environmental determinants. *Nature Genetics* **49(9)**:1319–1325 DOI 10.1038/ng.3931.

Wang Y, Zhu M, Ma H, Shen H. 2021. Polygenic risk scores: the future of cancer risk prediction, screening, and precision prevention. *Medical Review* **1(2)**:129–149 DOI 10.1515/mr-2021-0025.

Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera AV, Okada Y, Matsuda K, Yamanashi Y, Furukawa Y, Morisaki T, Murakami Y, Kamatani Y, Muto K, Nagai A, Obara W, Yamaji K, Takahashi K, Asai S, Takahashi Y, Suzuki T, Sinozaki N, Yamaguchi H, Minami S, Murayama S, Yoshimori K, Nagayama S, Obata D, Higashiyama M, Masumoto A, Koretsune Y, Martin AR, Finucane HK, Price AL. 2022. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nature Genetics* **54(4)**:450–458 DOI 10.1038/s41588-022-01036-9.

Wray NR, Goddard ME, Visscher PM. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* **17(10)**:1520–1528 DOI 10.1101/gr.6665407.

Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, Adams MJ, Agerbo E, Air TM, Andlauer TMF, Bacanu S-A, Bækvad-Hansen M, Beekman AFT, Bigdeli TB, Binder EB, Blackwood DRH, Bryois J, Buttenschøn HN, Bybjerg-Grauholm J, Cai N, Castelao E, Christensen JH, Clarke T-K, Coleman JIR, Colodro-Conde L, Couvy-Duchesne B, Craddock N, Crawford GE, Crowley CA, Dashti HS, Davies G, Deary IJ, Degenhardt F, Derks EM, Direk N, Dolan CV, Dunn EC, Eley TC, Eriksson N, Escott-Price V, Kiadeh FHF, Finucane HK, Forstner AJ, Frank J, Gaspar HA, Gill M, Giusti-Rodríguez P, Goes FS, Gordon SD, Grove J, Hall LS, Hannon E, Hansen CS, Hansen TF, Herms S, Hickie IB, Hoffmann P, Homuth G, Horn C, Hottenga J-J, Hougaard DM, Hu M, Hyde CL, Ising M, Jansen R, Jin F, Jorgenson E, Knowles JA, Kohane IS, Kraft J, Kretzschmar WW, Krogh J, Kutalik Z, Lane JM, Li Y, Li Y, Lind PA, Liu X, Lu L, MacIntyre DJ, MacKinnon DF, Maier RM, Maier W, Marchini J, Mbarek H, McGrath P, McGuffin P, Medland SE, Mehta D, Middeldorp CM, Mihailov E, Milaneschi Y, Milani L, Mill J, Mondimore FM, Montgomery GW, Mostafavi S, Mullins N, Nauck M, Ng B, Nivard MG, Nyholt DR, O'Reilly PF, Oskarsson H, Owen MJ, Painter JN, Pedersen CB, Pedersen MG, Peterson RE, Pettersson E, Peyrot WJ, Pistis G, Posthuma D, Purcell SM, Quiroz JA, Qvist P, Rice JP, Riley BP, Rivera M, Saeed Mirza S, Saxena R, Schoevers R, Schulte EC, Shen L, Shi J, Shyn SI, Sigurdsson E, Sinnamon GBC, Smit JH, Smith DJ, Stefansson H, Steinberg S, Stockmeier CA, Streit F, Strohmaier J, Tansey KE, Teismann H, Teumer A, Thompson W, Thomson PA, Thorgeirsson TE, Tian C, Traylor M, Treutlein J, Trubetskoy V, Uitterlinden AG, Umbricht

D, Van der Auwera S, Van Hemert AM, Viktorin A, Wang Y, Webb BT, Weinsheimer SM, Wellmann J, Willemsen G, Witt SH, Wu Y, Xi HS, Yang J, Zhang F, Arolt V, Baune BT, Berger K, Boomsma DI, Cichon S, Dannlowski U, de Geus ECG, DePaulo JR, Domenici R, Domschke E, Esko K, Grabe T, H. J, Hamilton SP, Hayward C, Heath AC, Hinds DA, Kendler KS, Kloiber S, Lewis G, Li QS, Lucae S, Madden PFA, Magnusson PK, Martin NG, McIntosh AM, Metspalu A, Mors O, Mortensen PB, Lewis R, C. M, Levinson DF, Breen G, Børglum AD, Sullivan PF. 2018. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics* **50**:668–681.

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. 2013. Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* **14(7)**:507–515 DOI 10.1038/nrg3457.

Xiao J, Cai M, Hu X, Wan X, Chen G, Yang C. 2022. Xpxp: improving polygenic prediction by cross-population and cross-phenotype analysis. *Bioinformatics* **38(7)**:1947–1955 DOI 10.1093/bioinformatics/btac029.

Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. 2017. Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics* **49(9)**:1304–1310 DOI 10.1038/ng.3941.

Yang S, Zhou X. 2020. Accurate and scalable construction of polygenic scores in large biobank data sets. *The American Journal of Human Genetics* **106(5)**:679–693 DOI 10.1016/j.ajhg.2020.03.013.

Zhang Y, Weng Y, Lund J. 2022. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics* **12(2)**:237 DOI 10.3390/diagnostics12020237.

Zhang J, Zhan J, Jin J, Ma C, Zhao R, O'Connell J, Jiang Y, Koelsch BL, Zhang H, Chatterjee N. 2024. An ensemble penalized regression method for multi-ancestry polygenic risk prediction. *Nature Communications* **15(1)**:3238 DOI 10.1101/2023.03.15.532652.

Zhang H, Zhan J, Jin J, Zhang J, Lu W, Zhao R, Ahearn TU, Yu Z, O'Connell J, Jiang Y, Chen T, Okuhara D, Aslibekyan S, Auton A, Babalola E, Bell RK, Bielenberg J, Bryc K, Bullis E, Coker D, Partida GC, Dhamija D, Das S, Elson SL, Eriksson N, Filshtein T, Fitch A, Fletez-Brant K, Fontanillas P, Freyman W, Granka JM, Heilbron K, Hernandez A, Hicks B, Hinds DA, Jewett EM, Kukar K, Kwong A, Lin K-H, Llamas BA, Lowe M, McCreight JC, McIntyre MH, Micheletti SJ, Moreno ME, Nandakumar P, Nguyen DT, Noblin ES, Petrakovitz AA, Poznik GD, Reynoso A, Schumacher M, Shastri AJ, Shelton JF, Shi J, Shringarpure S, Su QJ, Tat SA, Tchakouté CT, Tran V, Tung JY, Wang X, Wang W, Weldon CH, Wilton P, Wong CD, Garcia-Closas M, Lin X, Koelsch BL, Chatterjee N. 2023. A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nature Genetics* **55(10)**:1757–1768 DOI 10.1038/s41588-023-01501-z.

Zhou X, Chen Y, Ip FCF, Jiang Y, Cao H, Lv G, Zhong H, Chen J, Ye T, Chen Y, Zhang Y, Ma S, Lo RMN, Tong EPS, Weiner MW, Aisen P, Petersen R, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L, Green RC, Saykin AJ, Morris J, Shaw LM, Khachaturian Z, Sorensen G, Kuller L, Raichle M, Paul S, Davies P, Fillit H, Hefti F, Holtzman D, Mesulam MM, Potter W, Snyder P, Schwartz A, Montine T, Thomas RG, Donohue M, Walter S, Gessert D, Sather T, Jiminez G, Harvey D, Bernstein M, Thompson P, Schuff N, Borowski B, Gunter J, Senjem M, Vemuri P, Jones D, Kantarci K, Ward C, Koeppe RA, Foster N, Reiman EM, Chen K, Mathis C, Landau S, Cairns NJ, Householder E, Taylor-Reinwald L, Lee V, Korecka M, Figurski M, Crawford K, Neu S, Foroud TM, Potkin SG, Shen L, Faber K, Kim S, Nho K, Thal L, Buckholtz N, Albert M, Frank R, Hsiao J, Kaye J, Quinn J, Lind B, Carter R, Dolen S, Schneider LS, Pawluczyk S, Beccera M, Teodoro L, Spann BM, Brewer J, Vanderswag H, Fleisher A, Heidebrink JL, Lord JL, Mason SS, Albers CS, Knopman D, Johnson K, Doody RS, Villanueva-Meyer J, Chowdhury M, Rountree S, Dang M, Stern Y, Honig LS, Bell KL, Ances B, Carroll M, Leon S, Mintun MA, Schneider S, Oliver A, Marson D, Griffith R, Clark

D, Geldmacher D, Brockington J, Roberson E, Grossman H, Mitsis E, de Toledo-Morrell L, Shah RC, Duara R, Varon D, Greig MT, Roberts P, Onyike C, D'Agostino D, Kielb S, Galvin JE, Cerbone B, Michel CA, Rusinek H, de Leon MJ, Glodzik L, De Santi S, Doraiswamy PM, Petrella JR, Wong TZ, Arnold SE, Karlawish JH, Wolk D, Smith CD, Jicha G, Hardy P, Sinha P, Oates E, Conrad G, Lopez OL, Oakley M, Simpson DM, Porsteinsson AP, Goldstein BS, Martin K, Makino KM, Ismail MS, Brand C, Mulnard RA, Thai G, McAdams-Ortiz C, Womack K, Mathews D, Quiceno M, Diaz-Arrastia R, King R, Weiner M, Martin-Cook K, DeVous M, Levey AI, Lah JJ, Cellar JS, Burns JM, Anderson HS, Swerdlow RH, Apostolova L, Tingus K, Woo E, Silverman DHS, Lu PH, Bartzokis G, Graff-Radford NR, Parfitt F, Kendall T, Johnson H, Farlow MR, Hake AM, Matthews BR, Herring S, Hunt C, van Dyck CH, Carson RE, MacAvoy MG, Chertkow H, Bergman H, Hosein C, Hsiung G-YR, Feldman H, et al. 2023. Deep learning-based polygenic risk analysis for Alzheimer's disease prediction. *Communications Medicine* 3(1):49 DOI 10.1038/s43856-023-00269-x.

Zhou G, Qie X, Zhao H. 2023. A bayesian approach to correcting the attenuation bias of regression using polygenic risk score. BioRxiv 2023–11 DOI 10.1101/2023.11.27.568907.