



多基因遗传风险评分(PRS)在遗传疾病研究中的应用与前景：数据分析方法综述*

贺舒欣^{1,2)} 喻长顺^{1,2)} 贾晓冬^{1,2)} 陈建春^{1,2)} 闫克强^{1,2)**}

(¹⁾ 天津金域医学检验实验室, 天津 300392; ²⁾ 天津市神经系统疾病多组学精准诊断技术重点实验室, 天津 300392)

摘要 多基因遗传风险评分 (polygenic risk score, PRS) 是一种新兴的遗传数据分析方法。该方法通过对个体多个遗传变异位点的综合考虑, 对个体复杂疾病的遗传风险进行定量评估, 在遗传学领域受到广泛关注, 同时该方法的有效性也在临床应用中得到进一步验证。由于PRS的计算涉及大量的基因组数据分析, 其模型的数据选择、构建方法以及验证方法均存在较大差异。本综述结合目前已发表的PRS相关研究和算法, 对PRS模型以及其应用进行阐述。

关键词 多基因遗传风险评分, 全基因组关联研究, 遗传学, 复杂疾病

中图分类号 R466, Q3

DOI: 10.16476/j.pibb.2023.0389

随着新一代测序技术的发展, 成本更低的基因分型技术推动了大型遗传数据集的生成, 全基因组关联研究 (genome-wide association study, GWAS) 的出现也促进了研究人员对与常见复杂疾病相关的遗传变异的识别, 目前已报道了超过 50 000 个性状和疾病关联^[1]。GWAS 是指在人类全基因组范围内找出存在的序列变异, 从中筛选出与疾病相关的单核苷酸多态性 (single nucleotide polymorphism, SNP), 这些 SNP 被认作是评估复杂疾病的稳定且准确的影响因素^[2]。然而, 复杂疾病通常涉及到大量对表型影响微弱的遗传变异位点, 单个变异并不足以用来评估个体的患病风险。为解决上述问题, 多基因遗传风险评分 (polygenic risk score, PRS) 应运而生。PRS 模型通过将复杂疾病相关的 SNP 位点信息整合起来, 对每个 SNP 的微小效应进行叠加来衡量个体性状或复杂疾病的遗传风险, 用于复杂疾病的前期预防和精准诊疗。本综述对 PRS 模型的基本原理、构建和验证方法、应用和前景进行了全面探讨。综合而言, PRS 作为一种强大的遗传风险评估工具, 在解析复杂疾病的遗传基础以及实现个体化精准诊断和治疗方面, 均具有巨大潜力。

1 PRS模型的基本原理

PRS 通过计算多个遗传变异的累计影响以得到个体对某一表型的遗传风险的单值估计。根据每个 SNP 携带的等位基因数量 (0、1 或 2) 计算评分, 并依据每个 SNP 与表型的关联强度和方向进行加权, 最后求和得出预期风险评分^[3]。

通常假设 PRS 潜在的模型为加性模型:

$$PRS_i = \sum_{j=1}^M a_{ij} w_j$$

i 表示第 i 个个体, j 表示第 j 个 SNP, a_{ij} 为第 i 个个体的第 j 个 SNP 的等位基因拷贝数, w_j 为该 SNP 的权重。

目前大多数研究使用 GWAS 汇总统计数据进行分析, 基本流程如图 1 所示。GWAS 汇总统计数据主要包含 GWAS 分析中每个遗传变异所在染色体位置及坐标、效应等位基因型及其频率、其他等位基因型、效应值及 P 值、标准误差等。常见

* 天津市科技领军培育企业重大创新项目 (22YDPYSY00220) 资助。

** 通讯联系人。

Tel: 18813023833, E-mail: yankeqiang1993@163.com

收稿日期: 2023-10-08, 接受日期: 2024-02-07

效应值包括 β 值、比值比 (odds ratio, OR) 以及风险比 (hazard ratio, HR)。目前收录最全的 GWAS 数据库主要有 The NHGRI-EBI GWAS Catalog (<https://www.ebi.ac.uk/gwas/>) 和 IEU Open GWAS Project (<https://gwas.mrcieu.ac.uk/>)，除此之外也可以关注欧美或东亚 Biobank 的 GWAS 公开数据 (如日本 Biobank JENGER (<http://jenger.riken.jp/result>))。GWAS 采用病例/对照方法，将受疾病影响的病例组和健康对照组进行比较， P 值则是用于

检验零假设成立或表现更严重的可能性，代表的是目标 SNP 与表型的相关性是否具有统计学意义，在 GWAS 研究中通常以 $P \leq 5 \times 10^{-8}$ 作为评价相关与否的阈值^[4]。在 PRS 模型中，选择哪些 SNP 纳入分析是重要的一步：一些算法只包括 P 值低于一定阈值的 SNP，而其他算法可能会包括更多的 SNP。这种 P 值选择策略是不同 PRS 方法之间的主要差异之一。

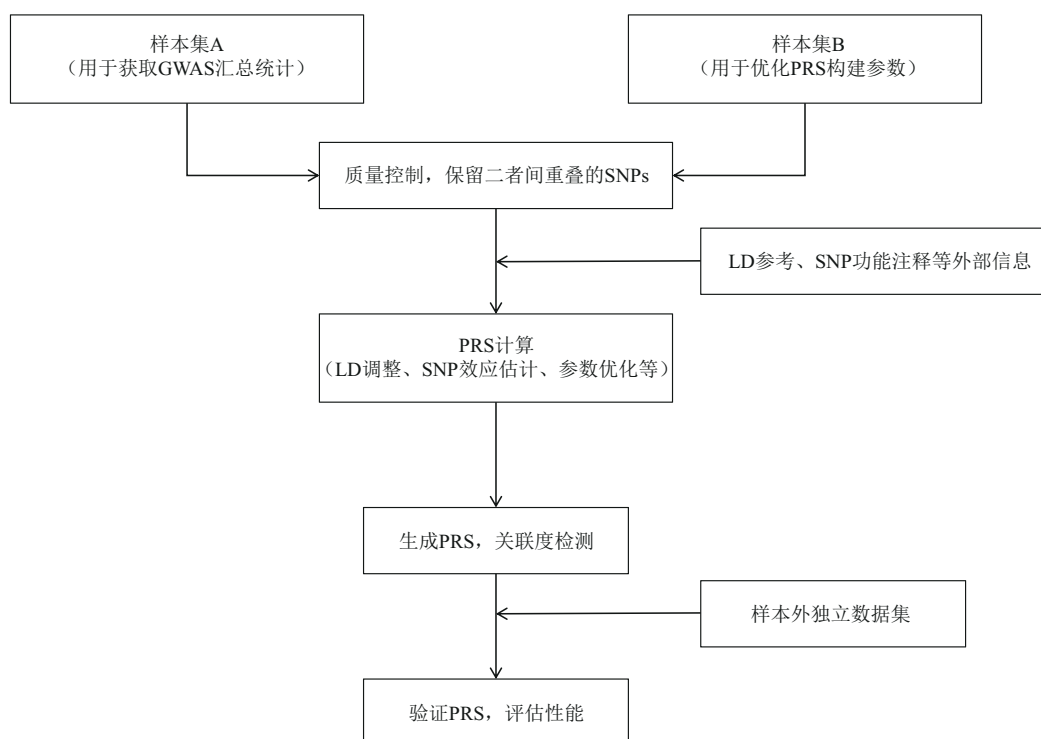


Fig. 1 The basic process of PRS model constructing based on SNP genetic variation sites

图1 基于SNP遗传变异位点利用GWAS汇总统计数据构建PRS的基本流程

LD: 连锁不平衡 (linkage disequilibrium)。

2 PRS模型构建方法

目前国内外已经开发了多种多基因评分方法 (表1)，主要目的在于更准确地估计 SNP 遗传力并纳入更多与基因组中相关的可用信息，根据构建模型的底层原理不同，可以大致将其分为以下几类。

2.1 基于GWAS汇总统计数据

2.1.1 经典计算方式: C+T method

经典的 PRS 构建方式是基于 GWAS 汇总数据，经连锁不平衡 (linkage disequilibrium, LD) 聚类后筛选得到各个基因座里 P 值最低的 SNP，最后留

下 P 值低于定义阈值并具有强关联信号的近似独立的 SNP 子集来构建 PRS。主要代表软件是 PLINK 和 PRSice，能够快速处理非常大的数据集。PRSice 通过多次测试自动计算任意 P 值阈值下的 PRS 以找到最合适的 P 值阈值，最后所选的 SNP 取决于参考的 LD 结构。2019 年 Choi 等^[5] 开发出了 PRSice-2 软件，较于上一代 PRSice 可以自动简化大规模数据的 PRS 分析，避免过度拟合同时大幅提高了计算效率。2019 年 Chaudhury 等^[6] 基于国际阿尔茨海默病基因组学计划 (IGAP) GWAS 产生的汇总统计数据，使用 PRSice2 预测迟发性阿尔茨海默病

(LOAD) 的发病风险, 结果表明, 结合了PRS的LOAD预测模型较仅包括遗传因素载脂蛋白E (apolipoprotein E, APOE) SNP、年龄和性别的预测模型在性能表现上有明显提高。

2.1.2 基于完全贝叶斯方法: SBayesR

SBayesR是基于完全贝叶斯模型来构建SNP效应分布, 利用GWAS汇总数据的特性无需调整队列直接学习遗传结构得到SNP效应权重, 用户使用时无需选择参数。同时, SBayesR通过预计算和高效存储稀疏的LD矩阵极大地提高了计算效率^[7]。Ni等^[8]研究显示, SBayesR较为适合应用在高度多基因性状的精神疾病等领域, SBayesR在精神分裂症和重度抑郁症的风险评估上有优异表现。

2.1.3 基于经验贝叶斯方法: LDpred2/PRS-CS

PRS-CS和LDpred在PRS构建上也同样具有高性能, 仅需要GWAS汇总统计数据和外部LD参考即可广泛应用。但与基于完全贝叶斯模型的方法不同, 它往往需要通过网格搜索优化特定的超参数, 同时通过马尔可夫链蒙特卡罗 (MCMC) 算法获得其他参数估计 (比如LDpred中的因果SNP分数以及PRS-CS中的全局收缩参数)。Ge等^[9]的研究表明, 当遗传结构稀疏时, 随着训练样本量的增加, PRS-CS受到的影响较小, 而LDpred会因LD中所有因果变异都变得具有高度统计学意义而导致预测性能不佳。不过这些在之后由Privé等^[10]推出的新版本LDpred2中都得到了很好的修复, 而且LDpred2可以在R包bigsnpr中实现。两种方法都已被证明可以很好地进行预测, 其中PRS-CS由于使用连续收缩先验而拥有较高的计算效率。PRS-CS和LDpred2都提供了无需调整样本的auto模式, 当训练样本量较大或难以获取独立验证集时可能会很有帮助。

2.1.4 基于频率论方法: DBSLMM/SBLUP

SBLUP (summary statistics BLUP) 基于线性混合模型, 通过最佳线性无偏预测在随机效应模型中重新估计SNP的效应值, 与只考虑无穷小情况的LDpred模型 (LDpred-inf) 相似。SBLUP分析主要通过GCTA软件实现, 需要GWAS汇总数据、LD参考和预先指定参数, 最后利用PLINK计算验证数据中个体的PRS^[11]。2020年Yang等^[12]开发了另一种同样无需调整样本优化参数就可以根据潜在的遗传结构自适应构建PRS的预测模型DBSLMM (deterministic Bayesian sparse linear

mixed model), 并且能将效应量较小的SNP也以合理的精度结合起来共同估计PRS, 实现了预测准确性的提升。除此之外, 基于C+T算法选择高效值的SNP也使DBSLMM具有较高的计算效率。

2.1.5 基于惩罚回归: lassosum

2017年Mak等^[13]指出, 在给定一些近似值并且仅使用GWAS汇总统计数据的情况下, 惩罚回归可用于估计多基因风险评分。lassosum是一种计算线性回归问题的LASSO/弹性网络 (elastic net, EN) 估计的方法, 需要调整参数以构建最佳PRS, 和PRS-CS一样都用到了LD区域块考虑局部LD矩阵。2021年Privé等^[14]提出了lassosum的新版本lassosum2, 并在R包bigsnpr中得以重新实现。lassosum2不仅使用与LDpred2完全相同的输入数据, 并且运行速度更快表现也更稳健。

一些研究针对不同的PRS模型构建算法进行评估。2021年Pain等^[15]在参考标准化框架内评估了多种PRS计算方法的预测效用, 发现DBSLMM比SBLUP在不同的遗传结构中相对更加稳定且占用内存更小, 无论参考基因型数据集样本量的或大或小都有更好的平均表现, 而SBayesR则在使用较大的LD参考时的性能会有明显提高。PRS-CS和LDpred2在不同表型之间较其他方法有更好表现, 不过在CPU时间上还有待提高^[16]。2021年Ni等^[8]在10种精神类疾病中比较了不同的多基因评分方法, 发现, lassosum作为单性状多基因预测的最佳方法之一, 虽然具有容易使用和运行时间短等优点, 且在多种疾病中也被证明有良好表现, 但有时会在不同疾病中表现出很大差异, 比如它对冠状动脉疾病和2型糖尿病几乎实现了最大的受试者工作特征曲线下面积 (area under the ROC curve, AUC) 值, 而对抑郁症的预测性能却低于传统C+T方法^[17]。此外, 通过增加考虑SNP的功能注释来计算权重也可能提高预测准确性, 如LDpred-funct算法, 但有研究显示, 在实践中与不使用功能注释的LDpred2相比并没有优势^[8]。

2.2 其他数据源

除了GWAS汇总统计数据, 个体级基因型数据也可以作为数据源进行PRS分析研究。2019年Privé等^[18]基于非常大的个体级SNP数据集开发了一种惩罚逻辑回归 (PLR) 方法, 与标准C+T方法相比, 其在大多数情况下都具有更高的预测性能和计算效率, 比如在利用英国生物银行150 000名个体预测乳腺癌的研究中只需几分钟即可拟合PLR。

与单变量 GWAS 汇总统计数据不同,使用真实的个体水平的基因型数据作为数据源联合估算 SNP 效应能显著提高 PRS 的预测性能,尤其是当邻近区域的相关 SNP 效应值极低时。另一种被广泛应用的依赖于个体水平的基因型和表型的 PRS 构建方法为贝叶斯稀疏线性混合模型 (Bayesian sparse linear mixed model, BSLMM),它是线性混合模型和稀疏回归模型的混合,并成功结合了两种模型的优点。BSLMM 假设所有 SNP 都具有非零效应,每个 SNP 效应大小来自于两个正态分布的混合,最后使用 MCMC 算法进行模型拟合。BSLMM 通过将 SNP 分为两类分别进行不同的收缩,在适当收缩效应值低的 SNP 的同时又避免了高效应值 SNP 的过度收缩,解释了 SNP 之间的连锁不平衡,能够适应所研究性状的不同遗传结构,证明 BSLMM 在疾病风险和复杂表型预测方面表现良好^[19-20]。不过这种仅通过 SNP 数据估算权重的方式对原始数据有一定的要求。由于实际研究中很难获得数十万人的数据集或合并单个队列数据,因而目前应用最广泛的仍是基于 GWAS 汇总统计数据的方法。

此外,相当一部分复杂性状的表达并非单纯由遗传变异决定,而是通过环境因素(辐射、长期隔

离等)、基因间相互作用、基因-环境相互作用等共同决定的,这种决定方式具有不可预测性的特点,进而限制了基于遗传变异计算的 PRS 的通用性。转录组测序数据、表观遗传组数据等基于高通量检测技术的组学数据可以动态地反映环境因素对人体基因表达造成的短期或长期的影响,对构建预警模型评估健康风险有潜在的应用价值^[21]。近年来已有研究将高通量检测技术的组学数据作为数据源以构建 PRS,其中大多数是基于机器学习开展的,例如贝叶斯岭 (Bayesian ridge, BR) 和弹性网络已被用于构建各种性状和疾病的遗传评分。2022 年 Xu 等^[22]发现,通过 EN 和 BR 设计的 PRS 模型对于血细胞特征预测做了极大的优化,其团队又于后一年针对约 50 000 名健康献血者的随机试验开发了基于血液 RNA 转录本、蛋白质和代谢物的遗传评分,采用 BR 方法构建了 17 227 个生物分子性状的遗传评分用于探究潜在的遗传机制和病理生理学^[23]。此外,也有研究基于 DNA 甲基化数据集通过轻量级梯度提升机 (LightGBM) 算法构建 PRS 模型预测肾透明细胞癌,癌症阶段表型预测 r^2 (相关系数的平方) 达到了 0.796^[24]。

Table1 PRS model construction methods

表1 PRS构建方法

方法	模型	SNP效应分布	软件	计算语言	调整样本	数据源	LD参考	SNP功能	引用数 (截至2023年7月7日)
C+T ^[25]	无模型,仅基于算法	—	PLINK PRSice-2	C/C++	是	GWAS汇总统计	是	—	29 873
Lassosum ^[14]	惩罚回归	$f(\beta) = y^T y + (1-s)\beta^T X_r^T X_r \beta - 2\beta^T r + s\beta^T \beta + 2\lambda \ \beta\ _1$	lassosum	R	是 (lassosum-auto无需)	GWAS汇总统计	是	—	290
PLR ^[19]	惩罚回归	$\beta_j \sim N\left(0, \frac{h_g^2}{ S_{causal} }\right); \beta_j \sim Laplace(0, \sqrt{\frac{h_g^2}{2 S_{causal} }})$	bigstatsr	R	是	个体级基因型数据	—	—	57
LDpred2 ^[11]	经验贝叶斯	$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{\pi M}\right), & \pi \\ 0, & 1-\pi \end{cases}$	bigsnpr	R	是 (LDpred2-auto无需)	GWAS汇总统计	是	—	251
PRS-CS ^[10]	经验贝叶斯	$\beta_j \sim N\left(0, \frac{\sigma^2}{N} \psi_j\right); \psi_j \sim G(a, \delta_j); \delta_j \sim G(b, \phi)$	PRS-CS	Python	是 (PRS-CS-auto无需)	GWAS汇总统计	是	—	698
SBayesR ^[8]	完全贝叶斯	$\beta_j \left \pi, \sigma_\beta^2 = \begin{cases} 0, & \pi_1 \\ \sim N(0, \gamma_2 \sigma_\beta^2), & \pi_2 \\ \vdots \\ \sim N(0, \gamma_C \sigma_\beta^2), & 1 - \sum_{c=1}^{C-1} \pi_c \end{cases}$	GCTB	C++	—	GWAS汇总统计	是	—	286

续表1										
方法	模型	SNP效应分布	软件	计算语言	调整样本	数据源	LD参考	SNP功能	引用数 (截至2023年7月7日)	
BSLMM ^[20]	完全贝叶斯	$\beta_j \sim \pi N(0, \sigma_i^2 \tau^{-1}) + (1 - \pi) N(0, \sigma_s^2 \tau^{-1})$	GEMMA	C/C++	-	个体级基因型数据	-	-	767	
DB-SLMM ^[12]	频率论	$\beta_{ij} \sim N(0, \sigma_i^2 \tau^{-1}); \beta_{sj} \sim N(0, \sigma_s^2 \tau^{-1})$	DBSLMM	C++	-	GWAS汇总统计	是	-	61	
SBLUP ^[13]	频率论	$\beta_j \sim N\left(0, \frac{h_g^2}{M}\right)$	GCTA	C++	-	GWAS汇总统计	是	-	267	
LDpred-funct ^[18]	频率论	$\beta_j \sim N(0, c * \sigma_j^2)$	LDpred-funct	Python	-	GWAS汇总统计	是	是	96	

列表包括基础模型、SNP效应大小估计原理、计算工具和所需数据。

3 PRS模型的验证

不同的PRS方法在SNP的纳入标准、效应值估算和LD衡量等方面的参数都需要借助由个体级基因型数据组成的样本进行测试,以实现PRS预测效果的最优化。PRS模型的构建中将目标样本随机分为训练数据集和测试数据集,使用训练样本优化参数;往往通过交叉验证和logistic回归模型来评价PRS模型的内部有效性,比如LDpred2、lassosum和PRS-CS都提供使用10折交叉验证来选择最优参数(达到最佳预测性能的P值阈值或收缩参数)^[26]。最后为了避免生成过拟合PRS模型和预测结果,研究往往需要样本外验证集通过预测发病风险模型的区分度和校准度等来评估最优PRS模型的预测性能,样本外验证集需要独立于基础和目标数据集^[27]。值得注意的是,由于PRS受种族、地域的影响,因此往往需要针对不同人群对PRS模型重新评估验证。

在测试集和验证集中评估PRS模型的预测能力主要根据以下4种指标^[28]。a. 模型拟合优度: 决定系数R²常被用于评价回归模型中预测值和实际样本观测值的符合程度,其取值在0到1之间,越接近1说明模型对于样本数据点的拟合优度越高,PRS对结局变量的解释程度也越高。b. 模型区分度: 一个好的风险预测模型能够正确区分不同发病风险的人群,AUC-ROC曲线就是最常用的判断模型区分度的重要评估指标,ROC概率曲线是由TPR(真阳率)与FPR(假阳率)绘制的,AUC即ROC曲线下面积,也叫C统计量,取值在0.5到1之间,AUC越大说明PRS模型的区分能力越好,

如果为0.5则表示PRS模型的预测概率与随机概率相同,即没有预测价值^[29]。近年来应用越来越多的指标还有净重新分类指数(net reclassification index, NRI)和附加NRI,常用于评估结合PRS后的新模型对旧模型的改善情况。c. 模型校准度: 预测模型的校准度作为反映了PRS模型预测风险与实际疾病发生风险的一致程度,比如Nam-D’Agostino卡方检验。校准度差的模型往往高估或低估发病风险,一个区分度很好的模型也可能有着较差的校准度。d. 模型预测能力: 基于PRS的结果分值可以将人群三等分、五等分或十等分,根据间距可分为低、中、高风险组并比较不同风险组人群发病的相对风险与绝对风险之间的差异,筛选疾病易感人群。

4 PRS模型的应用领域

通过整合遗传信息筛查高风险人群对疾病的早期防治有重要作用,PRS的临床意义越来越受到重视。随着各种PRS计算方法的不断开发和基因组学数据的不断积累,国内外已有相当多的研究表明,PRS在不同类型的人体性状或复杂疾病的风险预测、诊断治疗和预后生存估计中都具有良好的应用前景,促进了精准医学发展(表2)。

4.1 肿瘤性疾病

肿瘤的早期筛查一直是非常重要的研究热点,针对全体人群尤其是健康人群使用PRS预估癌症风险近年来发展迅速,目前已有多项研究建立PRS模型以量化肿瘤易感位点和临床特征对肺癌、乳腺癌、胃癌、肝癌和结直肠癌等疾病的遗传风险。

肺癌作为世界上最常见的恶性肿瘤之一,主要

环境风险因素是吸烟，但重度吸烟者中的70%并不会患肺癌，而约15%的肺癌患者也并非烟民，提示了遗传背景因素的重要作用。近年来，已有大量的肺癌相关的GWAS研究成果发表。2016年Cheng等^[30]的吸烟者肺癌风险研究中，基于38个SNP位点构建了PRS模型，*AUC*达到0.697。后续也有研究考虑了其他非遗传影响因素，2020年Huang等^[31]利用欧洲人后裔的肺癌GWAS数据中筛选出的18个SNP位点构建了PRS模型，评估了空气污染物和遗传风险之间可能存在的相互作用，结果表明，高空气污染暴露和高遗传风险的参与者患肺癌的风险最高，遗传因素和空气污染对肺癌的联合影响以剂量反应方式呈现。

据2020年全球癌症统计报告显示，女性中的乳腺癌数量已经超过肺癌，乳腺癌成为全球恶性肿瘤发病第一位^[32]。目前针对乳腺癌已经发表了多种PRS模型，*AUC*大多在0.60~0.64之间。根据乳腺癌协会联盟开发的包含313种遗传变异的PRS₃₁₃的*AUC*也仅为0.630 (95% *CI* 0.628~0.651)^[33]。2018年Khera等^[34]选择纳入更多遗传变异后将*AUC*提高到了0.690。Evans等^[35]于2022年的发表的一项研究表明，随着SNP数量的增加，PRS的人群的风险分层效果趋于改善，添加PRS₃₁₃后较单独使用乳腺X线摄影密度和TC风险预测模型(Tyrer-Cuzick risk prediction model，用于计算个人携带BRCA1或BRCA2突变的可能性)时可识别出潜在低风险(<1.4%)人群的比例由4.7%上升至24.89%，识别10年风险≥5%的病例的比例也升至40%以上。尽管PRS受种族、地域的影响，但最近的一项研究表明，基于非西班牙裔欧洲血统女性乳腺癌患者建立的PRS模型在拉丁裔女性中仍有良好的表现^[36]。

4.2 心脑血管疾病

冠心病作为一种复杂的多基因遗传病，是全球多重发病率和死亡率的主要原因之一。近些年来，随着高通量测序技术的进步，GWAS成功鉴定了多个冠心病易感位点，并且多项研究支持PRS模型可以有效地评估冠心病患病风险。随着越来越多遗传变异的纳入，PRS模型对冠心病的预测能力也不断提高。有报道称，通过构建包含663万个SNP的PRS模型，*AUC*达到了0.81^[34]。2018年Tikkanen等^[37]研究发现握力、体力活动和心肺健康与心血管事件呈负相关，高水平的心肺健康能够使高遗传风险人群的冠心病风险降低49%。有证据支持，与

非遗传性冠心病风险因素相比，遗传风险会随着年龄的增长而递减^[38]。2020年Elliott等^[39]比较了PRS模型与汇集队列风险方程(pooled cohort equations, PCE)在临床实践中预测普通人群患病风险的表现，发现将冠心病PRS模型添加到PCE中可以提高发病预测的准确性，改善较年轻的年龄组的区分度。2023年由中国科学院北京基因组研究所汪敏先研究组与美国博德研究所Amit V. Khera研究组合作开发的新模型GPSmult整合了冠心病和相关风险因素特征的GWAS多祖先汇总统计数据，使用样本数超过1 346万例，能够进一步改善约40%个体的风险预测准确性，特别是显著改善了东亚和南亚人群的患病风险预测^[40]。

4.3 神经类疾病

阿尔茨海默病(AD)是一种进行性神经退行性疾病，特征主要是记忆、语言和行为退化，因其起病隐匿性和具有很强的遗传因素，在发病前对高危人群进行针对性预防具有非常重要的意义。已有多团队发表了用于AD发病风险评估的PRS模型。2017年Chaudhury等^[41]使用PRSice建立了PRS模型，预测AD发病风险的准确率为72.9%。2021年Leonenko等^[42]不仅为AD应该被界定为多基因疾病提供了充分证据，还表明年龄和APOE-ε4等位基因状态是影响PRS识别高风险人群的重要因素，PRS分布极端的个体OR和*AUC*都非常高。

帕金森病(PD)是一种仅次于AD的常见神经退行性疾病，PD的发病风险同时关乎遗传变异与非遗传因素。目前已经发表的GWAS研究大多是针对欧洲人群，较少针对亚洲人群。2020年Foo等^[43]发表的一项主要针对亚洲人群的GWAS研究发现，欧亚两个人群的PD风险位点存在相当大的重叠面，78个欧洲人群PD风险位点中的63个在亚洲人群中具有同方向的OR值，另15个位点在亚洲人群中表现为反方向的OR值，Foo等还发现了2个新的PD风险位点(SV2C、WBSCR17)。2021年Sia等^[44]利用亚洲人群遗传位点发现了调整非遗传风险因素后PRS与中国人群的PD发病率存在显著相关性。由于上述研究进展，PRS已经可以用于PD的患病风险以及疾病进展情况的评估。

4.4 代谢性疾病

糖尿病是以高血糖、高心血管疾病风险为特点的慢性代谢性疾病，有研究认为，预计到2045年全球范围内2型糖尿病(T2DM)患者将达到6.29亿。2008年，Meigs等^[45]在临床风险因素的

Table 2 Research results on genetic risk prediction models for common complex diseases
表2 常见复杂疾病遗传风险预测模型的研究成果

疾病类型	数据收集	人群	验证样本量 (病例/对照)	PRS构建方法	遗传标志物数量	非遗传标志物	AUC (95%CI)		
							非遗传	遗传	合并
肿瘤性疾病	乳腺癌	欧洲	6 586/157 895	C+T	5 218	年龄、性别、祖先、基因分型阵列	—	—	0.690 [34]
	肺癌	个体级基因型数据 亚洲	1 937/1 984	Lasso惩罚回归	38	吸烟	0.610	0.653	0.697 [30]
心脑血管疾病	冠心病	欧洲	6 272/346 388	C+T	2 363	PCE (包括年龄、性别、种族、行为习惯、血压、血脂等数据)	0.76	0.61	0.78 [39]
				lassosum	40 079 (CAD); 29 7862 (CVD)				
神经类疾病	冠心病	欧洲	8 676/288 978	LDPred	6 630 150	年龄、性别、祖先、基因分型阵列	—	0.81	0.81 [34]
	阿尔茨海默病	欧洲	408/436	C+T	28	—	—	0.729	0.729 [40]
代谢性疾病	帕金森病	亚洲	333/25 313	C+T	6	性别、吸烟、红茶摄入量、咖啡因摄入量、单饱和脂肪摄入量和胆固醇摄入量等	—	0.63	0.63 [43]
	2型糖尿病	欧洲	255/2 179	P值	18	性别、年龄、家族	0.595	—	0.615 [44]
自身免疫疾病	1型糖尿病	个体级基因型数据 欧洲	5 853/288 978	LDPred	6 917 436	年龄、性别、祖先、基因分型阵列	—	0.73	0.73 [34]
			765/423	贝叶斯模型	41	—	—	0.84	0.84 [46]

CI: 置信区间 (confidence interval); CAD: 冠心病 (coronary artery disease); CVD: 心血管疾病 (cardiovascular disease)。

基础上叠加了基于18个风险等位基因的PRS评分,发现PRS与临床风险因素一起使用时明显对新发糖尿病病例的预测能力有显著的改善。2018年,Mahajan等^[46]汇总了来自32个欧洲后裔的全基因组基因分型GWAS数据并构建了PRS模型,按照该模型的评分,处于前2.5%的个体相较中位数个体而言患T2DM的风险提高3.4倍,相较末位2.5%的个体的患病风险提高9.4倍,并且评分还与多种肥胖等代谢紊乱的风险因素呈正相关。在2022年,已有研究通过构建跨祖先的PRS模型对非欧洲人群中的T2DM PRS进行全面评估^[47]。

4.5 自身免疫疾病

1型糖尿病(T1DM)涉及胰腺 β 细胞的自身免疫性破坏,主要发生在30岁以下的人群中,已知多种功能基因与T1DM的患病风险相关,其中,HLA基因与约50%的T1DM相关。2014年Winkler等^[48]结合HLA等位基因分型结果以及40个其他基因的变异,通过多变量逻辑回归和贝叶斯特征选择构建PRS模型,结果显示,其能够有效改善前瞻性随访儿童的风险分层效果,帮助具有家族史和高遗传风险评分的儿童实现早期临床前识别。年轻群体中肥胖率的升高往往使得T1DM和T2DM的区分变得困难,而误诊会导致更高的药物成本或副作用,甚至危及生命。2015年Oram等^[49]开发了可以准确识别需要胰岛素治疗的年轻糖尿病患者的遗传风险评分T1DM-PRS,其区分能力独立于胰岛自身抗体、体重指数(BMI)和诊断年龄,有助于糖尿病患者的临床分型。2021年Qu等^[50]使用PRSice-2识别了T1DM PRS较低的病例,并通过全基因组关联测试与对照进行比较鉴别出了13个高质量新位点,其中9个已被报告与肥胖的相关性状有关。

5 PRS模型的前景和挑战

PRS在疾病的精准预防、改善诊断和风险分层等方面已初显应用价值,结合临床特征构建的评分模型在临床实践上也有了一定表现^[39, 51],但在实现广泛应用方面仍存在挑战。基因与环境的相互作用也会影响复杂疾病产生和发展,很多情况下某些复杂性状的表现并不通过遗传决定的方式。潜在的与环境有相互作用的SNP若不具备主效应或与主效应方向相反则可能无法被识别,SNP与环境间互动方式的不同也可能导致这种相互作用的抵消,因此仅通过遗传位点构建的PRS模型在GWAS样本量较小时很有可能被环境因素左右预测效果^[52]。

PRS预测模型的精确度主要取决于特定性状或疾病的遗传度,准确度则主要取决于GWAS研究的样本量,增加GWAS样本量也能在一定程度上减轻环境风险因素对预测准确性的影响。研究中为了保证所构建的PRS模型足够准确,往往需要对数十万甚至数百万个样本做全基因组测序。基因组分析公司(Genomics PLC)与英国启动的涉及500万人的前瞻性研究“Our Future Health”项目合作,将多基因风险评分作为开发预防、检测和治疗疾病新方法的重要部分。此外,PRS构建方法和软件开发也存在进步的空间。传统C+T方法中使用阈值过于严格可能会导致遗传位点选择的丢失,过于宽松又会严重降低预测效果,而有时研究中会随意设定阈值又不能给予充分解释。LDpred2和PRS-CS等虽然能够纳入所有SNP到PRS的构建中,但在疾病的实际预测中常常并不如基于C+T方法设计的软件应用广,有时在预测性能上也难以看到较C+T方法的明显改善,并且运行速度和占用内存也是二者需要考虑和解决的问题。而且当前的机器学习系统由于其理论上的局限性,研究人员是在一种不透明的前提下使用深度学习模式,很难解释问题是出在程序、方法还是环境的改变上,这都给未来多基因风险评分生信工具的开发提供了新的考虑和广阔的发展空间。

由于复杂疾病之间的临床特征很有可能存在一定的关联,PRS跨疾病风险预测也具有重要的研究价值,比如T2DM是冠心病最重要的发病危险因素之一,还有研究显示将脑卒中表型及其相关亚表型都作为构建冠心病meta-PRS模型的重要组成部分并且预测表现良好^[53]。同时,多组学数据如转录组、表观遗传组数据等往往能够通过动态反应基因-环境相互作用揭示遗传研究以外的见解,机器学习和人工智能技术的不断发展与进步为整合不同类型数据共同预测疾病风险提供了可能。

参 考 文 献

- [1] Albiñana C, Grove J, McGrath J J, *et al.* Leveraging both individual-level genetic data and GWAS summary statistics increases polygenic prediction. *Am J Hum Genet*, 2021, **108**(6): 1001-1011
- [2] Nachman M W. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet*, 2001, **17**(9): 481-485
- [3] Ma Y, Zhou X. Genetic prediction of complex traits with polygenic scores: a statistical review. *Trends Genet*, 2021, **37**(11): 995-1011
- [4] Pe'er I, Yelensky R, Altshuler D, *et al.* Estimation of the multiple

- testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol*, 2008, **32**(4): 381-385
- [5] Choi S W, O'Reilly P F. PRSice-2: polygenic risk score software for Biobank-scale data. *Gigascience*, 2019, **8**(7): giz082
- [6] Chaudhury S, Brookes K J, Patel T, *et al.* Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment *Psychiatry. Transl Psychiatry*, 2019, **9**(1): 154
- [7] Lloyd-Jones L R, Zeng J, Sidorenko J, *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun*, 2019, **10**(1): 5086
- [8] Ni G, Zeng J, Revez J A, *et al.* A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biol Psychiatry*, 2021, **90**(9): 611-620
- [9] Ge T, Chen C Y, *et al.* Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*, 2019, **10**(1): 1776
- [10] Privé F, Arbel J, Vilhjálmsson B J. LDpred2: better, faster, stronger. *Bioinformatics*, 2021, **36**(22-23): 5424-5431
- [11] Yang S, Zhou X. Accurate and scalable construction of polygenic scores in large Biobank data sets. *Am J Hum Genet*, 2020, **106**(5): 679-693
- [12] Robinson M R, Kleinman A, Graff M, *et al.* Genetic evidence of assortative mating in humans. *Nat Hum Behav*, 2017, **1**: 0016
- [13] Mak T S H, Porsch R M, Choi S W, *et al.* Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol*, 2017, **41**(6): 469-480
- [14] Privé F, Arbel J, Aschard H, *et al.* Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *HGG Adv*, 2022, **3**(4): 100136
- [15] Pain O, Glanville K P, Hagenaars S P, *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet*, 2021, **17**(5): e1009021
- [16] Privé F, Vilhjálmsson B J, Aschard H, *et al.* Making the most of clumping and thresholding for polygenic Scores. *Am J Hum Genet*, 2019, **105**(6): 1213-1221
- [17] Márquez-Luna C, Gazal S, Loh P R, *et al.* Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nat Commun*, 2021, **12**: 6052
- [18] Privé F, Aschard H, Blum M G B. Efficient implementation of penalized regression for genetic risk prediction. *Genetics*, 2019, **212**(1): 65-74
- [19] Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet*, 2013, **9**(2): e1003264
- [20] Moser G, Lee S H, Hayes B J, *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet*, 2015, **11**(4): e1004969
- [21] 李旭. 基于组学数据的模拟空间环境健康监测及其应用研究 [D]. 哈尔滨: 哈尔滨工业大学生物医学工程系, 2016
- Li Xu. A Research and Application on the Simulated Space Environment Health Monitoring Using Multi-omics Data[D]. Harbin: Biomedical Engineering, Harbin Institute of Technology, 2016
- [22] Xu Y, Vuckovic D, Ritchie S C, *et al.* Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease. *Cell Genom*, 2022, **2**(1): 100086
- [23] Xu Y, Ritchie S C, Liang Y, *et al.* An atlas of genetic scores to predict multi-omic traits. *Nature*, 2023, **616**(7955): 123-131
- [24] 潘建桥. 基于机器学习和基因组学数据的多基因风险评分模型研究[D]. 大连: 大连海事大学电子与通信工程系, 2021
- Pan J Q. Research on Ploygenic Risk Score Model Based on Machine Learning and Genomics Data[D]. Dalian: Electronics and Communication Engineering, Dalian Maritime University, 2021
- [25] Purcell S, Neale B, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Human Genet*, 2007, **81**(3): 559-75
- [26] Choi S W, Mak T S, *et al.* Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*, 2020, **15**(9): 2759-2772
- [27] 邢小龙. 基于传统和遗传危险因素的脑卒中个体化发病风险预测研究[D]. 北京: 北京协和医学院流行病与卫生统计学系, 2021
- Xing X L. Prediction for Personalized Stroke Risk With Traditional and Genetic Risk Factors[D]. Beijing: Epidemiology and Health Statistics, Peking Union Medical College, 2021
- [28] 汪天培, 靳光付, 胡志斌, 等. 多基因遗传风险评分用于精准预防的研究进展. *中华疾病控制杂志*, 2021, **25**(9): 993-997
- Wang T P, Jin G F, Hu Z B, *et al.* Chinese Journal of Disease Control & Prevention, 2021, **25**(9): 993-997
- [29] Uno H, Cai T, Pencina M J, *et al.* On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*, 2011, **30**(10): 1105-1117
- [30] Cheng Y, Jiang T, Zhu M, *et al.* Risk assessment models for genetic risk predictors of lung cancer using two-stage replication for Asian and European populations. *Oncotarget*, 2016, **8**(33): 53959-53967
- [31] Huang Y, Zhu M, Ji M, *et al.* Air pollution, genetic factors, and the risk of lung cancer: a prospective study in the UK Biobank. *Am J Respir Crit Care Med*, 2021, **204**(7): 817-825
- [32] 王悠清. 2020年全球癌症统计报告. *中华预防医学杂志*, 2021, **55**(3): 398-398
- Wang Y Q. Chinese Journal of Preventive Medicine, 2021, **55**(3): 398-398
- [33] Mavaddat N, Michailidou K, Dennis J, *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet*, 2019, **104**(1): 21-34
- [34] Khera A V, Chaffin M, Aragam K G, *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*, 2018, **50**(9): 1219-1224
- [35] Evans D G R, van Veen E M, Harkness E F, *et al.* Breast cancer risk stratification in women of screening age: incremental effects of adding mammographic density, polygenic risk, and a gene panel. *Genet Med*, 2022, **24**(7): 1485-1494

- [36] Shieh Y, Fejerman L, Lott P C, *et al.* A polygenic risk score for breast cancer in US Latinas and Latin American women. *J Natl Cancer Inst*, 2020, **112**(6): 590-598
- [37] Tikkanen E, Gustafsson S, Ingelsson E. Associations of fitness, physical activity, strength, and genetic risk with cardiovascular disease: longitudinal analyses in the UK Biobank study. *Circulation*, 2018, **137**(24): 2583-2591
- [38] Jiang X, Holmes C, McVean G. The impact of age on genetic risk for common diseases. *PLoS Genet*, 2021, **17**(8): e1009723
- [39] Elliott J, Bodinier B, Bond T A, *et al.* Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. *JAMA*, 2020, **323**(7): 636-645
- [40] Patel A P, Wang M, Ruan Y, *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat Med*, 2023, **29**(7): 1793-1803
- [41] Chaudhury S, Patel T, Barber I S, *et al.* Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimer's disease. *Neurobiol Aging*, 2018, **62**: 244.e1-244.e8
- [42] Leonenko G, Baker E, Stevenson-Hoare J, *et al.* Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. *Nat Commun*, 2021, **12**(1): 4506
- [43] Foo J N, Chew E G Y, Chung S J, *et al.* Identification of risk loci for Parkinson disease in Asians and comparison of risk between Asians and Europeans: a genome-wide association study. *JAMA Neurol*, 2020, **77**(6): 746-754
- [44] Sia M W, Foo J N, Saffari S E, *et al.* Polygenic risk scores in a prospective Parkinson's disease cohort. *Mov Disord*, 2021, **36**(12): 2936-2940
- [45] Meigs J B, Shrader P, Sullivan L M, *et al.* Genotype score in addition to common risk factors for prediction of type 2 diabetes. *New Engl J Med*, 2008, **359**(21): 2208-2219
- [46] Mahajan A, Taliun D, Thurner M, *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet*, 2018, **50**(11): 1505-1513
- [47] Ge T, Irvin M R, Patki A, *et al.* Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Med*, 2022, **14**: 70
- [48] Winkler C, Krumsiek J, Buettner F, *et al.* Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes. *Diabetologia*, 2014, **57**(12): 2521-2529
- [49] Oram R A, Patel K, Hill A, *et al.* A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes Care*, 2016, **39**(3): 337-344
- [50] Qu H Q, Qu J, Bradfield J, *et al.* Genetic architecture of type 1 diabetes with low genetic risk score informed by 41 unreported loci. *Commun Biol*, 2021, **4**: 908
- [51] Patel A P, Wang M, Ruan Y, *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nat Med*, 2023, **29**(7): 1793-1803
- [52] Tang Y, You D, *et al.* IPRS: leveraging gene-environment interaction to reconstruct polygenic risk score. *Front Genet*, 2022, **13**: 801397
- [53] 刘钟应. 中国人群冠心病多基因遗传风险评分的构建及不同遗传风险下心血管健康获益评价[D]. 北京: 北京协和医学院流行病与卫生统计学系, 2021
- Liu Z Y. Development of a Polygenic Risk Score for Coronary Heart Disease and Evaluation of Cardiovascular Health Benefits Among the Chinese Population [D]. Beijing: Epidemiology and Health Statistics, Peking Union Medical College, 2021

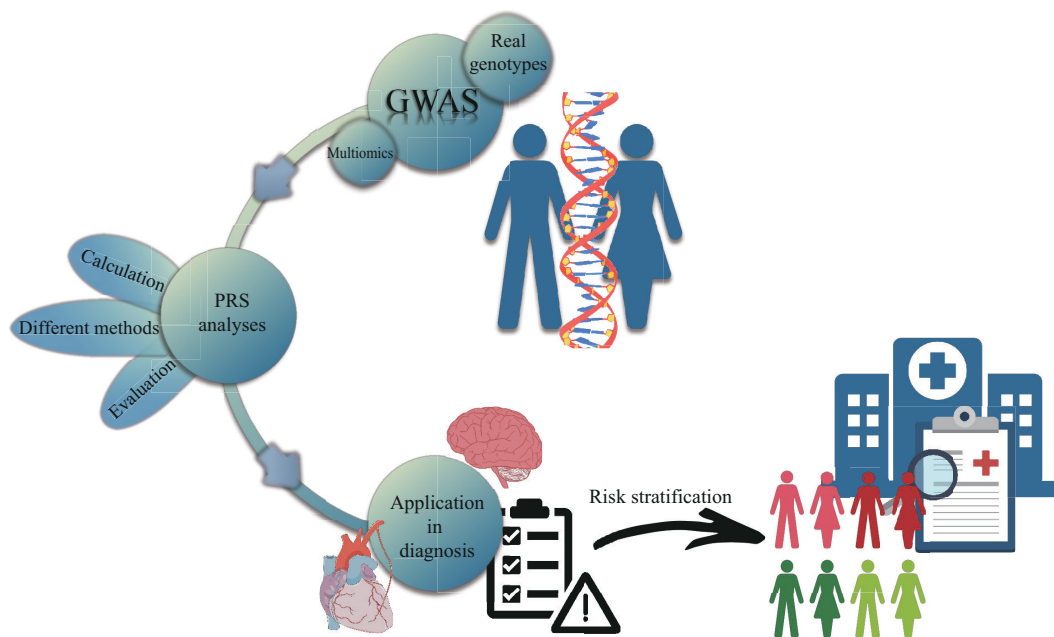
Application and Prospects of Polygenic Risk Score (PRS) in Genetic Disease
Research: a Review of Data Analysis Methods*

HE Shu-Xin^{1,2)}, YU Chang-Shun^{1,2)}, JIA Xiao-Dong^{1,2)}, CHEN Jian-Chun^{1,2)}, YAN Ke-Qiang^{1,2)**}

⁽¹⁾Tianjin Kingmed Center for Clinical Laboratory, Tianjin 300392, China;

⁽²⁾Tianjin Key Laboratory of Multiomics Precision Diagnosis Technology of Neurological Diseases, Tianjin 300392, China)

Graphical abstract



Abstract Lower-cost genotyping technology has promoted the generation of large genetic datasets with the evolving next-generation sequencing technology. The emergence of genome-wide association studies (GWAS) has facilitated researchers’ understanding of common complex diseases. GWAS refers to finding the sequence variations present in the human genome and screening out disease-related single nucleotide polymorphisms (SNPs). These SNPs are considered as the basis for assessing the stability of complex diseases. However, a single variation is not sufficient to assess an individual’s risk of disease. Polygenic risk score (PRS) is an emerging genetic data analysis method for quantitatively estimating an individual’s genetic risk for complex diseases by comprehensively considering multiple genetic variation sites. A single-value estimate of an individual’s genetic risk for a certain phenotype can be calculated as the cumulative impact of multiple genetic variants by building a

* This work was supported by a grant from Tianjin Science and Technology Leading Enterprises Cultivation Major Innovation Project (22YDPYSY00220).

** Corresponding author.

Tel: 86-18813023833, E-mail: yankeqiang1993@163.com

Received: October 8, 2023 Accepted: February 7, 2024

PRS model. The finally expected risk score is weighted by the strength and direction of association of each SNP with the phenotype based on the number of alleles carried by each SNP. With the continuous development of various PRS calculation methods and the constant accumulation of genomic data, PRS has received widespread attention in the field of genetics. So far, quite a few studies at home and abroad have shown that PRS is valuable in risk prediction of different types of human traits or complex diseases, and its effectiveness has been further verified in clinical applications. At present, many studies have established PRS models based on GWAS summary statistics to quantify the genetic risk of susceptibility loci and clinical characteristics on diseases such as lung cancer, breast cancer, coronary heart disease, diabetes and Alzheimer's disease. The disease-susceptible populations can be recognized through comparing the relative risk and absolute risk of the disease in different risk groups according to the population risk stratification results. Additionally, individual-level genotype data and omics data can also be used as data sources for PRS analysis research, especially the latter can dynamically reflect the short-term or long-term effects of environmental factors on human gene expression, and has potential application value in building early warning models to assess health risks. Since the calculation of PRS involves a large amount of genomic data analysis, there are big differences in the methods for data selection, model building and validation. Different PRS construction methods and software have different performances in disease risk prediction, and even the performance of same algorithm varies across diseases. It is worth noting that the PRS model often needs to be re-evaluated and verified for different groups of people, because PRS is affected by race and region. This review combines currently published PRS-related research and algorithms to describe the basic principles of PRS, compares their construction and verification methods, and discusses their applications and prospects. As a powerful genetic risk assessment tool, PRS has great potential in analyzing the genetic code of complex diseases and achieving precise diagnosis and personalized treatment.

Key words polygenic genetic risk score, genome-wide association study, genetics, complex disease

DOI: 10.16476/j.pibb.2023.0389