·述 评·

多基因风险评分在恶性肿瘤风险预测中的应用展望

马红霞*

南京医科大学公共卫生学院流行病学系,江苏 南京 211166

[摘 要] 多基因风险评分通过综合多个易感位点的累积效应预测个体的肿瘤发病风险,对肿瘤高危人群的筛查和有效防治具有应用价值。近年来,多基因风险评分的构建方法和应用范畴得到了进一步拓展和完善。本文就多基因风险评分应用于恶性肿瘤风险预测的最新进展进行简要介绍,总结其应用展望和挑战。

[关键词] 多基因风险评分;全基因组关联研究;恶性肿瘤;风险预测

[中图分类号] R730.1 [文献标志码] A [文章编号] 1007-4368(2020)04-467-03

doi:10.7655/NYDXBNS20200401

Application of polygenic risk scores in the risk prediction of cancer

MA Hongxia*

Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing 211166, China

[**Abstract**] Polygenic risk score predicts cancer risk by integrating the effects of multiple susceptibility loci, which has potential application in the screening of high-risk population. In recent years, the construction method and application of polygenic risk score have been further improved and expanded. This paper reviews the most up-to-date advances in the application of polygenic risk score to the risk prediction of cancer, and summarizes the prospects and challenges.

[**Key words**] polygenic risk score; genome-wide association study; cancer; risk prediction

[J Nanjing Med Univ, 2020, 40(04): 467-469]

恶性肿瘤是威胁人类健康最主要的公共卫生问题之一。根据国际癌症研究机构(International Agency for Research on Cancer, IARC)最新的统计数据,2018年全球新发恶性肿瘤1 810万人,因恶性肿瘤死亡的人数达960万[1]。因此,阐明恶性肿瘤发生的危险因素,建立切实有效的防治措施,对于预防肿瘤发生和降低肿瘤死亡率具有重要意义。疾病风险预测模型可基于个体所具有的危险因素,建立统计模型,用于评估其在未来某个时间段内发生某种疾病的概率。传统的肿瘤风险预测模型通常基于人口学特征、环境危险因素暴露、生活方式以及临床指标等。近十余年来,全基因组关联研究(genome-wide association study, GWAS)在恶性肿瘤遗传易感性研究方面取得丰硕成果[2],其阐明的遗传易感位点可作为生物标志物用于构建多基因风险评分(polygenic risk score, PRS),进而预测恶性肿瘤发病风险,为高危人群的筛查和有效防治提供有力工具[3]。

PRS是指依据个体基因变异,通过计算多个易感位点的累积效应量化个体对疾病易感程度的一种评估工具。传统PRS的构建方法主要有5种,本课题组前期已对其进行了系统介绍[4],其中应用最广泛的是以OR值作为权重的PRS,该方法通常纳入GWAS报道的易感位点,以位点效应作为权重,计算个体携带的危险等位基因数目的加权总和[4]。此外,近几年兴起的诸如支持向量机、惩罚回归模型、神经网络、随机森林等机器学习算法也被应用于PRS的构建[5]。这些方法首先通过设定关联显著性

[基金项目] 国家自然科学基金(81922061,81973123)

*通信作者(Corresponding author),E-mail:hongxiama@njmu.edu.cn

标准以及连锁不平衡界值进行位点筛选，而后使用机器学习算法构建PRS。值得一提的是，来自哈佛大学流行病学系的研究者提出的LDpred算法在构建PRS方面效能优异。与机器学算法不同的是，该方法无须进行位点筛选，而是基于贝叶斯理论，以遗传变异的关联效应和连锁不平衡关系作为先验信息，推测遗传变异的“独立”效应，以此构建PRS[6]。目前，已有多篇研究显示PRS在优化疾病高危人群筛选标准、制定个体化筛查方案以及提高预防性干预人群的临床获益等方面具有较好的应用前景[3]。本课题组前期也已经对PRS在复杂性疾病预防中的应用现况及主要分析方法进行了阐述[7]。同时，研究者们对PRS的应用方法及应用范围也进行了进一步拓展和完善。

首先是方法学上取得诸多进展，PRS的构建不断被优化。早期PRS的构建一般仅使用GWAS发现的独立易感位点而忽略了基因组其他遗传变异对肿瘤发生的效应，由于位点少，变异影响小，有研究者认为其预测效能及价值有限。2018年，Khera等[8]提出了全基因组多基因风险评分（genome-wide polygenic score，GPS），通过适当放宽位点纳入标准，提高风险评分的预测效能。同时，随着基因组测序成本的下降，使用低通量全基因组测序（low coverage whole genome sequencing，lcWGS）技术开展人群基因组学研究已成为可能。相比于芯片分型技术，lcWGS成本相似，并且可以避免芯片设计过程中位点筛选所导致的偏倚，但预测效能是否具有优势尚不确定。Homburger等[9]分别基于lcWGS分型信息和芯片分型信息构建乳腺癌GPS，发现两者构建的GPS高度相关，GPS每增加1个标准差，乳腺癌发病风险增高1.56倍。该效应与基于芯片构建的GPS效应相近[10]。因此，lcWGS有可能取代芯片分型技术用于GPS的构建和疾病风险预测，然而使用lcWGS数据构建GPS的预测效能是否优于芯片分型技术仍需大样本量人群研究加以评估。

其次，PRS的应用对象逐渐由欧美人群扩展至其他种族人群。近十年来开展的GWAS中有67%仅纳入欧洲人群，另外19%是基于东亚人群，仅有近4%的GWAS在非洲、中东、拉丁美洲或印第安人中开展[11]。然而，由于不同种族人群之间遗传背景的差异，依据欧洲人群GWAS所构建的PRS在其他种族人群的预测效能均有所降低[11]。因此，亟需在多个种族人群开展大样本GWAS，由此构建PRS，并在相应种族的队列研究中评估其风险预测效能。最近，笔者课题组基于19个中国人群肺癌易感单核苷酸多态性（sigle nucleotide polymorphism，SNP）构建PRS，并利用中国慢性病前瞻性队列研究（China Kadoorie Biobank，CKB）评估其在肺癌风险预测方面的应用价值[12]。研究显示，相比于PRS最低的5%的人群，PRS处于前5%的人群罹患肺癌的风险增高137%。并且PRS能够在吸烟的基础上进一步优化肺癌风险分层。该研究对中国人群肺癌的风险预测和高危人群筛查具有较好的指导意义[12]。

此外，肿瘤不同亚型的PRS构建受到越来越多的关注。目前，肿瘤PRS研究大多基于同一种肿瘤总体的易感位点（不同亚型的集合体）进行构建，未能进行亚型的进一步区分。然而，同种恶性肿瘤的不同组织亚型在致病因素方面往往存在异质性，因此基于组织亚型特异的危险因素进行风险预测可以为肿瘤的个体化预防提供更准确的信息。鉴于此，Mavaddat等为雌激素受体（estrogen receptor，ER）阳性和ER阴性乳腺癌构建组织亚型特异的PRS。研究者分别使用总乳腺癌权重、亚型特异性权重、混合权重（即存在亚型异质性的位点使用亚型特异性权重，否则使用总乳腺癌权重）等方法构建PRS，结果显示混合权重PRS的预测效能最佳。进一步应用于英国女性人群，发现PRS最高和最低1%的人群罹患ER阳性乳腺癌的风险分别为2%和31%，罹患ER阴性乳腺癌的风险分别为0.55%和4%。研究者认为亚型特异的乳腺癌PRS不仅有助于优化乳腺癌筛查方案，还有助于指导高危人群的预防性用药[10]。

最后，综合应用基因与环境因素共同构建风险预测模型，有助于改善现有的肿瘤筛查指南，提高筛查的效益。以结直肠癌为例，美国预防服务工作组（USPSTF）建议年龄在50~70岁的人群接受结肠镜检查，同时加强具有结直肠癌家族史人群的筛查。然而，80%以上的患者并无结直肠癌家族史。鉴于此，Jeon等[13]基于19个生活方式和环境暴露因素（E-score）以及63个结直肠癌易感SNP（G-score）共同构建风险预测模型。研究结果显示，对于具有结直肠癌家族史的人群，E-score和G-score的综合评分（后面简称风险评分）处于前10%的人群推荐筛查年龄为男性40岁，女性46岁；而风险评分最低的10%人群推荐筛查年龄为男性51岁，女性59岁。而在没有结直肠癌家族史的人群中，风险评分处于前10%的人群在男性44岁、女性50岁即可考虑结肠镜筛查，风险评分最低的10%人群可分别将男性和女

性筛查年龄推迟到56岁和64岁[13]。这些发现为进一步完善结直肠癌的筛查方案和精准预防提供了重要依据。

综上,PRS的人群应用已取得一定进展,但仍存在诸多挑战。PRS所纳入的遗传变异大多是标签SNP,难以捕获致病变异的真实效应;且大多数肿瘤易感SNP位于非编码区,其生物学机制尚不明确。幸运的是,DNA元件百科全书(ENCODE)计划、Roadmap表观基因组学计划等项目的开展为破译非编码区SNP功能提供了新机遇。在此基础上开展精细作图研究,能够有效识别致病变异[15]。在未来的研究中,精细作图能否提高PRS的预测效能仍需进一步评估。同时,现有的大多数肿瘤风险预测模型尚未考虑基因-环境交互作用,环境因素会随着时间的迁移而变化,使交互作用分析更具挑战性。因此,仍需开发新的暴露测量方法、统计学算法以及模型评价方法,以有效识别基因-环境交互作用,同时将时间依赖性环境暴露纳入风险预测模型,并对模型的科学性和临床应用价值进行评估和验证。此外,目前研究往往使用曲线下面积(area under the curve,AUC)评价PRS的预测效能。AUC代表患者PRS高于非患者的概率,能够评价PRS区分患者与非患者的能力,但无法衡量个体(或特定人群)罹患肿瘤的绝对风险[16]。因此,目前大多数研究均停留在理论层面,PRS能否最终应用于恶性肿瘤的病因预防和筛查仍需通过人群试验和卫生经济学方法加以评估。

[参考文献]

[1] BRAY F, FERLAY J, SOERJOMATARAM I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA Cancer J Clin, 2018, 68(6): 394-424

[2] BUNIELLO A, MACARTHUR J A L, CEREZO M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019[J]. Nucleic Acids Res, 2019, 47(D1): D1005-D1012

[3] TORKAMANI A, WINEINGER N E, TOPOL E J. The personal and clinical utility of polygenic risk scores[J]. Nat Rev Genet, 2018, 19(9): 581-590

[4] 王 铖,戴俊程,孙义民,等. 遗传风险评分的原理与方法[J]. 中华流行病学杂志,2015,36(10):1062-1064

[5] HO D S W, SCHIERDING W, WAKE M, et al. Machine learning SNP based prediction for precision medicine[J]. Front Genet, 2019, 10: 267

[6] VILHJALMSSON B J, YANG J, FINUCANE H K, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores[J]. Am J Hum Genet, 2015, 97(4): 576-592

[7] 杭 栋,沈洪兵. 多基因风险评分与复杂性疾病风险预测和精准预防:机遇和挑战[J]. 中华流行病学杂志,2019,40(9):1027-1030

[8] 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[J]. Nat Genet, 2018, 50(9): 1219-1224

[9] HOMBURGER J R, NEBEN C L, MISHNE G, et al. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores[J]. Genome Med, 2019, 11(1): 74

[10] MAVADDAT N, MICHAILIDOU K, DENNIS J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes[J]. Am J Hum Genet, 2019, 104(1): 21-34

[11] DUNCAN L, SHEN H, GELAYE B, et al. Analysis of polygenic risk score usage and performance in diverse human populations[J]. Nat Commun, 2019, 10(1): 3328

[12] DAI J, LV J, ZHU M, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations[J]. Lancet Respir Med, 2019, 7(10): 881-891

[13] JEON J, DU M, SCHOEN R E, et al. Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors[J]. Gastroenterology, 2018, 154(8): 2152-2164

[14] SCHAID D J, CHEN W, LARSON N B. From genome-wide associations to candidate causal variants by statistical fine-mapping[J]. Nat Rev Genet, 2018, 19(8): 491-504

[15] 蒋 祝,孙 洁,蒋 涛,等. 中国人群冠心病相关基因的精细定位研究[J]. 南京医科大学学报(自然科学版),2019,39(5):756-761

[16] COOK N R. Use and misuse of the receiver operating characteristic curve in risk prediction[J]. Circulation, 2007, 115(7): 928-935

[收稿日期] 2020-02-23