ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

A Modular Pipeline for the Construction and Validation of Polygenic Risk Scores in Oncology

Pascucci Eleonora[1], *Galarducci Riccardo*[2], *Boccia Stefania*[1,2], *Pastorino Roberta*[1,2]

(1) Section of Hygiene, Department of Life Sciences and Public Health, Catholic University of Sacred Hearth, Rome, Italy
(2) Department of Woman and Child Health and Public Health, Fondazione Policlinico Universitario A. Gemelli IRCCS, Roma, Italia

CORRESPONDING AUTHOR: *Pascucci Eleonora, eleonora.pascucci@unicatt.it*

INTRODUCTION

Polygenic Risk Scores (PRS) are statistical tools designed to estimate individual predisposition to complex diseases by aggregating the effects of numerous genetic variants (Single Nucleotide Polymorphisms SNPs). In oncology, PRS hold promise for enhancing cancer risk stratification and personalizing screening strategies. However, their effectiveness depends on a well-defined computational framework that guarantees high-quality data processing and consistent predictive performance.

OBJECTIVE

This study aims to describe a modular and reproducible pipeline for the construction and validation of PRS in cancer research, detailing each analytical step from genotype preprocessing to risk score validation. Given that cancer is characterized by a highly polygenic architecture, involving thousands of loci with small effect sizes, such efforts require analytical workflows capable of handling complex and large-scale genomic data in a reliable and scalable manner.

METHODS

The pipeline begins with raw Variant Call Format (VCF) files obtained through genotyping. To ensure statistical power for detecting associations and constructing robust and accurate PRSs, large sample sizes, typically comprising several thousand cases and controls, are essential. Maintaining a balanced case-control ratio of 1:1 is crucial to minimize bias and maximize model stability. When relevant covariates are available, propensity score matching [1] can be applied to further balance cases and controls on clinical or demographic characteristics. Quality control (QC) is implemented using PLINK, a tool for handling SNP data, to remove variants and individuals based on call rate (<98%), minor allele frequency (MAF<1%), Hardy-Weinberg equilibrium deviations ($p<1\times10^{-4}$), excess heterozygosity or relatedness (PI_HAT>0.2), and sex discrepancies. Population stratification is assessed using Principal Component Analysis (PCA) or Multi-dimensional Scaling (MDS) to control for confounding due to population structure, and outliers are optionally detected via unsupervised clustering methods. The resulting components are included as covariates in downstream models. Imputation is performed via the Michigan [2] or Helmholtz imputation [3] servers using ancestry-matched reference panels to enhance the density of genotype data. Post-imputation filtering excludes SNPs with low imputation quality ($R^2 < 0.3$) and extreme allele frequencies to preserve dataset integrity. To identify genetic variants associated with cancer susceptibility, genome-wide association studies (GWAS) are conducted using logistic regression models, adjusting for age, sex, and leading principal components to mitigate confounding due to population substructure. When multiple cohorts are available, GWAS are initially conducted independently within each dataset. Subsequently, meta-analysis is performed to combine effect size estimates across studies, using either a fixed-effects or random-effects model. The choice of model depends on the extent of between-cohort heterogeneity, which may arise from differences in environmental exposures or other context-specific factors influencing cancer risk. In cases where such heterogeneity is minimal, fixed-effects meta-analysis via inverse-variance weighting is applied; otherwise, a random-effects model is employed to account for variability in genetic effect estimates across cohorts. For PRS construction, we employ a Bayesian regression framework with continuous shrinkage (PRS-CS) as proposed by Ge et al. [4], which integrates GWAS summary statistics with an external linkage disequilibrium (LD) reference panel to infer posterior SNP effect sizes. This approach eliminates the need to specify p-value thresholds or perform LD clumping and produces a single, optimized polygenic model. The PRS is finally calculated by summing allele dosages weighted by GWAS-derived effect sizes. Score performance is internally validated on a held-

DOI: 10.54103/2282-0930/29434

out portion of the original dataset and externally tested on independent cohorts. Evaluation metrics include the area under the receiver operating characteristic curve (AUC), R^2, and calibration plots.

RESULTS

We are currently applying this pipeline to the development and validation of a PRS for gastric cancer (GC) risk in individuals of European ancestry. Despite the growing use of PRS in various malignancies, only few of them have focused on GC, mostly on Asian individuals, and no validated PRS currently exists for GC in European populations. To address this gap, we are leveraging individual-level genotype data from over 8,000 GC cases and more than 350,000 controls across multiple European cohorts, including the Helsinki Biobank, the Rotterdam Study, dataset from Hess et al. [5], and the Spanish sample from the Stomach cancer pooling (StoP) Consortium. These cohorts form the discovery dataset used to conduct GWAS and meta-analysis, followed by PRS construction using a Bayesian framework (PRS-CS). The resulting scores are being externally validated in independent datasets from the UK Biobank and three cohorts from StoP consortium (Rome, Latvia, and Lithuania).

CONCLUSIONS

This pipeline provides a comprehensive and adaptable framework for constructing PRS in oncology, supporting methodological transparency and interoperability. Its modular design ensures flexibility across various datasets and facilitates implementation in clinical research. Future directions include increasing cross-ancestry portability and integrating PRS within clinical decision-making tools.

REFERENCES

1. Maekawa, M., Tanaka, A., Ogawa, M. et al., Propensity score matching as an effective strategy for biomarker cohort design and omics data analysis. Plos one, 2024, 19(5), e0302109.
2. https://imputationserver.sph.umich.edu
3. Das S., Forer L., Schönherr S. et al., Next-generation genotype imputation service and methods. Nature Genetics, 2016, 48, 1284–1287.
4. Ge T., Chen C. Y., Ni Y. et al. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nature communications, 2019, 10(1), 1776.
5. Hess T., Maj C., Gehlen J. et al. Dissecting the genetic heterogeneity of gastric cancer. EBioMedicine, 2023, 92.

DOI: 10.54103/2282-0930/29434