Expanding polygenic risk scores to include automatic genotype encodings and gene-gene interactions

This manuscript (permalink) was automatically generated from lelaboratoire/rethink-prs-ms@24f7157 on August 1, 2019.

Authors

- **Trang T. Le**☯
  0000-0003-3737-6565 · trang1618 · trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Hoyt Gong**☯
  0000-0001-9339-4763 · hoytgong · GongHoyt
  Life Sciences and Management, Wharton School, University of Pennsylvania

- **Patryk Orzechowski**
  0000-0003-3578-9809
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Elisabetta Manduchi**
  0000-0002-4110-3714
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Jason H. Moore**†
  0000-0002-5015-1099 · EpistasisLab · moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104 · Funded by National Institutes of Health Grant Nos. LM010098, LM012601, AI116794

☯ — These authors contributed equally to this work.

† — Direct correspondence to jhmoore@upenn.edu.

Abstract

Polygenic Risk Scores (PRSs) are aggregation of genetic risk factors of specific diseases and have been successfully used to identify subgroups of individuals who are more susceptible to those diseases. While severall studies have focused on identifying the correct genetic variants to include in PRS, most existing statistical models focus on the marginal effect of the variants on the phenotypic outcome but do not account for the effect of gene-gene interactions. Here, we propose a novel calculation of the risk score that expands beyond marginal effect of individual variants on the phenotypic outcome. The Multilocus Risk Score (MRS) method effectively selects alternative genotype encodings and captures epistatic gene-gene interactions by utilizing an efficient implementation of the model-based Multifactor Dimensionality Reduction technique. On a diverse, unbiased collection of datasets, MRS outperforms the standard PRS in the majority of the cases, especially when at least two-way interactions between genes are present. Our findings suggest that more precise models that take epistatic interactions into account are necessary and will yield greater utility for polygenic risk profiling.

Introduction

As the field of traditional genomics rapidly expands its sequencing technologies and translational abilities, novel applications of genomic data are starting to arise in addressing disease burden. Beginning with the completion of the Human Genome Project in 2003, increased interest in cataloguing genomic data spurred the innovation of massively parallel, chip-based genotyping arrays. Leveraging these technologies, early researchers were able to characterize and categorize gene variants across millions of individuals internationally. In particular, the advent of projects such as the International HapMap Project and the 1000 Genomes Project sought to document genetic variants involved in specific diseases of the human genome through genotyping and whole-genome sequencing techniques, respectively [1,2]. As such, the gross information of nucleotide polymorphisms within publicly available databases has rapidly increased in the beginning of the 21st century with the rise in omics sequencing capabilities. This genomic information, coupled with functional genomics data on the transcriptome, epigenome and metabolome, has been touted to further drive precision medicine approaches using genetics.

Complementing the rapid growth in our understanding of human genetic variation was the emergence of genome-wide association studies (GWAS) in the early 2000s to identify gene variants associated with common human diseases. Non-candidate-driven in design, these observational studies carry out chip array genotyping across population subsamples to subsequently assay for phenotype signal association via statistical approaches in silico. GWAS have primarily sought to discern genetic association with various phenotypes of interest by studying single nucleotide polymorphisms (SNPs) and other DNA variants across the human genome [3,4,5].

In tandem with the movement towards precision medicine, the post-GWAS era strives to bring significant population-derived gene variants into individual level metrics actionable in health delivery settings. While GWAS indeed capture gene variants associated with a phenotype of interest on a population level, translating such results to personalized individual metrics of risk requires aggregating contributions of many gene variants in the form of polygenic risk scores (PRS). Importantly, PRS provide an ability to explain inherited risk for disease in an individual by representing a weighted sum aggregate of risk alleles based on measured loci effect contributions derived from GWAS studies [6,7]. In quantifying the effect of particular combinations of genetic SNP variants towards risk prediction, PRS offers a probabilisitic susceptibility value of an individual to disease. Such genetic risk estimation scores are central to clinical decision-making, serving to reinforce individual health management in heritable disease detection and early prevention of various adult-onset conditions. The utility of PRS scores have been demonstrated in previous studies towards disease risk stratification across leading heritable causes of death in the developed world [10,11,8,9]

For each SNP j of an individual’s genome of n possible SNPs for analysis, the PRS score is calculated via a summation across k selected SNPs as

$$PRS = \sum_{j=1}^{k} \beta_j \cdot SNP_j$$

where β_j is the weighted risk contribution of the loci gene variant derived from risk score model parameters and SNP_j represents the number of variant alleles (0, 1, or 2) at the j^{th} SNP. Various approaches towards predicting risk of the same disease exist across PRS studies based on the above equation; models may vary according to the specific statistical model used to produce the weights

$beta_j$ for individual genetic variations, the k with respect to the specification of genetic variants considered, and the ability of the PRS to generalize to the entire population [6]. Historically, PRS models have previously characterized genomic architecture in a dichotimous division of Mendelian monogenic and polygenic inheritance, in which either one or many gene perturbations give rise to disease phenotypes in an individual, respectively [12,7].

However, while such former PRS models arose from older sequencing technology and study design, a more realistic genetic archtechture of common adult-onset disease acknowledges dynamic interactions among a continuum of common low-risk to rare high-risk gene variants to cumultatively drive overall risk of an individual [13]. When only considering rare (minor allele frequency, MAF < 0.5%), high-risk gene variants, such genomic variation only contributes approximately 1-10% towards adult-onset disease incidence [14,7]. Often, more relevant and complete sources of genetic risk is captured from complex smaller interactions of both common (MAF > 5%) and low-frequency (MAF > 0.5% and < 5%) genetic variants each contributing individual, appreciable effects [15,16]. Existing standard multivariate categorical data analysis approaches fall short in handling such enormous possible gene interaction combinations with both linear and nonlinear effects. In this context, more robust and efficient methods towards a polygeneic risk calculation are necessary in capturing the overlap between context-dependent effects of both rare and common alleles on human genetic disorder.

With respect to better understanding the epistasis across an individual's genome, various statistical models have been designed with the intent of capturing high dimensional gene-gene (GxG) interactions. The Multifactor Dimensionality Reduction (MDR) method is one such nonparametric, model-free framework that addresses these challenges and has been extensively applied to detect nonlinear complex GxG interactions associated with individual disease [17]. By isolating a specific pool of genetic factors from all polymorphism and cross-valiating prediction scores averaged across identified high risk multi-locus genotypes, the original MDR approach is able to categorize multi-loci genotypes, whether in the 1D or 2D, into two groups of risk based on some threshold value. While created with the primary intention towards GxG interaction detection by reducing dimensionality interatively in inferring genotype encodings, the MDR model has additionally demonstrated applicability as a risk score calculation model in constructing PRS scores [18].

Modifications built on top of the MDR framework have been proposed in order to better capture multiple significant epistasis models and potential missed interactions owning to limitations of the original model in the higher dimensions. Model-Based Multifactor Dimensionality Reduction (MB-MDR) was formulated as a flexible GxG detection framework for dichotomous traits and unrelated individuals [19]. Rather than a direct comparison against a threshold level in the original MDR method, MB-MDR merges multi-locus genotypes exhibiting significant High or Low risk levels through association testing and adds an additional 'No evidence of risk' categorization. In comparison to the standard MDR framework which reveals at most one optimal epistasis model, the MB-MDR method flexibly weighs multiple models by producing a model list ranked with respect to statistical parameters.

In the present work, we aim to reformulate the PRS using the MB-MDR approach to better capture epistatic gene interactions of individual disease risk in a novel Multilocus Risk Score (MRS). In observing prediction accuracy results on an evidence-based simulated dataset from HIBACHI, we demonstrate the improved performance of our multi-model weighted epistasis framework over existing PRS towards characterizing more granular disease etiology.

Methods

Multifactor Dimensionality Reduction (MDR) and model-based MDR (MB-MDR)

MDR is a nonparametric method that detects multiple genetic loci associated with a clinical outcome by reducing the dimension of a genotype dataset by pooling multilocus genotypes into high-risk and low-risk groups [17]. Extended from the original MDR algorithm, MB-MDR addresses existing limitations of MDR by increasing detectability of important interactions and decreasing bias by allowing O labels for individuals with no evidence for abberant risk. Several improvements have been made to MB-MDR since it was first introduced in 2009, and its current implementation efficiently and effectively detects multiple sets of significant gene-gene interactions in relation to a trait of interest while efficiently controlling type I error rates.

In addition to the test statistic and P values associated with each genotype combination, another important output of MB-MDR is the HLO matrices generated from the affected- and unaffected-subjects matrices (in the case of binary outcome). Briefly, for each genotype combination, an HLO matrix is a 3 x 3 matrix with each cell containing H (high), L (low) or O (no evidence), indicating risk of an individual whose genotype pairs fall into that cell [20]. For an example binary outcome problem, a genotype combination SNP_1 and SNP_2 will have a χ^2 value, an associated P value and an HLO matrix that looks like

	$SNP_1 = 0$	$SNP_1 = 1$	$SNP_1 = 2$
$SNP_2 = 0$	O	O	O
$SNP_2 = 1$	O	H	L
$SNP_2 = 2$	O	L	H

We discuss in the following subsection how these values were utilized in the formulation of the Multilocus Risk Score (MRS).

[More on the significance of SNP combination vs. significance of H/L/O here...]

Multilocus Risk Score (MRS)

We apply the MB-MDR software [20] v.4.4.1 to simulated datasets of n individuals, p SNPs to obtain the significance level of each combination of SNPs. We let k_d denote the number of significant combinations for a specific model dimension d (e.g. $d = 2$ results in pairs of SNPs). In this study, no significance threshold is imposed at the SNP combination level and, thus, k_d reaches its maximum value of C_p^d (p choose d).

For each subject i ($i = 1, 2, \cdots, n$), the d-way interaction risk score is calculated as

$$MRS_d(i) = \sum_{j=1}^{k_d} \chi_j^2 \times \text{HLO}_j(X_{ij})$$

where χ_j^2 is the test statistic of each genotype combination j from a χ_j^2 test with one degree of freedom for the simulated binary trait, X_{ij} is the j^{th} genotype combinations of subject i and HLO_j represents the j^{th} recoded HLO matrix (1 = High, -1 = Low, 0 = No evidence). In this study, we selected the default multiple testing correction algorithms for an MB-MDR model where the χ_j^2 for each genotype combination is derived from the gammaMAXT algorithm for two tests: H versus LO and L versus HO. As an example, consider a pair $X_{*j} = (SNP_{j_1}, SNP_{j_2})$ with $\chi_j^2 = 8.3$ and corresponding HLO matrix of all O's except an L in the first cell. Then, all subjects' current risks would remain the same except the ones with $SNP_{j_1} = SNP_{j_2} = 0$ where their risks are subtracted by 8.3.

In this study, we consider 1-way and 2-way interactions and thus the combined risk is simply the total of the first two dimensions: $MRS = MRS_1 + MRS_2$. We will examine the combined risk MRS and also its components, MRS1 and MRS2, separately.

Mutual information and information gain

For a given simulated data set, we apply entropy-based methods to measure how much information about the phenotype is due to either marginal effects or the synergistic effects of the variants after subtracting the marginal effects. A dataset's main effect (i.e. marginal effect ME) can be measured as the total of mutual information between each genotype SNP_j and the phenotypic class y based on Shannon's entropy H [21]:

$$ME = \sum_{j}^{k} I(SNP_j; y) = \sum_{j}^{k} H(y) - H(y|SNP_j).$$

We measure the 2-way interaction information (i.e. degree of synergistic effects of genotypes on the phenotype) of each dataset by summing the pairwise information gain between all pairs of genetic attributes. Specifically, if we let X_j denote the j^{th} genotype combination (SNP_{j_1}, SNP_{j_2}), the total 2-way interaction gain (i.e. synergistic effects SE) is calculated as

$$SE = \sum_{j}^{k} IG(X_j; y) = \sum_{j}^{k} I(SNP_{j_1}, SNP_{j_2}; y) - I(SNP_{j_1}; y) - I(SNP_{j_2}; y),$$

where IG measures how much of the phenotypic class y can be explained by the 2-way epistatic interaction within the genotype combination X_j. We refer the reader to Ref. [22] for more details on the calculation of the entropy-based terms.

Simulated data

The primary objective of this data simulation process was to provide a comprehensive set of reproducible and diverse datasets for the current study. Each dataset, containing 1000 samples (rows) and 10 SNPs (columns), was generated in the following manner. The values in the matrix were randomly assigned from potential options: with 1/2 probability of drawing '1' (representing heterozygous minor alleles Aa), 1/4 probability of drawing '2' (representing a major homozygous allele AA) and 1/4 of drawing '0' (representing homozygous major allele AA). The binary endpoint for the data was determined using a recently proposed evolutionary-based method for dataset generation called Heuristic Identification of Biological Architectures for simulating Complex Hierarchical Interactions (HIBACHI) [23]. This method uses Genetic Programming (GP) to build different mathematical and logical models resulting in a binary endpoint, such that the objective function called fitness is maximized. The unique feature of HIBACHI is explainability, as each generated model represents a formula for generating the endpoint. This formula is later approximated using classifiers. In this study, to arrive at a diverse collection of datasets, we aim to maximize the difference in predictive performance of all pairs of ten pre-selected classifiers from the extensive library of scikit-learn [24] (Table 1). *[Great point! I did comparison for 12 datasets, but later excluded SVC and LinearSVC. Corrected.]* Therefore, we define the fitness function of HIBACHI as the difference in accuracy between two classifiers. In other words, the first classifier was supposed to perform as well as possible on the data while the second as bad as possible.

Table 1. Selected machine learning methods and their parameters.

Algorithm	Parameters
GaussianNB	‘priors’: {None,2}, ‘var_smoothing’: {1e-9, 1e-7, 1e-5, 1e-3, 1e-1, 1e+1}
BernoulliNB	‘alpha’: {1e-3, 1e-2, 1e-1, 1., 10., 100.},‘fit_prior’: {True, False}
DecisionTree	‘criterion’: {“gini”, “entropy”}, ‘max_depth’: [1, 11], ‘min_samples_split’: [2, 21], ‘min_samples_leaf’: [1, 21]
ExtraTrees	‘n_estimators’: {50,100}, ‘criterion’: {“gini”, “entropy”},‘max_features’: [0.1, 1], ‘min_samples_split’: range(2, 21,2),‘min_samples_leaf’: range(1, 21,2), ‘bootstrap’: {True, False}
RandomForest	‘n_estimators’: {50,100}, ‘criterion’: {“gini”, “entropy”},‘max_features’: [0.1, 1], ‘min_samples_split’: range(2, 21,4), ‘min_samples_leaf’: range(1, 21,5), ‘bootstrap’: {True, False}
GradientBoosting	‘n_estimators’: {50,100},‘learning_rate’: {1e-3, 1e-2, 1e-1, 0.5, 1}, ‘max_depth’: {2,4,8}, ‘min_samples_split’: [2, 21], ‘min_samples_leaf’: [1, 21], ‘subsample’: [0.1, 1], ‘max_features’: [0.1, 1]
KNeighbors	‘n_neighbors’: [1, 100], ‘weights’: {“uniform”, “distance”} , ‘p’: {1, 2}
LogisticRegression	‘C’: {1e-3, 1e-2, 1e-1, 1., 10.}, ‘solver’ : {‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’}
XGBoost	‘n_estimators’=100, ‘max_depth’: [2, 10], ‘learning_rate’: {1e-3, 1e-2, 1e-1, 0.5, 1.}, ‘subsample’: [0.05, 1], ‘min_child_weight’: [1, 20],
MLPClassifier	: ‘hidden_layer_sizes’: {(10),(11),(12),(13),(14),(15),(10,5)}, ‘activation’: {‘logistic’,‘tanh’,‘relu’}, ‘solver’: [‘lbfgs’], ‘learning_rate’: {‘constant’,‘invscaling’}, ‘max_iter’=1000, ‘alpha’:np.logspace(-4,1,4)

HIBACHI was run for 50 iterations with population of 500 individuals representing pairs of classifiers. At each iteration, five randomly chosen settings for each of the classifiers (Table 1) were chosen among all potential options and served as hyperparameters for that method. Each of the settings was analyzed using 5-fold cross-validation and the best set of hyperparameters for each classifier was considered for comparison. The best settings out of 5 for each classifiers were compared and the settings that maximized the difference in the fitness were promoted. Each experiment in which one machine learning method was expected to outperform the other was repeated five times. The results of each pair of classifiers were later averaged.

For each simulated dataset, after randomly splitting the entire data in two smaller sets (80% training and 20% holdout), we built the MRS model on training data to obtain the χ^2 coefficients and the HLO matrix, and then we calculated risk score for each individual in the holdout set. We assess the performance of the MRS by comparing the area under the Receiving Operator Characteristic curve (auROC) with that of the standard PRS method on the holdout set.

Manuscript drafting

This manuscript is collaboratively written using the Manubot software which supports open paper writing via GitHub with Markdown [25]. Manubot uses continuous integration to monitor changes and automatically update the manuscript. As a result, the latest version of this manuscript is always available at https://lelaboratoire.github.io/rethink-prs-ms/.

Availability

Detailed simulation and analysis code needed to reproduce the results in this study is available at https://github.com/lelaboratoire/rethink-prs/.

Results

Datasets

[Patryk Orzechowski]

[simulated datasets of $n = 1000$ individuals, $p = 10$ SNPs]

MRS outperforms standard PRS in the majority of simulated datasets

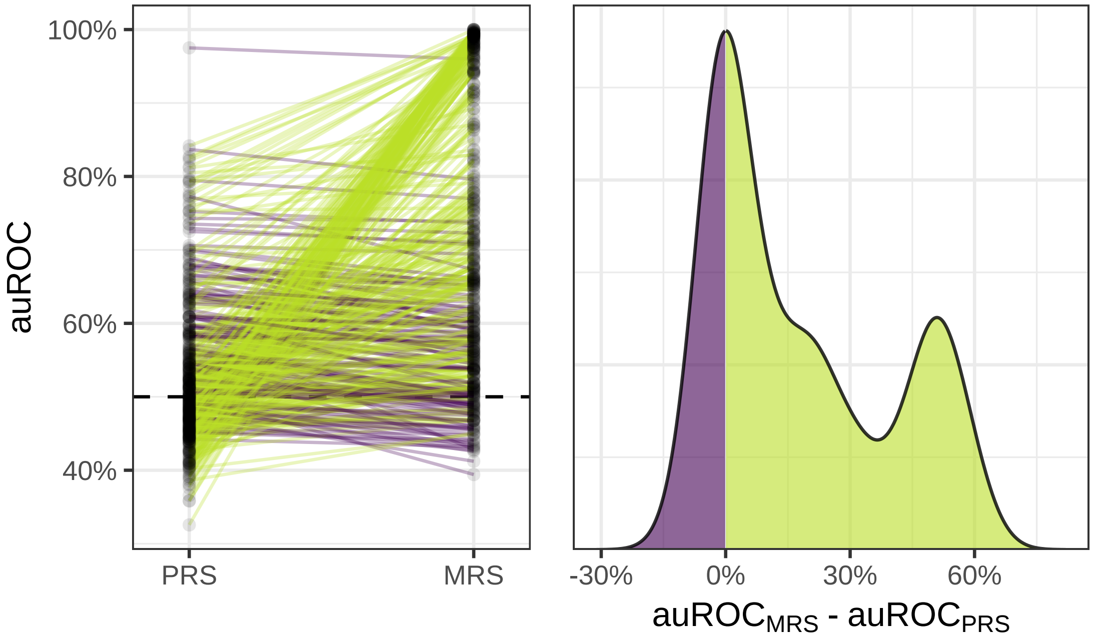

Figure 1: MRS produces improved auROC in the majority (335 green lines) of the 450 simulated datasets (each line represents a dataset). In many datasets, the standard PRS method performs poorly (auROC < 60%) while the new method yields auROC over 90%. This improvement in performance can be seen at the second peak (~50% auROC increase) in the density of the difference between the auROCs from the two methods (right).

In 335 out of 450 simulated datasets, MRS produces higher auROC compared to PRS (green lines, Fig. 1). In 363 datasets where the standard PRS method performs poorly (auROC < 60%), MRS performs particularly well (auROC > 90%) in 102 datasets. When MRS yields smaller auROC, the difference is small (3.3% ± 2.8%, purple lines/areas). Across all datasets, the improvement of MRS over PRS is significant (P < 10^{-15}) according a Wilcoxon signed rank test. To assess whether this improvement in performance correlates with the amount of interaction effects [contained] in each dataset, in the following section, we untangled the two components of MRS and test for the correlation between the difference in auROC and two entropy-based measures, main and interaction effect, of each dataset.

Assess improvement in performance

Individually, MRS1 and MRS2 both significantly outperformed the standard PRS method (both P values $< 10^{-15}$) according a Wilcoxon signed rank test. As the amount of main effects increases (Fig. 2 left column), MRS1 increasingly performs better than PRS, which is likely because encodings are inferred (top left). Meanwhile, MRS2's accuracy remain mostly similar to that of PRS (middle left). On the other hand, when the amount of interaction effects increases (Fig. 2 right column), MRS1 performs mostly on par to PRS while MRS2 increasingly performs better than PRS. Combining the gain from both MRS1 and MRS2, MRS's performance progressively increases compared to the standard PRS.

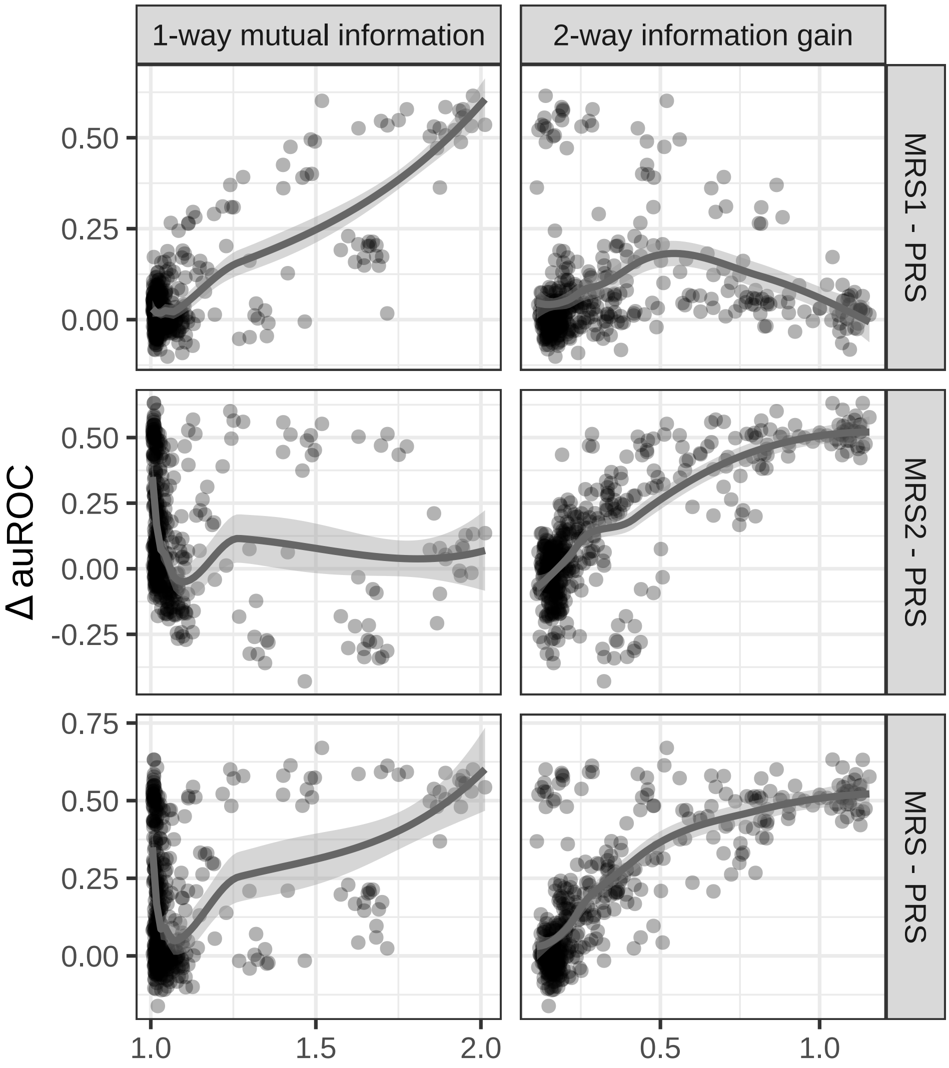

Figure 2: Combining 1-way (MRS1) and 2-way (MRS2) risk scores, MRS shows increasing outperformance to standard PRS as dataset contains more main and interaction effects.

Discussion

We introduce a Multilocus Risk Score (MRS) method to improve the performance of the standard PRS in disease risk stratification of patient populations. While PRS holds much promise for development of new precision medicine approaches by identifying high risk individuals who may benefit from prioritized interventions, one of its current limitations is the model simplicity [7]. As a first step towards addressing this issue and increasing comprehensiveness of risk profiling models, in this study, we developed a new applied MRS method from the MB-MDR framework that enables automatic genotype encodings and takes into account multiple models for detecting gene-gene (GxG) interactions. Utilizing the efficient implementation of MB-MDR, MRS automatically infers the genotype encodings and simultaneously computes the risk matrices of pairs of variants. Through comparing method performance on unbiased collections of simulated data, we demonstrate the robust polygenic risk profiling ability of MRS and suggest the importance of flexible, precise methods in better capturing epistasis behind individual patient risk.

We showed that the MRS method outperformed standard PRS in many of the simulated datasets, highlighting the importance of genotype encodings and consideration of epistasis. We further examined the association between this improvement and the amount of two-way epistatic effect induced in the binary phenotypic outcome. Appropriate phenotype encodings are important for improving the accuracy when there is a large amount of main effects of the variants on the phenotypic outcome. Meanwhile, inclusion of epistatic terms significantly increases the accuracy from PRS, especially when two-way interactions are present in the data. Although we only considered up to two-way GxG interactions, it is straightforward to incorporate higher order interactions (e.g. three-way, four-way) into MRS. However, preliminary analyses on the simulated datasets for such higher order interactions did not show significant improvement from the current MRS (results not shown). We also recommend estimating the computational expense prior to implementing high order interactions, especially for larger datasets encountered in practice.

Although MRS captures the improvements of MB-MDR in reporting polygenic risk profiles, there are three primary limitations. First, MRS has not been applied to real-world data. Although we compensated the lack of real data with a diverse, unbiased set of simulated datasets, a future study analyzing will prove beneficial to quantify the new MRS model’s utility in practice. Second, accounting for epistasis, in principle, is largely more computationally expensive compared to investigating solely main effects. Therefore, even with fast and efficient softwares, pre-selecting the variants (e.g. based on specific pathways or prior knowledge) will prove beneficial for accurate MRS computing when analyzing datasets containing a larger number of variants. However, we hope the promising preliminary results will open the doors to future approaches that encompass main and interaction effects and improve scalability.

Finally, we caution that a risk score model should be evaluated based on not only sensitivity and specificity but also with respect to potential clinical efficacy, and any genetic risk should be interpreted in aggregate with other risk factors. Future works focusing on gene-environment interactions with time-dependent risk factors will be crucial in order to communicate risk properly for preventive interventions.

In conclusion, MRS enhances the predictive capacity of current risk profiling model for complex diseases with polygenic architectures. While there is much work left to do to improve the personal and clinical utility of general risk profiling framework, we highlight that more comprehensive models that infer proper genotype encodings and account for epistatic effects greatly improve the prediction efficiency and affords new opportunities for more accurate clinical prevention.

Acknowledgement

We thank Dr. Kristel Van Steen and Aldo Camargo for their helpful responses to our inquires about the MB-MDR software.

References

1. **The International HapMap Project** *Nature* (2003-12) https://doi.org/dgd
DOI: 10.1038/nature02168 · PMID: 14685227

2. **An integrated map of genetic variation from 1,092 human genomes** *Nature* (2012-10-31) https://doi.org/f4k2v2
DOI: 10.1038/nature11632 · PMID: 23128226 · PMCID: PMC3498066

3. **Chapter 11: Genome-Wide Association Studies**
William S. Bush, Jason H. Moore
PLoS Computational Biology (2012-12-27) https://doi.org/gfr9pz
DOI: 10.1371/journal.pcbi.1002822 · PMID: 23300413 · PMCID: PMC3531285

4. **Genome-wide association studies for common diseases and complex traits**
Joel N. Hirschhorn, Mark J. Daly
Nature Reviews Genetics (2005-02) https://doi.org/bhcc36
DOI: 10.1038/nrg1521 · PMID: 15716906

5. **Genome-wide association studies: theoretical and practical concerns**
William Y. S. Wang, Bryan J. Barratt, David G. Clayton, John A. Todd
Nature Reviews Genetics (2005-02) https://doi.org/fcqz33
DOI: 10.1038/nrg1522 · PMID: 15716907

6. **What Are Polygenic Scores and Why Are They Important?**
Leo P. Sugrue, Rahul S. Desikan
JAMA (2019-05-14) https://doi.org/gfx8wg
DOI: 10.1001/jama.2019.3893 · PMID: 30958510

7. **The personal and clinical utility of polygenic risk scores**
Ali Torkamani, Nathan E. Wineinger, Eric J. Topol
Nature Reviews Genetics (2018-05-22) https://doi.org/gd46bh
DOI: 10.1038/s41576-018-0018-x · PMID: 29789686

8. **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder** *Nature* (2009-07-01) https://doi.org/cb5f2j
DOI: 10.1038/nature08185 · PMID: 19571811 · PMCID: PMC3912837

9. **Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting**
Pradeep Natarajan, Robin Young, Nathan O. Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, Valentin Fuster, Dermot F. Reilly, Adam Butterworth, … Sekar Kathiresan
Circulation (2017-05-30) https://doi.org/gbgqhb
DOI: 10.1161/circulationaha.116.024436 · PMID: 28223407 · PMCID: PMC5484076

10. **Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States**
Paige Maas, Myrto Barrdahl, Amit D. Joshi, Paul L. Auer, Mia M. Gaudet, Roger L. Milne, Fredrick R. Schumacher, William F. Anderson, David Check, Subham Chattopadhyay, … Nilanjan Chatterjee
JAMA Oncology (2016-10-01) https://doi.org/gf4w2z
DOI: 10.1001/jamaoncol.2016.1025 · PMID: 27228256 · PMCID: PMC5719876

11. **Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts**
Tyler M Seibert, Chun Chieh Fan, Yunpeng Wang, Verena Zuber, Roshan Karunamuni, J Kellogg Parsons, Rosalind A Eeles, Douglas F Easton, ZSofia Kote-Jarai, Ali Amin Al Olama, …
BMJ (2018-01-10) https://doi.org/gcsrzs
DOI: 10.1136/bmj.j5757 · PMID: 29321194 · PMCID: PMC5759091

12. **Beyond Mendel: an evolving view of human genetic disease transmission**
Jose L. Badano, Nicholas Katsanis
Nature Reviews Genetics (2002-10) https://doi.org/cjfzf2
DOI: 10.1038/nrg910 · PMID: 12360236

13. **The continuum of causality in human genetic disorders**
Nicholas Katsanis
Genome Biology (2016-11-17) https://doi.org/f9fzd3
DOI: 10.1186/s13059-016-1107-9 · PMID: 27855690 · PMCID: PMC5114767

14. **Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits**
Zhihong Zhu, Andrew Bakshi, Anna A.E. Vinkhuyzen, Gibran Hemani, Sang Hong Lee, Ilja M. Nolte, Jana V. van Vliet-Ostaptchouk, Harold Snieder, Tonu Esko, Lili Milani, … Jian Yang
The American Journal of Human Genetics (2015-03) https://doi.org/f64772
DOI: 10.1016/j.ajhg.2015.01.001 · PMID: 25683123 · PMCID: PMC4375616

15. **Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data**
Huwenbo Shi, Gleb Kichaev, Bogdan Pasaniuc
The American Journal of Human Genetics (2016-07) https://doi.org/f8xxkd
DOI: 10.1016/j.ajhg.2016.05.013 · PMID: 27346688 · PMCID: PMC5005444

16. **Common SNPs explain a large proportion of the heritability for human height**
Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, … Peter M Visscher
Nature Genetics (2010-06-20) https://doi.org/fjjm4v
DOI: 10.1038/ng.608 · PMID: 20562875 · PMCID: PMC3232052

17. **Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer**
Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi, L. Renee Bailey, William D. Dupont, Fritz F. Parl, Jason H. Moore
The American Journal of Human Genetics (2001-07) https://doi.org/bh3x75
DOI: 10.1086/321276 · PMID: 11404819 · PMCID: PMC1226028

18. **Risk score modeling of multiple gene to gene interactions using aggregated-multifactor dimensionality reduction**
Hongying Dai, Richard J Charnigo, Mara L Becker, J Steven Leeder, Alison A Motsinger-Reif
BioData Mining (2013-01-08) https://doi.org/gb5fmn
DOI: 10.1186/1756-0381-6-1 · PMID: 23294634 · PMCID: PMC3560267

19. **Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data**

Jestinah M Mahachie John, François Van Lishout, Kristel Van Steen
European Journal of Human Genetics (2011-03-16) https://doi.org/bndfnk
DOI: 10.1038/ejhg.2011.17 · PMID: 21407267 · PMCID: PMC3110049

20. **An efficient algorithm to perform multiple testing in epistasis screening**
François Van Lishout, Jestinah M Mahachie John, Elena S Gusareva, Victor Urrea, Isabelle Cleynen, Emilie Théâtre, Benoît Charloteaux, Malu Luz Calle, Louis Wehenkel, Kristel Van Steen
BMC Bioinformatics (2013-04-24) https://doi.org/f4v3n7
DOI: 10.1186/1471-2105-14-138 · PMID: 23617239 · PMCID: PMC3648350

21. **A Mathematical Theory of Communication**
C. E. Shannon
Bell System Technical Journal (1948-07) https://doi.org/b39t
DOI: 10.1002/j.1538-7305.1948.tb01338.x

22. **Epistasis Analysis Using Information Theory**
Jason H. Moore, Ting Hu
Methods in Molecular Biology (2014-11-03) https://doi.org/gf484q
DOI: 10.1007/978-1-4939-2155-3_13 · PMID: 25403536

23. **A heuristic method for simulating open-data of arbitrary complexity that can be used to compare and evaluate machine learning methods**
Jason H. Moore, Maksim Shestov, Peter Schmitt, Randal S. Olson
Biocomputing 2018 (2017-11-17) https://doi.org/gf5b6s
DOI: 10.1142/9789813235533_0024

24. **Scikit-learn: Machine Learning in Python**
Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, … Édouard Duchesnay
Journal of Machine Learning Research (2011-10) https://hal.inria.fr/hal-00650905

25. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653