Comparison of predictive abilities and model performance of biotx.ai's algorithmic pipeline with Polygenic Risk Scores for large GWAS (n > 10,000)

Abstract / Executive Summary

Polygenic Risk Scores (PRS) are the industry and research standard for predicting traits based on genetic data, when taking into account variation in multiple genetic variants. For each individual a score is calculated based on the variants associated with a trait and their respective impact on the trait. A previous white paper compared PRS to biotx.ai, a novel approach that uses data-mining of contextual information to create polygenic disease models. We showed that biotx.ai outperforms PRS in terms of ROC curve, explanatory value and the amount of data needed for analysis. The GWAS used in the previous white paper was small (471 subjects). In the present work we show that the advantages of biox.ai over PRS hold for large GWAS (>10,000 subjects).

Comparison PRS / biotx.ai

Polygenic Risk Scores

In human genetics Polygenic Risk Scores are calculated by computing the sum of risk alleles corresponding to a phenotype of interest in each individual, weighted by the effect size estimate of the most powerful GWAS on the phenotype. Studies have shown that substantially greater predictive power can be achieved by using PRS rather than a small number of genome-wide significant SNPs (Choi et al., 2018). Even though PRS compound a large number of SNPs, the score is entirely additive, that is each SNPs contribution to the score is considered separately, complex interactive patterns are not taken into account (see Box 1 below).

Box 1. Additive versus interactive modeling

The name 'Polygenic Risk Score' itself suggests that PRS models polygenic interactive effects. This is, however, not the case. In PRS, the separate effects of a large number of SNPs are added into one risk score. This does not capture any true interactions in which the effect of a SNP is contingent on the presence of one or more other variants, e.g. *rs_XXX (G/G)* leads to an increased risk for developing Diabetes, but only if *rs_YYY (A/C)* and *rs_ZZZ (T/T)* are present.

Biotx.ai

Biotx.ai consists of two complementary modules that are in feedback with each other. The contextual module uses information mined from the scientific literature, pathway libraries and protein co-expression data and an evaluation module that estimates predictive power of a feature based on that contextual information with extreme computational efficiency. When testing for monogenic effects the number of features equals the number of SNPs, but when testing for polygenic interactions, the number of features is the number of SNPs exponentiated with the complexity of the interaction one tests for. This can easily lead to a feature space that is so large that it becomes unmanageable. A combination of using the biological contextual information and computationally efficient algorithms (over 80.000 times as fast as logistic regression) allows for a massive reduction of the feature space and time required to scan it for association hypotheses.

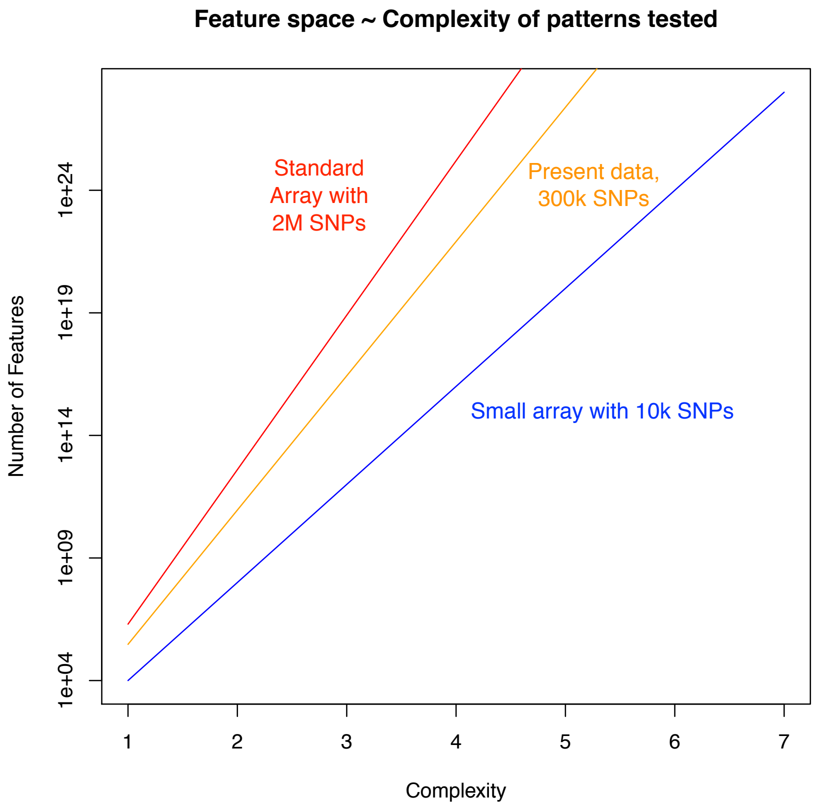

Figure 1. Number of features as a function of array size (number of variants) and complexity of the interactions tested for. Testing for complex patterns with standard methods leads to an unmanageably large feature space even for small arrays.

The final result is a manageable set of polygenic hypotheses, each containing up to 7 SNPs that, in interaction, better predict a given phenotype. In contrast to PRS, which merely adds the effects of different SNPs, biotx.ai models true interactions (see Box 1).

Table 1. Key differences of Polygenic Risk Scores (PRS) and biotx.ai

Polygenic Risk Scores	**biotx.ai**
Effects of multiple SNPs are only accounted for only by addition	True modeling of interactions
Can use summaries from previous GWAS (which results in overfitting)	Incorporates contextual information and previous research via NLP
Uses p-value threshold for inclusion of SNPs	True predictive power calculated via bootstrapping over validation data
Tens of thousands of SNPs required for predictions, functional relations are opaque	Produces concise polygenic term with verifiable functional insights

Data

The Mega-GWAS ALS I data (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000101.v5.p1, **dbGaP Study Accession:** phs000101.v5.p1) was used for the comparison. The data set contains 12319 subjects (7015 males, 5214 females, 90 ambiguous) and 471303 genotypes variants.

The dataset was submitted to a series of quality control steps that aim to removal of biases in the downstream analysis. In a first stage we filtered SNPs and individuals based on their missingness in the dataset. This ensures that SNPs are excluded that have a high proportion of subjects where genotyping information was unavailable or of poor quality. Similarly, individuals where a large proportion of SNPs could not be measured were excluded. This was achieved in to two steps where first a lenient threshold of 0.2 (i.e. >20%) was applied to remove the clear outliers, followed by a more stringent threshold of 0.02 (i.e. >2%). SNP filtering was performed before individuals were filtered. Next all variants not on autosomal chromosomes were removed. Next variants that deviate from the Hardy-Weinberg equilibrium were removed in a two-step process whereby we first applied a lenient threshold of 1e-6 followed by a more stringent threshold of 1e-10. This is a common indicator of genotyping errors. Thereafter individuals were filtered out based on their heterozygosity rates which can indicate sample contamination. Individuals deviating by more than 3 standard deviations from the mean of the rate from all samples were filtered out. To assess the heterozygosity rate per sample, those variants that were in linkage disequilibrium with each other were extracted by scanning the genome at a window size of 50 variants, step size of 5 and a pairwise correlation threshold of 0.2. Next related individuals were removed. To achieve this their identity by descent coefficients (IBD) was calculated and only one individual per related cluster would be kept. A total of 93,591 variants and 816 samples were removed at this stage. Next the small proportion of missing genotypes were imputed where and 11 individuals where this was not possible were removed. Finally the population structure of the samples was analyzed in two stages to identify internal stratifications, which was used to filter out any individuals not predicted to close enough to a European reference cohort (175 samples were removed).

Finally the dataset consisted of 11316 subjects (9354 cases, 1962 controls) that was split into training, validation and test sets. The exact same sets were used for both methods.

Method

For PRS results of a GWAS conducted with SAIGE (Zhou et al., 2018) were used to calculate PRS for 9933 different p-value thresholds for the subjects in the training, validation and test set. The PRS of the subjects in the training set was then used to train a separate logistic regression classifier for each p-value threshold. The validation data set was used to determine which p-value threshold produces the best classifier and this classifier was then used to predict the test set. This classifier is based on the PRS of 83598 different SNPs.

For biotx.ai the training set was used to generate and filter hypothesis. After 75000 iterations the most performant under 500 polygenic hypotheses were obtained. The validation set was then used to further reduce this to less than 100 hypotheses. The remaining hypotheses were summarized in a term that was used to train a LASSO (Tibshirani, 1996, Hastie and Qian, 2014) regression model on the training data. This model, based on less than 80 SNPs in interaction, then predicted the test set.

Results

The Receiver Operating Characteristics (ROC) were used to evaluate the quality of prediction of both approaches. Biotx.ai Area under Curve is 19 perecentage points (34%) larger than that of the PRS with Sensitivity and Specificity being 11 percentage points lower (-14%) and 32 percentage points (78%) higher at the Youden Point.

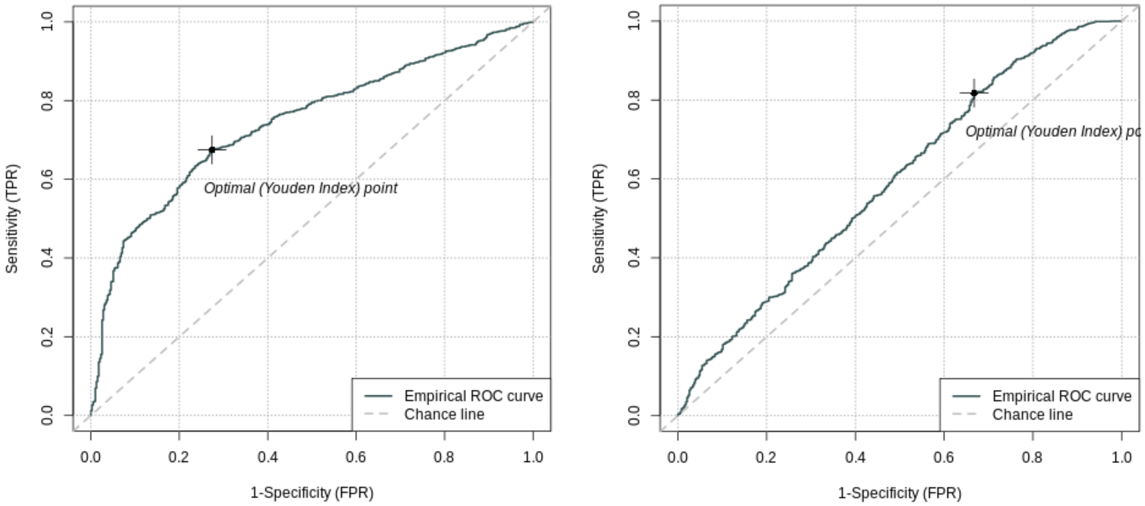

Figure 2. ROC Curve for biotx.ai (left) and PRS (right)

Table 2. Key Statistics for both methods at the optimal (Youden Index) point

Method	AUC	Sensitivity	Specificity
biotx.ai	0.75	0.674	0.73
PRS	0.56	0.78	0.41

Due to the imbalanced sample (9354 cases, 1962 controls), accuracy and misclassification are not good metrics for model evaluation. A model that predicts each subject to be a case would have an accuracy of 0.83, but no scientific value – this is one of the reasons why ROC is used within the machine learning community and sensitivity and specificity within the medical community.

Readers of the previous white paper will notice that the gap between the performance of biotx.ai and PRS is slightly smaller in the present research (19 vs. 26 percentage points, 34% vs. 46%). One possible explanation for this would be that the large sample size is more suited for PRS, which reduces the gap. This explanation can, however, be rejected, because the change in the gap is not due to PRS performing better on the big data, but due to biotx.ai performing slightly worse than on the previous data. Biotx.ai's ability to detect complex polygenic patterns means that the approach performs better when large portions of the variance are due to such polygenic interactions. In the absence of such polygenic interactions, the advantage diminishes. The slight difference in performance compared to the PPMI data can thus be bests explained by the less polygenic nature of the present ALS data.

Beyond prediction biotx.ai also provides insights about the genes associated with the disease. Due to its ability to detect true interactions between SNPs, at least 49 SNPs were associated with the disease that would not have shown any significance in a purely additive approach like PRS.

Conclusion

On this large GWAS of 12319 subjects, biotx.ai outperforms PRS by 19 percentage points or 34%. This shows that the advantages of biotx.ai's approach over PRS are not limited to small data sets (for which PRS is, admittedly, not suited). Rather, biotx.ai's ability to detect complex patterns of SNPs that impact a disease offer a general advantage. This is not surprising, because many diseases are polygenic, i.e. are caused by the complex interactions of many different SNPs. This means that the biotx.ai approach is able to associate new variants with the diseases that would have not shown up under an additive approach such as PRS. PRS models also require a large set of SNPs which leads to overfitting and limits their use in clinical practice. Biotx.ai generates more parsimonious models which do not have such limitations.

References

Choi, S. W., Mak, T. S. H., & O'reilly, P. (2018). A guide to performing Polygenic Risk Score analyses. *BioRxiv*, 416545.

Hastie, T., & Qian, J. (2014). Glmnet vignette. *Retrieve from http://www web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf. Accessed October*, *30*, 2019.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

Wei Zhou, Jonas B. Nielsen, Lars G. Fritsche, Maiken B. Elvestad, Brooke Wolford, Maoxuan Lin, Kristian Hveem, Hyun Min Kang, Goncalo R. Abecasis, Cristen J. Willer*, Seunggeun Lee* (2018) *Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies.* Nature Genetics 50, 1335–1341

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.