Challenges of Predicting Individual Risk in Genome-Wide Association Studies

Agnes M Nielsen* Helle K Pedersen† Ramneek Gupta†
Line H Clemmensen*

Abstract

Genome-wide association studies (GWASs) traditionally concern the relations between genetic variants known as SNPs and traits such as diseases. However, there has recently been a shift in focus towards clinical translation and personalized medicine making it relevant to also consider risk prediction approaches. In this study, a GWAS cohort containing ∼3000 type 2 diabetes cases and ∼3000 non-diabetic controls is analyzed with this shift in mind. A traditional univariate association study as well as risk prediction from traditional logistic regression and the non-linear machine learning algorithm random forest are studied. In contrast to findings of genome-wide significant associations, the predictive performance is not necessarily aided significantly by the information carried in the SNPs. In this paper, we discuss the statistical challenges in transitioning from associations on a population scale to prediction for individuals. This will be done in the context of GWA data in which factors like a limited number of observations and a large number of variables each with low effect-sizes make individual prediction a challenging problem.

Key Words: Genome-wide association study, random forest, prediction

1. Introduction

The relations between genetic variants called single nucleotide polymorphisms (SNPs, pronounced "snips") and biological traits such as diseases are traditionally considered by genome-wide association studies (GWASs). A SNP is a change in a single base-pair in the DNA that occurs often (typically >1%) in a population. The studies typically consider whether each SNP is univariately associated with a trait [1]. However, there has recently been a shift in focus towards clinical translation and personalized medicine. This makes it relevant to consider prediction from genetics such that personal risk can be ascertained.

Risk prediction from GWAS will here be considered in the example of type 2 diabetes. It is a metabolic disease, which can be delayed or prevented by lifestyle changes making it important to identify high risk individuals early from genetics. Risk SNPs for type 2 diabetes have previously been identified by univariate association [2]. We will perform risk prediction and discuss the challenges in going from associations at a population level to individual predictions in this context.

We have also performed a study of simulated SNP data to further illustrate the challenges of moving from association to prediction.

2. Methods

In this study, the GENEVA cohort was analyzed. It contained 6,033 observations of 909,622 SNPs and was obtained through the database of Genotypes and Phenotypes (dbGaP) (study accession phs000091.v2.p1) [3]. The data originated from

*Department of Applied Mathematics and Computer Science, Technical University of Denmark
†Department of Bio and Health Informatics, Technical University of Denmark

two studies: The Nurse Health Study (NHS) containing only women and the Health Professionals Follow-up Study (HPFS) containing only men. The individuals were genotyped on the Affymetrix 6.0 platform and the data was pre-processed using standard quality control filters: Excluded SNPs with >5% missing values, SNPs with minor allele frequency <5%, SNPs which deviate from the Hardy-Weinberg equilibrium (p<0.0001) and individuals who were <98% genotyped. After pre-processing it consisted of 5,827 individuals (observations), who had both phenotype and genotype information and either no diabetes or type 2 diabetes. After pruning, the data set was comprised of 652,400 SNPs and 20 clinical variables, of which age, sex and bmi were used.

Three subsets of SNPs were tested: i) SNPs previously associated with type 2 diabetes, ii) SNPs associated with type 2 diabetes in this cohort with genome-wide significance, and iii) SNPs identified for type 2 diabetes relevance using random forests. Known type 2 diabetes SNPs were taken from Morris et al. (2012), where 56 (of the 62 reported SNPs) were directly genotyped in the GENEVA cohort or could be represented by proxy SNPs. A proxy SNP is highly correlated with the identified SNP on a population scale. They were found using SNAP Proxy Search with HapMap (release 22), CEU population and $R^2 \leq 0.5$ [4]. To identify SNPs associated with type 2 diabetes a traditional univariate association analysis was performed where SNPs a were tested using logistic regression models adjusted for age and sex in PLINK [5, 6]. The SNPs with p-values below the genome-wide significance threshold ($p < 5 \cdot 10^{-8}$) were taken forward in the analyses. The last subset of SNPs was identified using random forests [7]. Here, the SNPs were selected by iteratively reducing the number of SNPs and creating new forests with the SNPs ranking highest according to the permutation importance (Figure 1). 50 SNPs were selected for the prediction. The result was three sets of selected SNPs whose performance can be compared to each other.

The predictions were performed with logistic regression and random forest using either known diabetes SNPs, SNPs identified using random forest, genome-wide significant SNPs or without SNPs. Each set of SNPs was included over two sets of clinical variables: age and sex, and age, sex and bmi. This means $2 \times 4 \times 2 = 16$ models were tested.

The process was cross-validated using 10-fold cross-validation and the predictions were evaluated using receiver operating characteristic (ROC) curves and the area under the curve (AUC).

2.1 Simulation Study

A simulation study was performed. First a data set resembling GWA data was simulated consisting of 6,000 observations and 15,000 explanatory variables ("SNPs"). The variables were simulated to take the values 0, 1 and 2 such that the distributions of the alleles were close to the Hardy-Weinberg equilibrium. They were simulated in highly correlated blocks of size 1 to 40 variables. A linear combination of 30 randomly selected variables was made with the parameter for each variable simulated to have odds ratios between 1/1.3 and 1.3, which are close to the odds ratios expected in a GWAS. The linear combination was transformed using the logistic function yielding values between 0 and 1. These were used for probabilities for binomial samples for the response variable. Of the 6,000 observations, 600 were used for testing the prediction and the remaining 5,400 were used for selecting the significant variables and training the prediction model. The significant variables

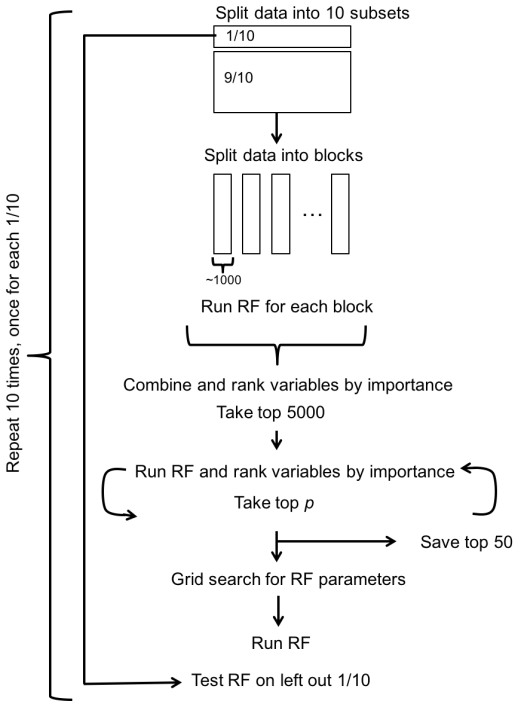


Figure 1: Diagram of the applied variable selection approach using random forest. The columns represent variables and rows observations [8]. RF is random forest and p is the number of variables selected in the iterative selection. The variable p takes the values 4000 down to 1000 with a jump of 1000 and then reduced by 100 until $p = 100$ is reached.

($p < 5 \cdot 10^{-8}$) were used for prediction using a multivariate logistic regression. This was repeated 10 times.

3. Results

We performed a traditional univariate association study as well as risk prediction using traditional logistic regression and random forest, a non-linear machine learning algorithm. In this cohort, genome-wide significant associations were found for five SNPs of which four are present in the data set used for prediction (plotted above the threshold-line in Figure 2). 239 SNPs were selected using random forest, 50 in each of the 10 folds. These are highlighted with green in Figure 2.

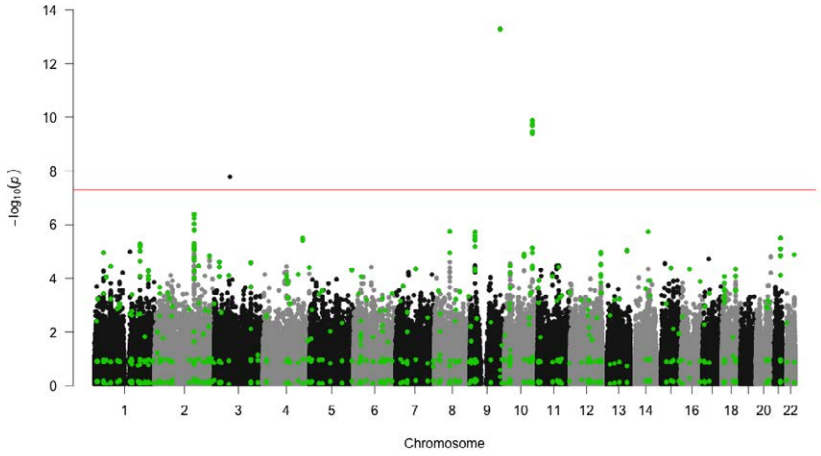


Figure 2: Manhattan plot. Each point is the p-value for a SNP plotted by its chromosome location. The green marks the SNPs selected using random forest. The red line indicates the genome-wide significant threshold. Consequently, the five SNPs above this line are said to be genome-wide significant.

In Figure 3, the ROC curves for the predictions are shown. The two curves close to the diagonal (A) are for linear regression and random forest, respectively, only including age and sex of the individuals. They predict diabetes poorly and have an AUC close to 0.5. The next group of curves from the diagonal (B) are the models including age, sex and one of the sets of SNPs: Previously identified SNPs, newly identified using random forest or genome-wide significant SNPs. The inclusion of the SNPs improve the prediction over only age and sex (Table 1). The last group of curves are models including age, sex and bmi (C). For these models the inclusion of SNPs does not improve the prediction and the curves are all close to each other and with similar AUCs (Table 1).

3.1 Simulation Study

The 10 simulation runs have each resulted in between 0 and 77 significant variables (mean of 24.9 with standard deviation 28.9). These are used for prediction of the test sets with logistic regression giving an average AUC of 0.5705 with standard deviation 0.0387. The resulting ROC curves are shown in Figure 4.

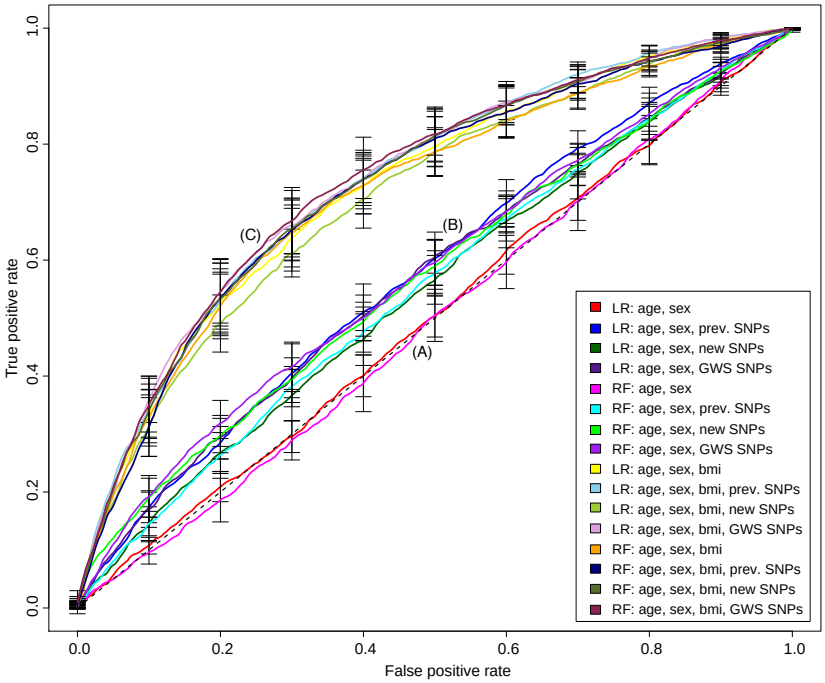


Figure 3: Average ROC curves from the cross-validation of the models including age, sex and in some bmi, genome-wide significant SNPs (GWS SNPs), newly identified SNPs using random forest (new SNPs) or previously identified SNPs (prev. SNPs). Here LR is logistic regression and RF is random forest. For each of the sixteen models bars indicate one standard deviation around the mean from the cross-validation. The dashed black line indicates the diagonal (indicating a non-informative model with random performance).

Table 1: Mean and standard deviation of AUC over 10-fold cross-validation for models including clinical variables and previously identified SNPs (prev.), newly identified SNPs by the random forest method (new) or genome-wide significant SNPs (GWS). The newly identified SNPs are the 50 SNPs identified within each fold. P-values are from a Welch two-sample test were the model is compared to logistic regression with only phenotype variables (age and sex alone or together with bmi).

	Logistic Regression		Random Forest	
	mean AUC (s.d)	p-value	mean AUC (s.d)	p-value
age + sex	0.5053 (0.0204)		0.4974 (0.0262)	0.4651
+ prev. SNPs	0.5783 (0.0215)	3.695×10^{-7}	0.5541 (0.0185)	2.665×10^{-5}
+ new SNPs	0.5500 (0.0284)	0.0009	0.5709 (0.0179)	5.156×10^{-7}
+ GWS SNPs	0.5707 (0.0133)	3.248×10^{-7}	0.5792 (0.0192)	1.366×10^{-7}
age + sex + bmi	0.7264 (0.0313)		0.7178 (0.0273)	0.5229
+ prev. SNPs	0.7355 (0.0261)	0.4581	0.7257 (0.0305)	0.9931
+ new SNPs	0.7091 (0.0295)	0.2195	0.7309 (0.0265)	0.7332
+ GWS SNPs	0.7375 (0.0286)	0.4174	0.7387 (0.0263)	0.3546

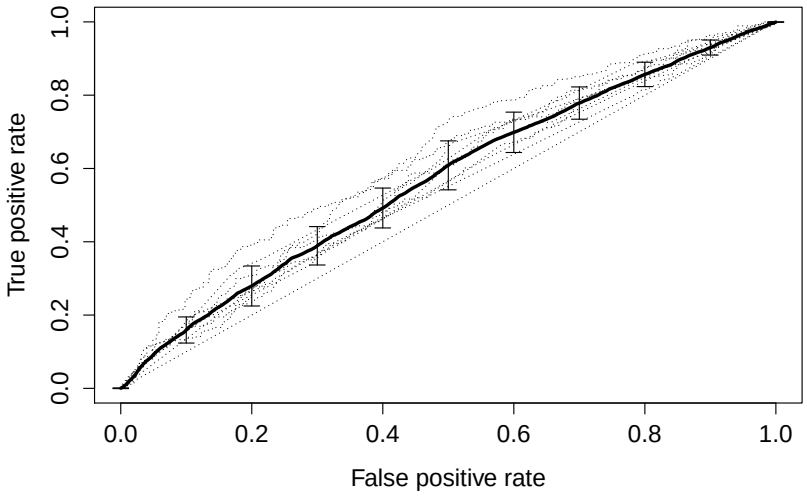


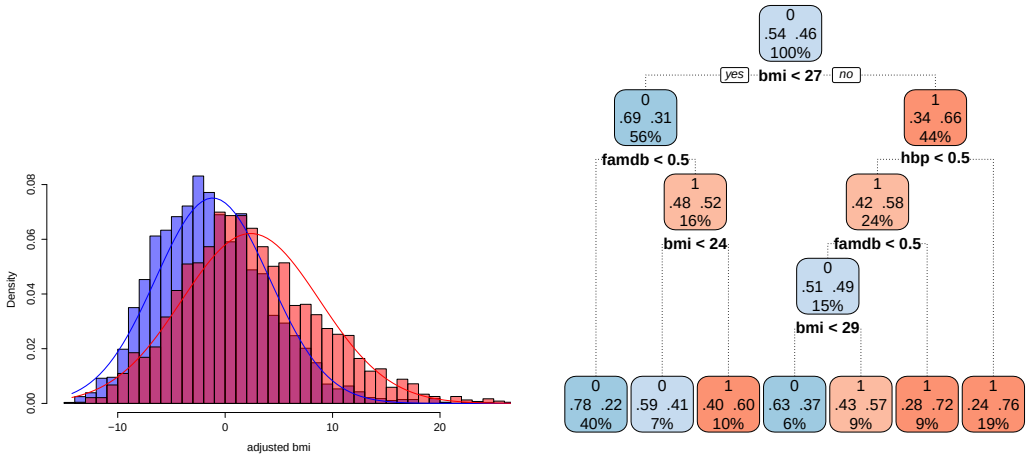
Figure 4: Performance for the simulation study. ROC curves for predicting diabetes status in the hold-out test data for each of the 10 simulations. The applied logistic regression models each included all genome-wide significant variables in the given simulation.

4. Discussion

The results from both the diabetes example and the simulation study demonstrate the challenge of moving from associations to predictions. Even though a strong association exists it is not necessarily possible to obtain a good performance of the prediction. In the diabetes example, genome-wide significant SNPs are identified, indicating that there are associations between these SNPs and the disease. However, including these SNPs, or other SNPs believed to be predictive of the disease, does not result in a good performance. When comparing the performance of the model including only age and sex to models also including the SNPs it is seen that the SNPs improve the prediction and the SNPs hold predictive information. When bmi is also included, no significant improvement in prediction performance is seen when adding the SNPs. This is also the case for the SNPs selected using random forest, although the selection method was hoped to utilize any non-linear signal. In order for any variable to have significant associations with disease status, the variables must describe a difference in distributions of the two groups, diabetics and non-diabetics. However, the two groups can overlap in the space spanned by the variables. For the sake of being able to predict, the overlap must be minimized, making the problem more challenging.

A visual way of considering this challenge is plotting the two populations in the diabetes example, diabetic and non-diabetics, against one variable. This is exemplified in Figure 5a where bmi adjusted for age and sex is plotted. The bmi is adjusted by using the residuals of the linear regression of bmi given age and sex. The two populations are clearly different (Welch Two Sample t-test, $p < 2.2 \cdot 10^{-16}$) but at the same time highly overlapping making it difficult to predict the correct population based only on the adjusted bmi.

In the simulated data study, the response is created from a subset of the explanatory variables with odds ratios at a level expected in a GWAS. This means they each have a low effect-size and that there are many noise variables, making



(a) Histogram (b) Classification Tree

Figure 5: (a) A histogram of diabetics (red) and non-diabetics (blue) plotted against bmi adjusted for age and sex. (b) Classification tree for predicting type-2 diabetes (1 is case, 0 control) using clinical variables. Here "famdb" is the family history of diabetes and "hbp" is reported high blood pressure.

it challenging to find the true associations and even more so to predict. This is evident as the number of associations found has a high variance and the response is not easily predicted giving that the performance is only slightly better than random.

A suggestion for future success in prediction is to first define subgroups based on clinical variables and then search for subgroups for which the genetic data can aid the prediction. This leads to more personalized medicine and give the genes the possibility to be useful for only a subgroup of the whole population. In Figure 5b, subgroups are defined for the diabetes example using a classification tree. It is seen that the performance of the prediction from clinical variables is good in some subgroups and that there is potential for aiding the prediction in other ill-performing nodes. If the genetics can aid the predictions in one of the subgroups then they are useful even if they cannot improve the performance for the whole population. The problem with this approach is the limited number of observations left in the subgroups. In the example, some nodes consist of as few as nine observations (Figure 5b). It thereby worsens an already common problem in GWAS of having a low sample size.

Acknowledgments

The dataset was provided by the GENEVA Genes and Environment Initiatives in Type 2 Diabetes project. Funding support for the GWAS of Gene and Environment Initiatives in Type 2 Diabetes was provided through the NIH Genes, Environment and Health Initiative [GEI] (U01HG004399). The human subjects participating in the GWAS derive from The Nurses' Health Study and Health Professionals' Follow-up Study and these studies are supported by National Institutes of Health grants CA87969, CA55075, and DK58845. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the Gene Environment Association Studies, GENEVA Coordinating Center (U01

HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH GEI (U01HG004424). The datasets used for the analyses described in this manuscript were obtained from dbGaP at [http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap] through dbGaP accession number [phs000091].

References

[1] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLoS Computational Biology*, vol. 8, 2012.

[2] A. P. Morris, B. F. Voight, T. M. Teslovich, T. Ferreira, A. V. Segre, V. Steinthorsdottir, R. J. Strawbridge, H. Khan, H. Grallert, A. Mahajan, *et al.*, "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes," *Nature genetics*, vol. 44, no. 9, p. 981, 2012.

[3] M. Mailman, M. Feolo, Y. Jin, and M. Kimura, "The ncbi dbgap database of genotypes and phenotypes," *Nature genetics*, vol. 39, no. 10, pp. 1181–1186, 2007.

[4] A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, and P. I. de Bakker, "Snap: a web-based tool for identification and annotation of proxy snps using hapmap," *Bioinformatics*, vol. 24, no. 24, pp. 2938–2939, 2008.

[5] S. Purcell, "Plink v1.07." `http://pngu.mgh.harvard.edu/purcell/plink/`, 2009.

[6] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly, and P. Sham, "Plink: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[7] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[8] A. M. Nielsen, "Application of Machine Learning on a Genome-Wide Associations Studies Data set," Master's thesis, Technical University of Denmark and Royal Institute of Technology, Denmark and Sweden, 2015.