Risk Prediction Using Genome-Wide Association Studies

Charles Kooperberg,* Michael LeBlanc, and Valerie Obenchain

Fred Hutchinson Cancer Research Center, Seattle, Washington

Over the last few years, many new genetic associations have been identified by genome-wide association studies (GWAS). There are potentially many uses of these identified variants: a better understanding of disease etiology, personalized medicine, new leads for studying underlying biology, and risk prediction. Recently, there has been some skepticism regarding the prospects of risk prediction using GWAS, primarily motivated by the fact that individual effect sizes of variants associated with the phenotype are mostly small. However, there have also been arguments that many disease-associated variants have not yet been identified; hence, prospects for risk prediction may improve if more variants are included. From a risk prediction perspective, it is reasonable to average a larger number of predictors, of which some may have (limited) predictive power, and some actually may be noise. The idea being that when added together, the combined small signals results in a signal that is stronger than the noise from the unrelated predictors. We examine various aspects of the construction of models for the estimation of disease probability. We compare different methods to construct such models, to examine how implementation of cross-validation may influence results, and to examine which single nucleotide polymorphisms (SNPs) are most useful for prediction. We carry out our investigation on GWAS of the Welcome Trust Case Control Consortium. For Crohn's disease, we confirm our results on another GWAS. Our results suggest that utilizing a larger number of SNPs than those which reach genome-wide significance, for example using the lasso, improves the construction of risk prediction models. *Genet. Epidemiol*. 34:643–652, 2010. © 2010 Wiley-Liss, Inc.

Key words: Crohn's disease; elastic net; GWAS; lasso; model selection

Contract grant sponsors: Wellcome Trust; National Institute of Diabetes and Digestive and Kidney Diseases; National Institutes of Health; Contract grant numbers: CA74841; CA53996; CA125489; CA90998.

*Correspondence to: Charles Kooperberg, Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, 1100 Fairview Avenue N, M3-A410, Seattle, WA 98109-1024. E-mail: clk@fhcrc.org

Received 9 December 2009; Revised 25 March 2010; Accepted 29 March 2010
Published online 14 September 2010 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/gepi.20509

INTRODUCTION

It is the hope that genome-wide association studies (GWAS) not only identify regions in the genome (e.g. single nucleotide polymorphisms (SNPs)) that are associated with phenotypes but also that the genetic variants that are identified can contribute to (risk-)prediction models for those phenotypes. In some GWAS publications, there have been initial attempts to look at the predictive power of the few identified ''top-SNPs'' [e.g. Lin et al., 2009; Miyake et al., 2009; Myocardial Infarction Genetics Consortium, 2009; Zheng et al., 2008]. The International Schizophrenia Consortium [2009] combined large numbers of variants. A recent commentary suggested that more and larger studies would help yield more effective prediction models [Kraft and Hunter, 2009]. The results of Gail [2009] are in line with the commentary. In Evans et al. [2009], it was investigated how models, combining marginal results for SNPs in a sensible way, can be used to compute a prediction score. Wei et al. [2009] recently carried out a similar experiment using Support Vector Machines.

In our article, we take this several steps further: (i) we use sparse regression models, which deal with correlation between SNPs, and yield estimates of the probability of disease; (ii) by varying some aspects of the cross-validation, we critically examine the effect of selecting significant SNPs in the same study as on which the prediction models are constructed; and (iii) we evaluate one of the constructed prediction models on an entirely different GWAS.

Most published GWAS have identified one to a few SNPs associated with a phenotype. Further meta-analyses of GWAS studies with the same phenotype often identify additional SNPs, but for most phenotypes the number of consistently confirmed SNPs is less than a dozen. Exceptions are some continuous phenotypes, such as lipids, for which GWAS have more power. The lack of power of GWAS suggests that there *may* very well be many more SNPs associated with some phenotypes that have smaller effect sizes. It would be expected that such SNPs are among the SNPs that have larger, but statistically insignificant, test-statistics. Another argument for the existence of more SNPs that are associated with a disease is that often the Q–Q plots for the P-values start showing deviations from the identity (diagonal) well before the effects are significant.Sparse regression methods, such as the lasso [Tibshirani, 1996] and the elastic net [Zou and Hastie, 2005], are increasingly used in high-dimensional settings [Hastie et al., 2001]. The advantage of those approaches is that in regression models they simultaneously carry out variable selection, and provide estimates of the coefficients of the selected variables. In this article, we explore the use of such methods for the construction of risk prediction models using GWAS data. In Wu et al.

[2009], the lasso is used for finding significant SNPs in GWAS data. In Park and Hastie [2008], sparse regression methods are used to identify gene × gene interactions in smaller genetic association studies. However, to the best of our knowledge, sparse regression methods have not yet been used to construct prediction models in GWAS to estimate the probability of disease and validate these probabilities on an independent data set.

When constructing such models, it is important to keep the ''training data'' and ''test data'' strictly separate. As such, we cannot start from the consensus list of disease-associated SNPs, as typically all data sets were used to obtain such a list. Instead we need to use the training data to select the set of SNPs that will be used in constructing the prediction model *and* the actual construction of that model. Additionally, the high bar that typically exists for a SNP to be declared genome-wide significant (e.g. $P<10^{-7}$) is not necessary for a SNP to be useful in risk prediction. In particular, a group of SNPs that have promising False Discovery Rates, but are not genome-wide significant will likely make a positive contribution to a risk prediction model, as some of these SNPs will predict, while the other ones just add some noise.

Our experiments consist of two parts. In the initial phase, we used the Welcome Trust Case Control Consortium (WTCCC) GWAS data with about 3,000 controls and 2,000 cases for several diseases [Welcome Trust Case Control Consortium, 2007], and split those in a training and a test data set. Model selection (based on cross-validation) was performed on the training data, and the selected model was evaluated on the test data. This article contains the results for the Crohn's disease GWAS; the results for type 1 diabetes and type 2 diabetes data are mostly similar; we refer to the results for these phenotypes in a few places. It should be noted that Wei et al. [2009] used a Support Vector Machine approach for the WTCCC type 1 diabetes data. Test data log-likelihood, receiver operator characteristic (ROC) curves, and the area under the curve (AUC) are used to evaluate the models. In addition, we check whether estimated probabilities of ''a subject being a case'' correspond to the fraction of subjects that are a case. Other methods for assessing risk models, such as positive- and negative-predicted value and mis-classification rate exist [e.g. Pepe, 2003], though it has been argued that well-calibrated probabilities are of critical importance in individual risk prediction [Cook, 2007].

As a confirmation experiment, we used the complete WTCCC Crohn's disease data, consisting of UK subjects, as our training data, and the National Institute of Diabetes and Digestive and Kidney diseases (NIDDK) GWAS data on Crohn's disease, consisting of US subjects, as our test data. We should note here that the NIDDK data were genotyped on the Illumina platform, and the WTCCC data on the Affymetrix platform; thus to apply the WTCCC-derived model to the NIDDK data, we needed to impute the Affymetrix SNPs on the NIDDK data.

Overall our results suggest that prediction models (when applied to GWAS cohort data) like the one we develop may provide well-calibrated risk estimates. But the predictive value overall is somewhat limited, as demonstrated by the overall AUC of these models; they may be most useful for designing trials, and weighing risk-benefits for preventive treatments. For individual risk prediction, the genetic factors would presumably be more useful if combined with other established risk factors than if the genetic factors were used by themselves.

METHODS

DATA PROCESSING

We obtained GWAS data from the Welcome Trust Case Control Consortium [Welcome Trust Case Control Consortium, 2007] and data from the NIDDK GWAS on Crohn's disease [Duerr et al., 2006; Rioux et al., 2007] from dbGaP. We refer to these data sets as the WTCCC and the NIDDK data, respectively.

WTCCC DATA

The WTCCC data consist of 2,000 cases for each of seven diseases and 3,000 shared controls. The subjects in this study, from the UK, were genotyped on the Affymetrix 5.0 platform. Separately for the Crohn's disease data, the type 1 diabetes data, and the type 2 diabetes data, we removed all samples with call rate smaller than 0.95, removed SNPs that had a Hardy Weinberg *P*-value smaller than 10^{-5}, SNPs that had more than 10% missing data, SNPs that were in the WTCCC list of ''bad SNPs'' that was provided with the data, and SNPs with minor allele frequency smaller than 0.05.

The small amount of remaining missing data was imputed using a probabilistic imputation based on a three SNP sliding window. That is, if SNP i was missing for participant j, we calculated the probability distribution of SNP i given the values of SNP $i-1$ and $i+1$ (for the appropriate case or control portion of the data) and imputed a random realization of this probability distribution. This imputation was quick and easy, and allowed us to carry out initial experiments using multiple imputation (not reported here). Other more sophisticated imputation methods, such as MACH [Li et al., 2006] could have been used as well. Given the small amount of missing data, this would not have had any qualitative effect on our results.

NIDDK DATA

The NIDDK GWAS consists of 792 cases of Crohn's disease and 932 controls. These subjects, Caucasians in the US, were genotyped on the Illumina HapMap300. As for the WTCCC data, we removed subjects with large percentages of missing data, and SNPs that had high missingness rates, or very small Hardy–Weinberg *P*-values. SNPs with small minor allele frequency were retained.

All prediction models that were applied on the NIDDK data used (a subset of) the 3,000 marginally most significant SNPs in the WTCCC Crohn's disease data. As most of the SNPs that were part of the Affymetrix 5.0 panel of SNPs are not part of the HapMap300 panel of SNPs, we used MACH [Li et al., 2006] to impute the data. In particular, we used ten sets of probabilistic inputs based on the CEPH HapMap data, using MACH options `--greedy --phase`. Since some of the SNPs in the prediction models may be in the same haplotype blocks, we did not want to impute either the ''best'' imputation, or use marginal posterior probabilities, as those approaches do not acknowledge joint imputation probabilities between two SNPs. The results reported in this article are based on the average posterior probability of a subject being a case over the ten imputed data sets.

PREDICTION METHODS

Let Y_i be a binary indicator for the phenotype of subject $i=1,\ldots,n$, and let X_{ij} be the value of SNP $j=1,\ldots,p$ for subject i, coded as 0,1,2 for the number of minor alleles. We write $Y=(Y_1,\ldots,Y_n)^t$, $X_j=(X_{1j},\ldots,X_{nj})^t$, and $x_i=(X_{i1},\ldots,X_{ip})^t$. All approaches that we consider are carried out on data sets of the p most significant SNPs ($p\leq 3{,}000$). These most significant predictors are pre-selected on just the training data, and this selection is repeated each of the cross-validation steps. (As the results, we report here typically do not change beyond $p\sim 2{,}000$ and computations are getting increasingly slow, we did not systematically investigate $p>3{,}000$.) For the remaining, assume that we have ordered the SNPs, and that X_1 is the marginally most significant SNP, X_2 the next most significant one, and so on.

Since for most phenotypes that are studied in GWAS the signal is small, and often other risk factors are known, we focus on modeling the probability that a subject is a case using logistic regression, rather than looking at classification. We also believe that a probabilistic risk estimate can convey more subtilty than a simple classification. If a classification rule is needed probabilistic estimates can be thresholded taking mis-classification costs in consideration. The simplest approach to model the probabilities is to fit a linear logistic regression model on the p pre-selected SNPs:

$$\text{logit}(P(Y_i=1|x_i))=\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}.$$

Traditionally, parameters in this model are estimated using maximum likelihood. When large numbers of predictors are used, the logistic regression model is known to overfit the data. Instead, we consider the lasso and the elastic net, two examples of penalized regression methods. Let $l(\beta;\ Y_i,\ x_i,\ i=1,\ldots,n)$ be the logistic log-likelihood. The lasso and elastic net estimates of β are the maximizers of

$$l(\beta;\ Y_i,\ x_i,\ i=1,\ldots,n)-\lambda_1\sum_{j=1}^{p}|\beta_j|,$$

and

$$l(\beta;\ Y_i,\ x_i,\ i=1,\ldots,n)-\lambda_1\sum_{j=1}^{p}|\beta_j|-\lambda_2\sum_{j=1}^{p}\beta_j^2,$$

respectively, where λ_1 and λ_2 are selected using cross-validation. Both of these approaches effectively carry out model selection, as the l_1 penalty $\lambda_1\sum_{j=1}^{p}|\beta_j|$ will set many of the coefficients β_j to 0. The potential advantage of the elastic net is that when many of the predictors are highly correlated, the l_2 penalty $\lambda_2\sum_{j=1}^{p}\beta_j^2$ encourages averaging of multiple-correlated predictors, while the lasso would select just a single predictor. The elastic net penalty can be viewed as a combination of the lasso penalty and the l_2 penalty form ridge regression, an early penalized regression method [Hoerl and Kennard, 1970]. We applied both approaches, using the R [R Development Core Team, 2009] package `glmnet`. The implementation of this package has the advantage that computation time does not significantly increases even when 1,000s of values of λ_1's are used in the lasso, making the optimization over this parameter using cross-validation straightforward.

In addition to these penalized regression methods, we considered traditional (stepwise) fitting of logistic regression models. In particular, we considered fitting models of the p marginally most significant predictors (referred to as "generalized linear models" (GLM)), fitting models of the p marginally most significant predictors, omitting any predictor that is correlated with correlation larger than $R=0.9$ to a more significant predictor (referred to as "filtered GLM"), stepwise addition of predictors (considering the top p marginally most significant SNPs) using AIC and BIC to select the number of SNPs (referred to as "stepwise GLM-AIC" and "stepwise GLM-BIC"). The parameter p for GLM and filtered GLM was selected using cross-validation of the log-likelihood. We investigated alternative cut-offs for the correlation between predictors for the filtering and found that the actual value has little influence over a range of about (0.7, 0.95).

CROSS-VALIDATION

There are two model selection steps in our procedure: selecting the parameters λ_1 (lasso and elastic net), λ_2 (elastic net), and p (stepwise GLM), and pre-selecting the marginally most significant SNPs that will be considered by the modeling approaches. The parameters were all selected using 10-fold cross-validation on the training data. We considered three approaches to pre-selecting the marginally most significant SNPs:

1. Pre-select those SNPs based on the training and test data combined.
2. Pre-select those SNPs based on the complete training data.
3. Repeatedly pre-select those SNPs for each cross-validation step, using only nine-tenths of the training data that are used to fit the model. The parameters (λ_1 and λ_2 for the lasso and the elastic net, p for some of the GLM procedures) are selected as those that maximize the log-likelihood on the one-tenth validation data averaged over the 10-folds. After the parameters are selected, the SNPs are once more selected, now using the complete training data, and the final model is estimated using the complete training data and the parameters selected in the cross-validation procedure. This model is the model that is evaluated on the test data.

For most of our experiments, the WTCCC data is split in a training and a test data set, as described above. Using Approach 3, we also analyze the data using the WTCCC data as the training set and the NIDDK data as the test data set.

In the model selection literature, the consensus is that such a pre-selection of SNPs should only use the data that is trained for in a particular validation run (as in Approach 3), and that it should not include the 10% of the training data used for validation (as in Approach 2), and definitely not the test data (as in Approach 1) [Hastie et al., 2001]. Since each of these approaches to pre-selection have occurred frequently in the GWAS literature, we wanted to quantify the magnitude of the problem that using Approaches 1 or 2 causes. For example, in Evans et al. [2009], the AUC values in their Table I appear to follow the strict selection rules of Approach 3, but the AUC values in their Table II, which include variants that were identified in the same study as the one on which they are evaluated, would be considered closer to Approach 1. The International Schizophrenia Consortium [2009] used a separate training and test set when constructing their model, more in line with Approach 2. The model developed in Myocardial Infarction Genetics Consortium [2009] is a combination of Approaches 1 and 3, as some of the included variants were identified in the study on which the risk model was evaluated, while others were identified in other studies.

EVALUATION

For each of the models, we evaluated the predicted probability of disease using the logistic regression model for the selected model on test data. We summarize these fitted probabilities using test data log-likelihoods, receiver operating characteristic (ROC) curves, and the AUC. For Figures 4 and 6, we fitted a generalized additive model with smoothing splines for logistic regression, using default settings in the R package gam, of fitted probability on case-control status.

RESULTS

After the initial data processing described in the METHODS section, there were 4,686 subjects (1,748 cases and 2,938 controls) in the WTCCC data, which we divided in a training set of 2,808 subjects (1,045 cases and 1,763 controls) and a test set of 1,878 subjects (703 cases and 1,175 controls).

CROSS-VALIDATION AND THE SELECTION OF SIGNIFICANT PREDICTORS

In Figure 1, we show the log-likelihood for the lasso with the three approaches to pre-select the most significant predictors in conjunction with cross-validation for the WTCCC Crohn's disease data. For Approach 1, all training and test data is used to pre-select the most significant predictors. The effect of this is that the test data is more similar to the training data than would be the case for new independent data. As a result, when the number of SNPs considered increases substantially, even the test results show overfitting. This overfitting, which is very apparent when the number of SNPs considered is larger than 100, is already evident from comparing to Approach 3 when the number of SNPs reaches 10. This result raises concern regarding predictive models for GWAS, where the same data is used to discover the SNPs and to construct the prediction models, even if only the "top established hits" are used.

Approach 2 uses the complete training data but not the test data to pre-select the most significant SNPs. As a result, the 90% learning part of the training data used during cross-validation is more similar to the 10% validation part of the training data than to the test data. As s result of this similarity between the validation data and the training data, the selected model is larger than it is in Approach 3 (i.e. the λ_1 in the lasso is smaller than what it is in Approach 3), as a result the model using the training data overfits more. In particular, the model uses too many SNPs and the fitted model when applied to the test data, which is more different from the training data, performs worse. Since the cross-validation results are used to select the model size, this would results in a too complex prediction model being selected.

Approach 3 pre-selects the top SNPs for each cross-validation fold. The effect of this is that the validation 10% of the data is as (dis)similar to the training data as the test data is to the training data. As a result, we are not given the illusion that model fits keep on improving with

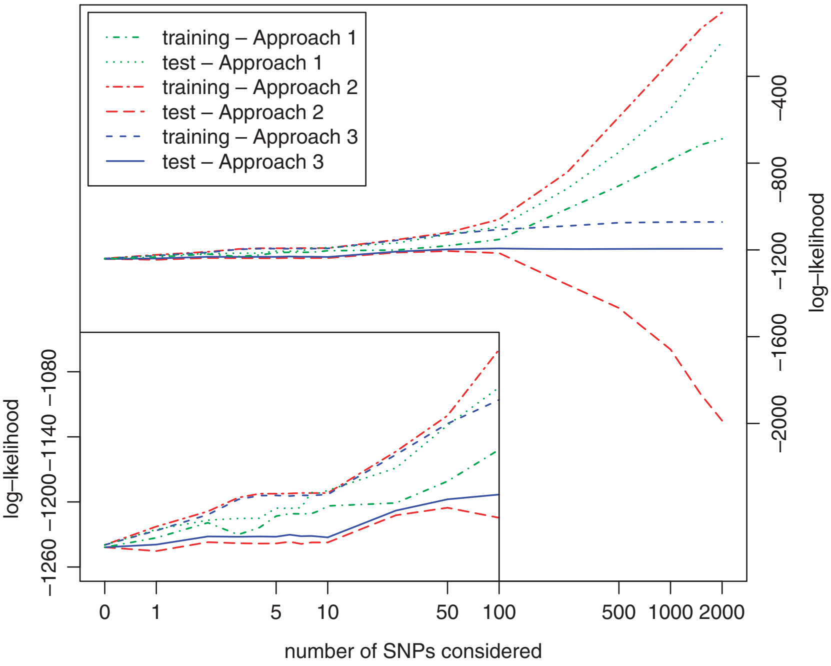

Fig. 1. Log-likelihood for the WTCCC Crohn's disease data using three different ways to carry out the pre-selection of significant SNPs in relation to the cross-validation. The training data log-likelihood was rescaled by a factor of 1,878/2,808 to be on the same scale as the test data log-likelihood. WTCCC, Welcome Trust Case Control Consortium; SNPs, single nucleotide polymorphisms.

increased model size (such as for Approach 1), and we do not overfit the data (such as for Approach 2).

We also compare the test data log-likelihood results that are displayed in Figure 1 with what we would obtain, if we used the test data to select the best parameters after the most significant SNPs are pre-selected on the complete training data (data not shown). This approach would generally be considered ''cheating.'' The difference between the ''cheating'' approach and the ''correct'' Approach 3 is much smaller than the differences between Approaches 1 and 3 or the differences between Approaches 2 and 3, suggesting that incorporating pre-selection of the significant SNPs in cross-validation is in fact very important.

CHOICE OF MODELING METHOD

When larger numbers of predictors (SNPs) are used to model a regression outcome, sparse regression methods, such as the lasso [Tibshirani, 1996] and the elastic net [Zou and Hastie, 2005], can be used both to carry out model selection, and for the estimation of the parameters in a (generalized) linear model. These sparse regression methods are alternatives to standard GLM, where model selection can be implemented by either restricting the generalized linear model to the most significant SNPs, or by using stepwise regression, where at each step of the algorithm the most significant SNP is selected for inclusion in the model.

We obtained prediction models using the lasso, the elastic net, GLM, filtered GLM, and stepwise GLM on a training data set of 60% of the WTCCC data. When we applied these models to the remaining 40% of the WTCCC data as test data, leading to test-sample-validated predictions, we noted that the lasso and the elastic net gave very similar results, both with respect to the test data log-likelihood and AUC (results for Crohn's disease in Table I and Fig. 2; the results for the diabetes data are similar).

The model that was selected using stepwise GLM with AIC has a much worse likelihood than the models obtained using other methods because on a few of the test samples some of the correlated SNPs that perform well in the training data yield very bad results. (A few estimated probabilities of a cases close to 0 ''sink'' the log-likelihood, which is all that AIC uses to select the model). This does not effect the AUC, which is comparable to results obtained by the lasso. The filtered GLM, which could be considered as a very simple regularization method for the regression, performs the best of the GLM methods, considering both AUC and log-likelihood.

For most of the elastic net models displayed the λ_2, penalty parameter selected by cross-validation was 0; for models using these number of SNPs, the elastic net results are thus the same as the lasso results. For the few models that were different between the lasso and the elastic net, the differences in log-likelihood were small, and the differences in AUC were negligible. Therefore, we do not display the results obtained using the elastic net. Use of the lasso and cross-validation limits the amount of overfitting, but while the test-data results flatten off when more than 100 SNPs are considered, the training data results still improve, suggesting some overfitting (Fig. 2). For type 2 diabetes, there is much less signal in the WTCCC data. The results for type 1 diabetes are somewhat better than for Crohn's disease because of the strong dependence of type 1 diabetes on the HLA genes. In fact, on the test data for type 1 diabetes, we achieved an AUC of 0.88, very similar to the AUCs of between 0.87 and 0.89 that Wei et al. [2009] achieved on the validation part of the training data using Support Vector Machines.

TABLE I. Number of SNPs used in the prediction models with non-zero coefficients, log-likelihood, and AUC for the test data ($n = 1{,}878$: 1,175 controls and 703 cases) for the WTCCC Crohn's disease data using the lasso

Method	SNPs used	Log-likelihood	AUC
No SNPs used	0	−1,241.77	0.500
GLM	18	−1,223.12	0.606
Filtered GLM[a]	26	−1,224.94	0.626
Stepwise GLM AIC	38	−1,287.68	0.631
Stepwise GLM BIC	14	−1,236.35	0.614
Lasso top 1 SNPs considered	1	−1,239.33	0.528
Lasso top 2 SNPs considered	2	−1,231.84	0.551
Lasso top 5 SNPs considered	4	−1,232.09	0.569
Lasso top 10 SNPs considered	6	−1,232.71	0.568
Lasso top 25 SNPs considered	14	−1,207.98	0.612
Lasso top 50 SNPs considered	25	−1,197.59	0.630
Lasso top 100 SNPs considered	33	−1,193.20	0.637
Lasso top 250 SNPs considered	91	−1,196.24	0.634
Lasso top 500 SNPs considered	155	−1,195.61	0.635
Lasso top 1,000 SNPs considered	176	−1,195.04	0.636
Lasso top 2,000 SNPs considered	177	−1,194.78	0.637

WTCCC, Welcome Trust Case Control Consortium; AUC, area under the curve; GLM, generalized linear models; SNPs, single nucleotide polymorphisms.

[a] For filtered GLM SNPs with correlation larger than 0.9 with a more significant SNP were omitted. The highest rank SNP among the selected SNPs for filtered GLM was 58.

The number of SNPs that have nonzero coefficients in the lasso models levels off at about 175 for Crohn's disease (Table I). When 100 of the top SNPs were considered as predictors, 33 ended up having nonzero coefficients in the lasso models. A paired t-test on the likelihoods of the test set, suggests that the model where the top 100 SNPs were considered was significantly better than the model where the top 25 SNPs were considered ($p \sim 0.025$). While in meta-analysis with other Crohn's disease data sets such a number of disease-associated SNPs in GWAS have been identified, this is a much larger number than what was identified using just the WTCCC data. Similar results were obtained for the WTCCC data for type 1 and 2 diabetes. An advantage of the lasso over other machine-learning techniques, such as support vector machines, is the effective selection of SNPs when some parameters corresponding to SNPs get set to zero. In Figure 3, we display for several of the lasso models, which of the SNPs that were considered were used with nonzero coefficients and which SNPs were not used. We note that most SNPs used in models where fewer SNPs are considered are also used in the models where more SNPs are considered, as is evident from the vertical stripes in Figure 3. Some of the highly significant SNPs are not used in a prediction model: for example, the column for SNP 5 is completely light-colored for lasso models, indicating that this SNP is not

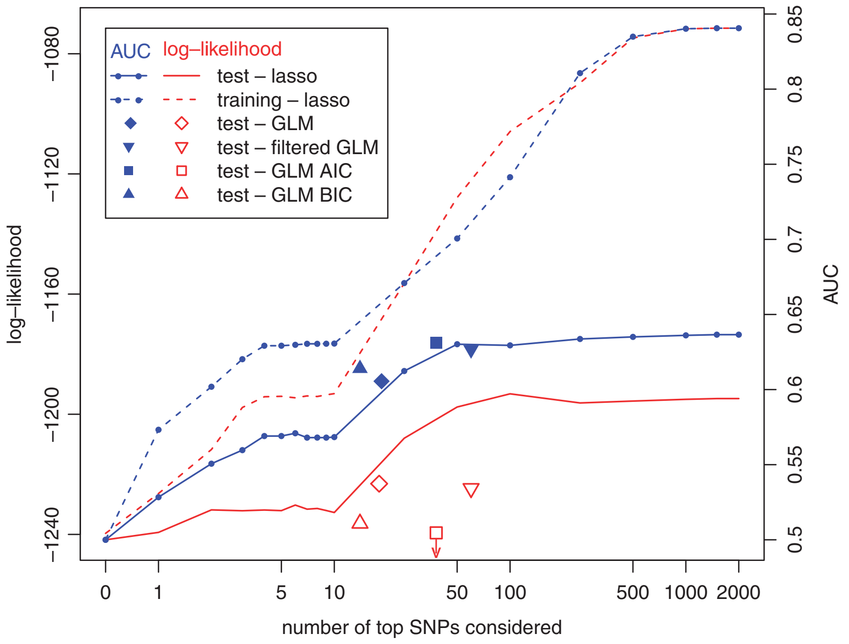

Fig. 2. Log-likelihood and AUC for the WTCCC Crohn's disease data for prediction models for test and training data. The training data log-likelihood was rescaled by a factor of 1,878/2,808 to be on the same scale as the test data log-likelihood. Note that not all SNPs considered have nonzero coefficients, see Table I. The log-likelihood for stepwise GLM using AIC (GLM-AIC) is –1,287.7. The insert figure at the left bottom vertically expands the curves for the models with 100 SNPs or less. WTCCC, Welcome Trust Case Control Consortium; AUC, area under the curve; GLM, generalized linear models; SNPs, single nucleotide polymorphisms.

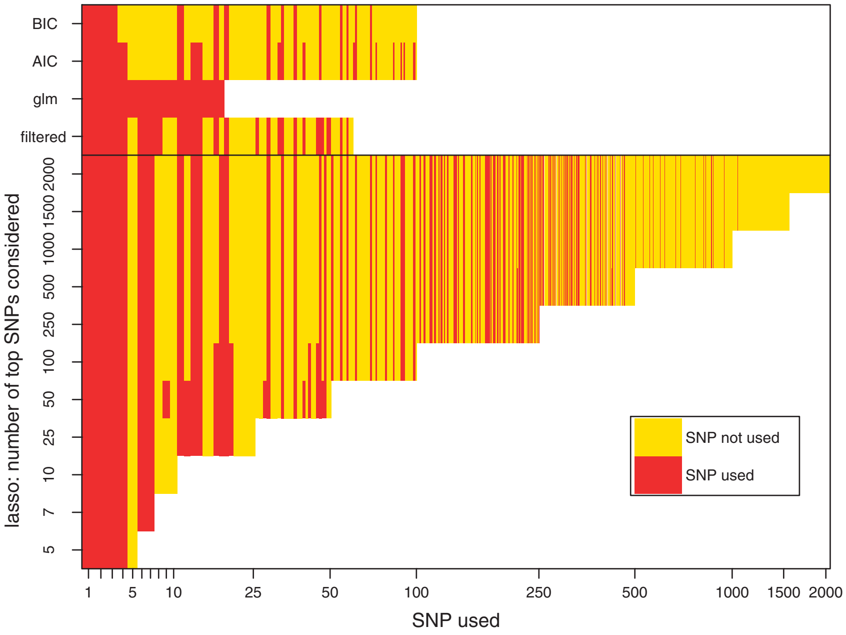

Fig. 3. Which SNPs are and are not used with nonzero coefficients for the lasso model and other prediction models for the WTCCC Crohn's disease data. The SNPs are ordered on the horizontal axis by significance. The vertical stripes suggest that frequently the same SNPs are selected. WTCCC, Welcome Trust Case Control Consortium; SNPs, single nucleotide polymorphisms.

used in any of those models. Typically, these SNPs are highly correlated with more significant SNPs. (For example, SNP 5 has correlation larger than 0.99 with SNP 4.) On the other hand, a few SNPs that are less significant are used: for example, several SNPs with ranks larger than 500 have nonzero coefficients. These SNPs maybe less significant in marginal models, but are more significant in models where the other SNPs are already included.

CALIBRATION OF PREDICTION PROBABILITIES

In Figure 4, we show the results of a smoothed regression estimate of the fraction of the data that is a case as a function of the estimated probability. We would like the curves for the test data to follow the dashed diagonal; if there was no signal in the data, the curves would be horizontal. This is the case for the test data for the two models, while the curves for the training data become much worse as the number of SNPs considered increases. The WTCCC data is case-control data, but in this article, we treat it as cohort data, which is reasonable as both cases and controls are samples of the UK population of cases and controls (be it obtained in different ways). As such, the "incidence" in this data is 40% by design. The population incidence in the US is between 1 in 500 and 1 in 1,000; with different population probabilities, the axes in Figure 4 would both be rescaled, but the angle of the curve would remain unchanged. To practically use a prediction model, we would want to combine this prediction model with established risk factors, such as age, ethnicity, smoking, and family history. A prediction model using genetics could be seen as a refinement of the family history risk factor.

APPLYING A MODEL OBTAINED FROM ONE GWAS TO ANOTHER POPULATION

We applied the results from the lasso on the WTCCC data to the data from a GWAS on Crohn's disease carried out by the NIDDK. While the populations for both GWAS studies are Caucasian, the population for the WTCCC GWAS is from the UK, and the one for the NIDDK GWAS from the USA. There are also many technical differences between these two GWAS; for example, since the NIDDK data were generated on a different platform, genotypes were imputed. See the Methods section for details. Figure 5 compares the AUC for models considering different numbers of SNPs for the test data part of the WTCCC data and the NIDDK data. We note that, surprisingly, the AUC for the NIDDK data is larger than for the WTCCC data. Figure 6 shows the results of a smoothed regression estimate of the fraction of the data that is a case as a function of the estimated probability for models considering 100 SNPs, arguably the best model size for the WTCCC data. We note that the probabilities appear well calibrated. (In fact, the models using both 50 and 250 SNPs appeared even better calibrated.)

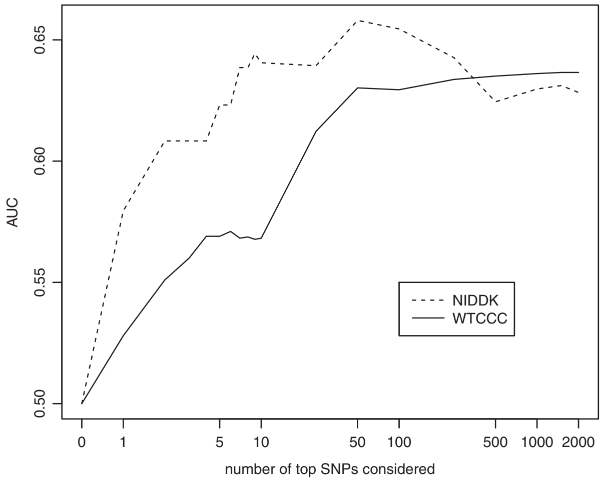

Fig. 5. Comparison of test data AUC for the NIDDK and WTCCC data. The model for the NIDDK data is trained on the complete WTCCC data, the model for the test part of the WTCCC data is trained on the training part of the WTCCC data. WTCCC, Welcome Trust Case Control Consortium; AUC, area under the curve; NIDDK, National Institute of Diabetes and Digestive and Kidney diseases.

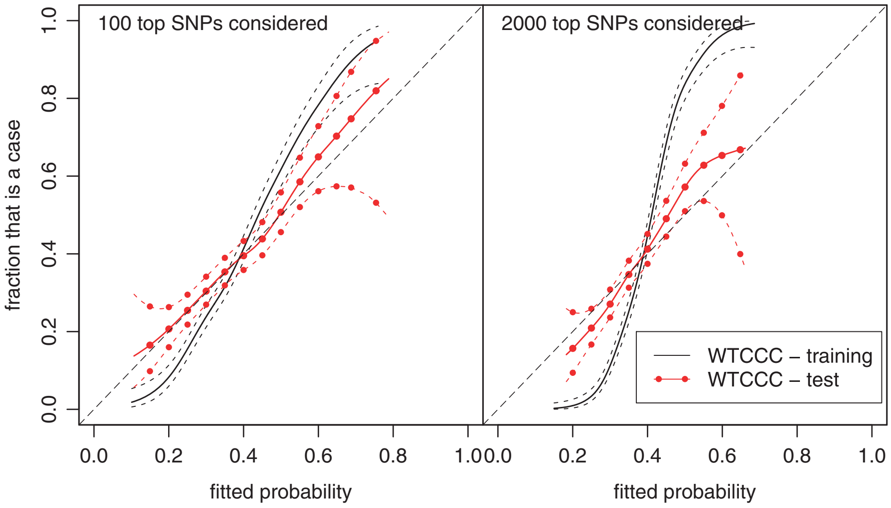

Fig. 4. Smoothed estimates of the probability of being a case as a function of the predicted probability of being a case with 95% confidence intervals for the WTCCC Crohn's disease data. The steeper curves for the training data suggest some overfitting, while the test data appears better calibrated. WTCCC, Welcome Trust Case Control Consortium.

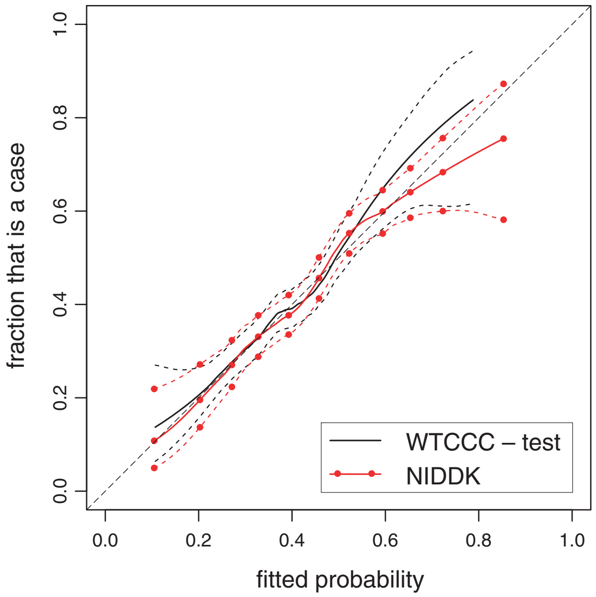

Fig. 6. Smoothed estimates of the probability of being a case as a function of the predicted probability of being a case with 95% confidence intervals for the NIDDK and WTCCC data for the lasso model that considers 100 SNPs. The model for the NIDDK data is trained on the WTCCC data, the model for the WTCCC data is trained on another part of the WTCCC data. WTCCC, Welcome Trust Case Control Consortium; NIDDK, National Institute of Diabetes and Digestive and Kidney diseases; SNP, single nucleotide polymorphism.

DISCUSSION

We believe the results from our experiments on three of the diseases studied in the WTCCC GWAS (using a 40% test sample for validation) and full test data from the NIDDK Crohn's disease GWAS convincingly support the feasibility for constructing multi-SNP risk models from GWAS (albeit with modest predictive strength). These models include SNPs that would have failed a genome-wide significance test. We obtained similar results for the WTCCC GWAS of type 1 diabetes and type 2 diabetes, suggesting that our results are generalizable to other diseases.

Our model building strategy incorporated several important but generally accepted statistical components.

1. The genotyped SNPs were first filtered with respect to missingness, Hardy–Weinberg equilibrium, and minor allele frequency.
2. Smaller numbers of SNPs <3,000 were pre-selected for more model building; this step controls the variability of subsequent regression analysis.
3. The penalized regression modeling (e.g. lasso) incorporates variable selection and estimation properties that further controls the variability of the risk estimates.
4. To avoid models, which are overly optimistic (with too many SNPs), similar to the "winners curse", we used cross-validation of the entire model building process, including the pre-selection of the larger number of SNPs that are considered by the penalized regression modeling, to obtain relatively unbiased estimates of prediction error.

In other words, our GWAS risk model statistical recipe includes a small number of ingredients: quality control, variance control, well structured SNP combinations and prediction optimism adjustment via cross-validation.

We believe that applying the model fit on the WTCCC GWAS to the NIDDK GWAS provides a strong confirmation of our approach because of the differences between the two studies: (i) the WTCCC GWAS and NIDDK GWAS were carried out on different genotyping platforms, so that the NIDDK prediction is more than 90% based on imputed SNPs; (ii) both studies were carried out on different populations, the WTCCC GWAS was carried out in the UK, the NIDDK GWAS within the US; (iii) there is no information whether other important characteristics, such as disease adjudication for the cases and risk profile of the controls, was comparable. Notwithstanding these differences, the estimated probabilities from the WTCCC GWAS-derived model, calibrated well on the NIDDK cohort. It has been argued that whether a probability estimate is well calibrated is more important for individual risk prediction than a focus on classification methods or AUC [Cook, 2007].

The predictive ability of the models that we derive, as measured by the area under the curve (AUC), is modest and the greatest utility of these models is probably for risk calculation in research studies. However, we should put that in perspective in that these genetic factors can be used in addition to already established other risk factors. Also, many frequently used risk prediction models have modest AUCs. For example, the often used Gail model for prediction of breast cancer risk [Gail et al., 1989] has an AUC of only about 0.58–0.6 [Rockhill et al., 2001]; nevertheless, it is used frequently for identifying risk groups for research studies and even sometimes for individual risk prediction, see for example `www.cancer.gov/bcrisktool`, although the use for individual prediction is sometimes questioned [e.g. Spiegelman et al., 1994].

Constructing risk prediction models with many predictors requires some form of regularization, as well as careful model selection using, for example, cross-validation. The lasso and the elastic net achieve this in an automated way, but the relatively good performance of the filtered GLM suggests that even some simpler regularization can do a decent job. The lack of regularization in (unfiltered) GLM and the less careful model selection using stepwise GLM can sometimes lead to less impressive results.

We think that the success of the lasso algorithm for risk prediction in GWAS is not surprising. The performance of the lasso and several variants has been studied in considerable theoretical detail in the statistics literature. For instance, in early work, Donoho and Johnstone [1994] showed near-minimax risk of the predictions for case of orthogonal predictors. And while orthogonality is not true for GWAS SNPs (since $p > n$) there is typically small correlation between SNPs that are not close together in the genome.

Theorem 2 in Zou and Hastie [2005] implies that the elastic net is a stabilized version of the lasso. It also suggests that if the predictors that are considered are uncorrelated the elastic net solution should be similar to the lasso. Thus, the observation that most of the top SNPs in a GWAS are fairly uncorrelated explains the fact that the elastic net and the lasso perform equivalently in our experiments.

More recently, mathematical studies have shown success of the adaptive lasso algorithm [Zou, 2006], which adaptively weights the penalty function based on initial estimators of the regression coefficients. Such a strategy directly relates to our pre-selection of a smaller number of SNPs to include in the regression based on univariate regression *p*-values. As the sample size gets large, good performance of the procedure (for both correctly selecting SNPs associated with outcome and prediction error) is obtained if the SNPs that are associated to the phenotype have only low correlation to those SNPs that are not associated with the phenotype. We believe that this is a reasonable assumption for most SNPs in GWAS.

We also establish that to get unbiased prediction results, it is critical to have a strict separation of training and test data: cases in the test data should not even be used to identify the most significant SNPs.

In estimating odds ratios related to the most significant SNPs, this effect is generally known as the ''winners curse'' [Zhong and Prentice, 2008; Zõllner and Pritchard, 2007]. Unfortunately, in risk prediction models, we have noted that many GWAS publications publish small ''prediction models'' on the hand-full of SNPs implicated in the same publication.

ACKNOWLEDGMENTS

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. The NIDDK IBDGC Crohn's Disease Genome-Wide Association Study was conducted by the NIDDK IBDGC Crohn's Disease Genome-Wide Association Study Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This manuscript was not prepared in collaboration with Investigators of the NIDDK IBDGC Crohn's Disease Genome-Wide Association Study study and does not necessarily reflect the opinions or views of the NIDDK IBDGC Crohn's Disease Genome-Wide Association Study study or the NIDDK. Data sets used in this study were dbGaP accession phs000130.v1.p1.

REFERENCES

Cook NR. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115:928–935.

Donoho DL, Johnstone IM. 1994. Ideal spatial adaptation via wavelet shrinkage. Biometrika 81:425–455.

Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Data LW, Kistner EO, Schumm P, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314:1461–1463.

Evans DM, Visscher PM, Wray NM. 2009. Harnessing the information contained with genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet 18:3525–3531.

Gail MH. 2009. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. J Natl Cancer Inst 101:959–963.

Gail MH, Brintom LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81:1879–1886.

Hastie T, Tibshirani R, Friedman J. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67.

Kraft P, Hunter DJ. 2009. Genetic risk prediction—are we there yet? N Engl J Med 360:1701–1703.

Li Y, Ding J, Abecasis GR. 2006. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet 79:S2290.

Lin X, Song K, Lim N, Yuan X, Johnson T, Abderrahmani A, Vollenweider P, Stirnadel H, Sundseth SS, Lai E, Burns DK, Middleton LT, Roses AD, Matthews PM, Waeber G, Cardon L, Waterworth DM, Mooser V. 2009. Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score—the CoLaus Study. Diabetologia 52:600–608.

Miyake K, Yang W, Hara K, Yasuda K, Horikawa Y, Osawa H, Furuta H, Ng MC, Hirota Y, Mori H, Ido K, Yamagata K, Hinokio Y, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Wang HY, Tanahashi T, Nakamura N, Takeda J, Maeda E, Yamamoto K, Tokunaga K, Ma RC, So WY, Chan JC, Kamatani N, Makino H, Nanjo K, Kadowaki T, Kasuga M. 2009. Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. Am J Hum Genet 54:236–241.

Myocardial Infarction Genetics Consortium. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 41:334–341.

Park MY, Hastie T. 2008. Penalized logistic regression for detecting gene interactions. Biostatistics 9:30–50.

Pepe MS. 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffiths AM, Targan SR, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, Rotter JI, Duerr RH, Cho JH, Daly MJ, Brant SR. 2007. Genome-wide association study identifies new susceptibility loci for Crohn's disease and implicates autophagy in disease pathogenesis. Nat Genet 39:596–604.

Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. 2001. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. J Natl Cancer Inst 93:358–366.

Spiegelman D, Colditz GA, Hunter D, Hertzmark E. 1994. Validation of the Gail et al. model for predicting individual breast cancer risk. J Natl Cancer Inst 86:600–607.

The International Schizophrenia Consortium. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288.

Wei Z, Wang K, Qu HG, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SFA, Polychronakos C, Hakonarson H. 2009. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 5:e1000678.

Welcome Trust Case Control Consortium (WTCCC). 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nat Genet 447:661–678.

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25:714–721.

Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, Adami HO, Hsu FC, Zhu Y, Balter K, Kader AK, Turner AR, Liu W, Bleecker ER, Meyers DA, Duggan D, Carpten JD, Chang BL, Isaacs WB, Xu J, Grõnberg H. 2008. Cumulative association of five genetic variants with prostate cancer. N Engl J Med 358:910–919.

Zhong H, Prentice RL. 2008. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics 9:621–634.

Zõllner S, Pritchard JK. 2007. Overcoming the winners curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80:605–615.

Zou H. 2006. The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J R Stat Soc B 67:301–320.