frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 29 February 2012
doi: 10.3389/fgene.2012.00026

Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all!

Paola Sebastiani [1] *, ***Nadia Solovieff*** [2,3,4] ***and Jenny X. Sun*** [1]

[1] *Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA*
[2] *Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetics Research, Massachusetts General Hospital, Boston, MA, USA*
[3] *Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA*
[4] *Department of Psychiatry, Harvard Medical School, Boston, MA, USA*

Edited by:
Albert Tenesa, University of Edinburgh, UK

Reviewed by:
Lei Zhang, University of Shanghai for Science and Technology, China
John Whittaker, GlaxoSmithKline, UK

****Correspondence:***
Paola Sebastiani, Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118, USA.
e-mail: sebas@bu.edu

One of the most popular modeling approaches to genetic risk prediction is to use a summary of risk alleles in the form of an unweighted or a weighted genetic risk score, with weights that relate to the odds for the phenotype in carriers of the individual alleles. Recent contributions have proposed the use of Bayesian classification rules using Naïve Bayes classifiers. We examine the relation between the two approaches for genetic risk prediction and show that the methods are mathematically related. In addition, we study the properties of the two approaches and describe how they can be generalized to include various models of inheritance.

Keywords: genetic risk prediction, genetic score, Naïve Bayes classifier, classification score, classification rule

INTRODUCTION

Several statistical methods have been proposed to capture the complex genetic bases of common diseases. These approaches include standard regression models in which the contribution of several genetic variants is summarized by a genetic risk score (GRS; Meigs et al., 2008; Purcell et al., 2009; Paynter et al., 2010), multivariate regression models and "machine learning type" approaches such as support vector machines (Wei et al., 2009; Wu et al., 2011), Naïve Bayes classifiers (NBC; Okser et al., 2010), classification and regression trees, random forests (Bureau et al., 2005; McKinney et al., 2006), rule induction (Sebastiani and Perls, 2010; Stengard et al., 2010), multifactor dimensionality reduction (Moore et al., 2006), and Bayesian networks (Rodin and Boerwinkle, 2005; Sebastiani et al., 2005; Jiang et al., 2011; Kang et al., 2011a). NBCs use a simple but surprisingly effective Bayesian rule that classifies a subject at risk of a trait if the posterior probability of the trait, given the individual's genetic profile, is maximal (Hand, 2009). The classification rule can be built using a large number of genetic variants, such as single nucleotide polymorphisms (SNPs), by assuming that the SNPs are conditionally independent given the trait (Sebastiani et al., 2012). This hypothesis is often mistaken for "marginal independence" but marginal and conditional independence have no relation (Whittaker, 1990).

In this manuscript we show that there is a mathematical link between NBCs and logistic regression models that use a GRS to summarize the contribution of many SNPs to the susceptibility to a genetic disease. The link between these two approaches also highlights their limitations. We discuss how the directed graphical model underlying a NBC can be extended to include interactions between genes and/or environmental risk factors by maintaining the computations scalable to genome-wide genotype data and even whole genome sequence data.

METHODS AND RESULTS

We describe two approaches – logistic regression and Bayesian classifier – to define a classification score and a rule to be used for genetic risk prediction of a dichotomous trait denoted as T or "not T." The classification score for genetic risk prediction is a function that maps a set of SNPs $\Sigma = \{S_1, \ldots, S_k\}$ into real numbers. The classification rule links the output of the score function to the events T or "not T." Formally, with S denoting the space of SNPs and R the real numbers:

$$\text{Classification score: } \mathrm{Sc}(\Sigma) : S \rightarrow R$$

$$\text{Classification rule: } \mathrm{Sc}(\Sigma) \; > \; \tau \Rightarrow \text{Classify as } T$$

LOGISTIC REGRESSION WITH A GENETIC RISK SCORE

A logistic regression model that includes the general effects of k biallelic SNPs $\Sigma = \{S_1, \ldots, S_k\}$ to model the odds for a dichotomous trait T is defined by the logit equation:

$$\log\left(\frac{p(T|\Sigma)}{1 - p(T|\Sigma)}\right) = \alpha_0 + \sum_{j=1}^{k} (\alpha_{1j} X_{j\mathrm{AB}} + \alpha_{2j} X_{j\mathrm{BB}})$$

$$\text{where } X_{j\mathrm{AB}} = \begin{cases} 1 \text{ if } S_j \text{ genotype} = \mathrm{AB} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{and } X_{j\mathrm{BB}} = \begin{cases} 1 \text{ if } S_j \text{ genotype} = \mathrm{BB} \\ 0 \text{ otherwise} \end{cases}$$

We assume that the alleles of the SNPs are ordered in lexicographical order (A < C < G < T), and A represents the first allele and B the second allele regardless of their frequency. The logit equation is the classification score that can be used to define a classification rule based on a threshold τ:

$$\text{Classification score: } \mathrm{Sc}(\Sigma) = \log\left(\frac{p\,(T|\Sigma)}{1 - p\,(T|\Sigma)}\right) = \alpha_0 + \sum\nolimits_j \left(\alpha_{1j}{}^{X_{j\mathrm{AB}}} + \alpha_{2j}{}^{X_{j\mathrm{BB}}}\right)$$

$$\text{Classification rule: } \mathrm{Sc}(\Sigma) \;>\; \tau \Rightarrow \text{Classify as } T$$

and τ can be determined to optimize sensitivity and specificity by receiver operating characteristic (ROC) curve analysis.

The coefficients of the logistic score are typically estimated by maximum likelihood (McCullagh and Nelder, 1989), or Bayesian methods using large sample approximations or Gibbs sampling (Balding, 2006). By definition, the intercept α_0 represents the log-odds for the trait T for the referent group with all SNPs genotypes equal to AA, while each parameter α_{1j} represents the log-odds ratio for the trait T between the AB genotype and the AA genotype of the jth SNP, and each parameter α_{2j} represents the log-odds ratio for T between the BB and AA genotype of the jth SNP, assuming the other SNP genotypes fixed. When $\alpha_{2j} = 2\alpha_{1j}$ for all $j = 1, \ldots, k$, then the logistic regression encodes the additive effects of the SNPs, and each parameter α_{1j} represents the log-odds ratio for T for each additional copy of the B allele relative to the referent genotype AA.

It is well known that when the data are from a case–control study design, the intercept does not provide the correct estimate of the odds for T in the populations and several corrections have been proposed to limit this problem (Jewell, 2003). Bias of the intercept term is not a problem when the logistic regression model is meant to be used for classification because different intercepts will simply shift the logistic function and classification scores that differ only by the intercept term lead to equivalent classification rules. We state this property formally because it will be used further.

Property 1: Irrelevance of the intercept term of a logistic regression model for classification

Let $\mathrm{Sc}_1(\Sigma)$ and $\mathrm{Sc}_2(\Sigma)$ be two classification scores defined as:

$$\mathrm{Sc}_1(\Sigma) = \log\left(\frac{p\,(T|\Sigma)}{1 - p\,(T|\Sigma)}\right) = \alpha_0 + \sum\nolimits_j \left(\alpha_{1j} X_{j\mathrm{AB}} + \alpha_{2j} X_{j\mathrm{BB}}\right)$$

$$\mathrm{Sc}_2(\Sigma) = \log\left(\frac{p\,(T|\Sigma)}{1 - p\,(T|\Sigma)}\right) = \beta_0 + \sum\nolimits_j \left(\alpha_{1j} X_{j\mathrm{AB}} + \alpha_{2j} X_{j\mathrm{BB}}\right)$$

The two classification scores can be used to define equivalent classification rules by using the relation:

"if $\mathrm{Sc}_1(\Sigma) > \tau \Rightarrow$ classify as T" if and only if
"if $\mathrm{Sc}_2(\Sigma) > \tau + \beta_0 - \alpha_0 \Rightarrow$ classify as T" □

We note however that the correct estimate of the intercept term is necessary to be able to interpret the prediction from the logistic model in terms of prevalence of the trait in the population.

One of the limitations of multivariate logistic regression is that the number of covariates is bounded above by the sample size. It is expected that many common genetic complex traits may be determined by hundreds of genetic variants (Kraft and Hunter, 2009), so that the sample size needed to build reliable logistic regression models for risk prediction can be prohibitively large.

A naïve but very popular alternative is to collapse the contribution of the k SNPs into a GRS to be used in a univariate logistic model. A GRS is typically defined as the weighted sum of the genotypes:

$$\mathrm{GRS} = \mathrm{GRS}(\Sigma) = \sum_{i=1}^{k} \left(w_i X_{i\mathrm{AA}} + v_i X_{i\mathrm{AB}} + z_i X_{i\mathrm{BB}}\right)$$

with weights that can be appropriately chosen. The variables X_{iAB} and X_{iBB} are defined as above, and $X_{iAA} = 1$ if the ith SNP genotype is AA and 0 otherwise. See **Table 1** for a summary of three possible weighting schemes. The GRS is then used as risk factor to define a classification score using a univariate logistic regression:

$$\mathrm{Sc}(\Sigma) = \log\left(\frac{p(T|\mathrm{GRS})}{1 - p(T|\mathrm{GRS})}\right) = \gamma_0 + \gamma_1\,\mathrm{GRS}$$

Case 1. Although this is often referred to as the "unweighted genetic score," the heterozygote genotype is always assigned a weight 1, while the homozygous genotype for the risk allele is assigned weight 2 and the other genotype is assigned weight 0. By adopting this weighting scheme, we are simply counting the number of risk alleles each subject carries. The risk allele of each SNP is determined by a "one-SNP-at-a-time" association analysis, typically under an additive genetic model. Using the same notation and lexicographical order of the SNPs that we used earlier, the risk allele of each SNP will be the A allele if the regression coefficient α_i of the logistic regression model

$$\log\left(\frac{p(T|S_i)}{1 - p(T|S_i)}\right) = \alpha_{0i} + \alpha_i (X_{i\mathrm{AB}} + 2X_{i\mathrm{BB}})$$

is negative, and the B allele if α_i is positive. In the first case ($\alpha_i < 0$), each copy of the B allele decreases the odds for T, while in the second case ($\alpha_i \geq 0$) each copy of the B allele increases the odds for T. With this definition, the GRS is only a function of the different number of risk alleles regardless of their individual genetic effects, and two identical GRS values can represent genetic profiles that are substantially different. See **Figure 1** for an example.

The slope γ_1 in the classification score:

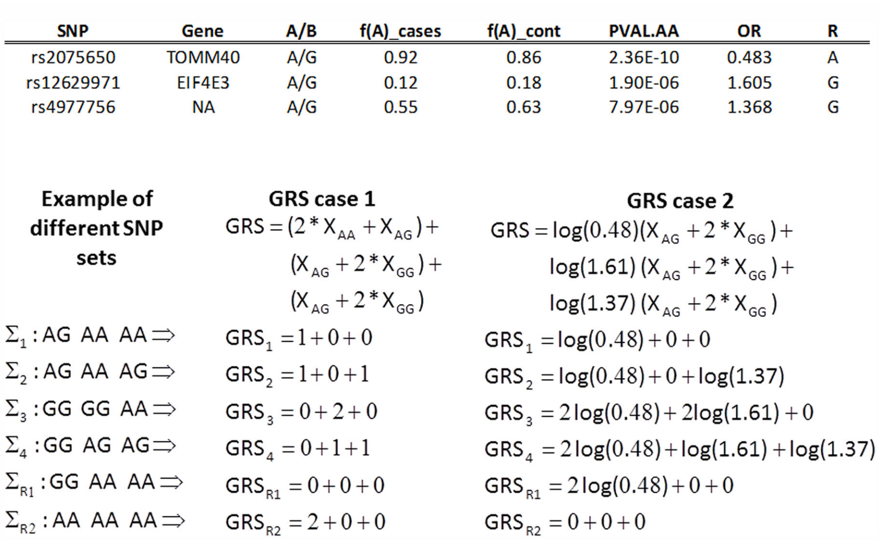
$$\mathrm{Sc}(\Sigma) = \log\left(\frac{p(T|\mathrm{GRS})}{1 - p(T|\mathrm{GRS})}\right) = \gamma_0 + \gamma_1\,\mathrm{GRS} \tag{1}$$

measures the association of the GRS with the trait T in terms of log-odds ratio for T between two GRS that differ by 1, and it is often estimated to test whether the GRS is significantly associated with T. However, the value of γ_1 is irrelevant for classification because two classification scores defined as in Eq. 1 that differ by the slope will produce equivalent classification rules. This is stated in the next property.

Table 1 | Example of choice of weights for the weighted genetic risk score.

Case	w_i	v_i	z_i	Comments
1	$2\delta(A=R)$	1	$2\delta(B=R)$	R denotes the risk allele and $\delta(X=Y)=1$ if $X=Y$ is true and 0 otherwise
2	0	$v_i = \log\dfrac{\frac{p(T\vert X_i=1)}{1-p(T\vert X_i=1)}}{\frac{p(T\vert X_i=0)}{1-p(T\vert X_i=0)}}$	$2v_i$	$X_i=0$ when the ith SNP genotype is AA, and $X_i=1$ when the genotype is AB. This is the standard coding for an additive model
3	0	$\log\dfrac{\frac{p(T\vert S_i=AB)}{1-p(T\vert S_i=AB)}}{\frac{p(T\vert S_i=AA)}{1-p(T\vert S_i=AA)}}$	$\log\dfrac{\frac{p(T\vert S_i=BB)}{1-p(T\vert S_i=BB)}}{\frac{p(T\vert S_i=AA)}{1-p(T\vert S_i=AA)}}$	The two weights represent the log-odds ratio relative to the referent genotype AA. This is the coding for genotypic model.

Case 1 is known as the "unweighted score" and case 2 is typically referred to as the "weighted genetic risk score." Case 3 is the most general and flexible but it does not seem to be used.



SNP	Gene	A/B	f(A)_cases	f(A)_cont	PVAL.AA	OR	R
rs2075650	TOMM40	A/G	0.92	0.86	2.36E-10	0.483	A
rs12629971	EIF4E3	A/G	0.12	0.18	1.90E-06	1.605	G
rs4977756	NA	A/G	0.55	0.63	7.97E-06	1.368	G

Example of different SNP sets	GRS case 1	GRS case 2
	$\mathrm{GRS} = (2*X_{AA}+X_{AG}) + (X_{AG}+2*X_{GG}) + (X_{AG}+2*X_{GG})$	$\mathrm{GRS} = \log(0.48)(X_{AG}+2*X_{GG}) + \log(1.61)(X_{AG}+2*X_{GG}) + \log(1.37)(X_{AG}+2*X_{GG})$
Σ_1 : AG AA AA $\Rightarrow$	$\mathrm{GRS}_1 = 1+0+0$	$\mathrm{GRS}_1 = \log(0.48)+0+0$
Σ_2 : AG AA AG $\Rightarrow$	$\mathrm{GRS}_2 = 1+0+1$	$\mathrm{GRS}_2 = \log(0.48)+0+\log(1.37)$
Σ_3 : GG GG AA $\Rightarrow$	$\mathrm{GRS}_3 = 0+2+0$	$\mathrm{GRS}_3 = 2\log(0.48)+2\log(1.61)+0$
Σ_4 : GG AG AG $\Rightarrow$	$\mathrm{GRS}_4 = 0+1+1$	$\mathrm{GRS}_4 = 2\log(0.48)+\log(1.61)+\log(1.37)$
Σ_{R1} : GG AA AA $\Rightarrow$	$\mathrm{GRS}_{R1} = 0+0+0$	$\mathrm{GRS}_{R1} = 2\log(0.48)+0+0$
Σ_{R2} : AA AA AA $\Rightarrow$	$\mathrm{GRS}_{R2} = 2+0+0$	$\mathrm{GRS}_{R2} = 0+0+0$

FIGURE 1 | Example of GRS (case 1 and case 2 in Table 1) based on three SNPs associated with exceptional longevity. The table on top reports the A/B alleles for the three SNPs, the frequencies of A allele in cases and controls, and the *p*-value for the additive model (Column PVAL.AA) and the odds ratio (OR) for exceptional longevity in carriers of the B allele. The two bottom panels show the calculations of the GRS with weights as in case 1 (left), and case 2 (right). Note that the GRS on the left is only a function of the different number of risk alleles regardless of their individual genetic effects, so the genetic profiles Σ_2, Σ_3, and Σ_4 have the same score while the case 2 GRS assigns different weights to non-referent genotypes and the scores are different. The profile Σ_{R1} denotes the referent group in case 1, while Σ_{R2} denotes the referent group in case 2. The data for this example are taken from Sebastiani et al. (2012).

Property 2: Irrelevance of the slope of a univariate logistic regression model for classification

Let $\mathrm{Sc}_1(\Sigma)$ and $\mathrm{Sc}_2(\Sigma)$ be two classification scores defined as:

$$\mathrm{Sc}_1(\Sigma) = \log\left(\frac{p(T|\mathrm{GRS})}{1-p(T|\mathrm{GRS})}\right) = \gamma_0 + \gamma_1\,\mathrm{GRS}$$

$$\mathrm{Sc}_2(\Sigma) = \log\left(\frac{p(T|\mathrm{GRS})}{1-p(T|\mathrm{GRS})}\right) = \beta_0 + \beta_1\,\mathrm{GRS}$$

The two classification scores can be used to define equivalent classification rules by using the relation:

$$\text{“}\mathrm{Sc}_1(\Sigma) > \tau \Rightarrow \text{classify as T”, if and only if}$$

$$\text{“}\mathrm{Sc}_2(\Sigma) > \beta_0 + \beta_1\frac{\tau-\gamma_0}{\gamma_1} \Rightarrow \text{classify as T”} \qquad \square$$

The GRSs labeled 2 and 3 in **Table 1** weight SNP alleles in different ways to reflect their individual associations with the trait *T*.

Case 2. The GRS can be written as:

$$\mathrm{GRS} = \sum_{i=1}^{k} v_i(X_{i\mathrm{AB}} + 2X_{i\mathrm{BB}})$$

where each weight v_i is the maximum likelihood estimate of the regression coefficient in the univariate logistic regression:

$$\log\left(\frac{p(T|X_i)}{1-p(T|X_i)}\right) = \alpha_{i0} + v_iX_i;\; X_i = \begin{cases} 1 & \text{if } S_i = \mathrm{AB} \\ 2 & \text{if } S_i = \mathrm{BB} \\ 0 & \text{otherwise} \end{cases}$$

that measures the association between SNP S_i and the trait T with an additive genetic model. Therefore, each weight

$$v_{\mathrm{i}} = \log \frac{\dfrac{p(T|X_i = 1)}{1 - p(T|X_i = 1)}}{\dfrac{p(T|X_i = 0)}{1 - p(T|X_i = 0)}}$$

estimates the log-odds ratio for T for each copy of the B allele in an additive genetic model. Note that this formulation of the GRS does not require the specification of the risk allele of the SNPs, and the weighted genetic score will increase by v_i for each copy of the B allele of SNP S_i, if this is a risk allele, and decrease by v_i for each copy of the B allele if this is the protective allele. See the example in **Figure 1**.

The classification score based on this GRS is computed using the logistic regression in Eq. 1, with parameters γ_0, γ_1 that can be estimated by maximum likelihood or Bayesian methods. The slope represents the odds ratio (OR) for T for a unit change of the GRS. In general, the OR for T between two genetic profiles $\Sigma_1 = \{S_{11}, \ldots, S_{k1}\}$ and $\Sigma_2 = \{S_{12}, \ldots, S_{k2}\}$ associated with GRS_1 and GRS_2 is

$$\log \left(\frac{p(T|\mathrm{GRS}_1)/(1 - p(T|\mathrm{GRS}_1)}{p(T|\mathrm{GRS}_2)/(1 - p(T|\mathrm{GRS}_2)} \right) = \gamma_1 \sum_{i=1}^{k} \log \left(\frac{p(T|S_{i1})/(1 - p(T|S_{i1})}{p(T|S_{i2})/(1 - p(T|S_{i2})} \right)$$

and this equation shows that the log-odds ratio for T between two weighted GRSs is an average of log-odds ratios of the individual genetic effects rescaled by the coefficient γ_1.

The classification rule

$$\text{if } \mathrm{Sc}_1(\Sigma) = \log \left(\frac{p(T|\mathrm{GRS})}{1 - p(T|\mathrm{GRS})} \right) > \tau \Rightarrow \text{classify as } T,$$

based on the score

$$\mathrm{Sc}_1(\Sigma) = \log \left(\frac{p\,(T|\mathrm{GRS})}{1 - p\,(T|\mathrm{GRS})} \right) = \gamma_0 + \gamma_1\, \mathrm{GRS}$$

is equivalent to:

$$\text{if } \sum_{i=1}^{k} \log \frac{p(T|S_i)/(1 - p(T|S_i)}{p(T|S_i = \mathrm{AA})/(1 - p(T|S_i = \mathrm{AA})} > \frac{\tau - \gamma_0}{\gamma_1} \Rightarrow \text{classify as } T$$

So the classification rule that uses the weighted GRS in case 2 is essentially based on an average of the individual log-odds ratio for T of each SNP genotype relative to the referent genotypes.

Case 3. The GRS is:

$$\mathrm{GRS} = \sum_{i=1}^{k} (v_i X_{i\mathrm{AB}} + z_i X_{i\mathrm{BB}})$$

where v_i and z_i are the MLE estimate of the regression coefficients of the univariate logistic regression

$$\log \left(\frac{p(T|S_i)}{1 - p(T|S_i)} \right) = \alpha_{i0} + v_i X_{i\mathrm{AB}} + z_i X_{i\mathrm{BB}};$$

$$X_{i\mathrm{AB}} = \begin{cases} 1 & \text{if } S_i = \mathrm{AB} \\ 0 & \text{otherwise} \end{cases}; X_{i\mathrm{BB}} = \begin{cases} 1 & \text{if } S_i = \mathrm{BB} \\ 0 & \text{otherwise} \end{cases}$$

that measures the genotypic association between SNP S_i and the trait T. Therefore

$$v_i = \log \frac{\dfrac{p(T|S_i = \mathrm{AB})}{1 - p(T|S_i = \mathrm{AB})}}{\dfrac{p(T|S_i = \mathrm{AA})}{1 - p(T|S_i = \mathrm{AA})}}; \; z_i = \log \frac{\dfrac{p(T|S_i = \mathrm{BB})}{1 - p(T|S_i = \mathrm{BB})}}{\dfrac{p(T|S_i = \mathrm{AA})}{1 - p(T|S_i = \mathrm{AA})}}$$

are the log-odds ratio for T between the AB and AA genotypes, and BB and AA genotypes. See **Figure 2** for an example. The classification score and classification rule are derived as in case 2 and can be interpreted as average of the log-odds ratios of individual SNPs genotypes. Compared to case 2, the weights based on genotype associations allow for more general model of associations that are not restricted to linear increase of the log-odds for T. Note also that when the SNPs included in a GRS (case 2 and 3) are independent, the two scores should be approximately equivalent to multivariate logistic regression with additive (case 2) or genotypic association (case 3). In addition, if the SNPs included in the GRS have similar effects, then the GRS in case 1 and 2 should be approximately equivalent.

NAÏVE BAYES CLASSIFIERS

The classification score based on a NBC is the posterior probability of the trait T that is calculated using the formula:

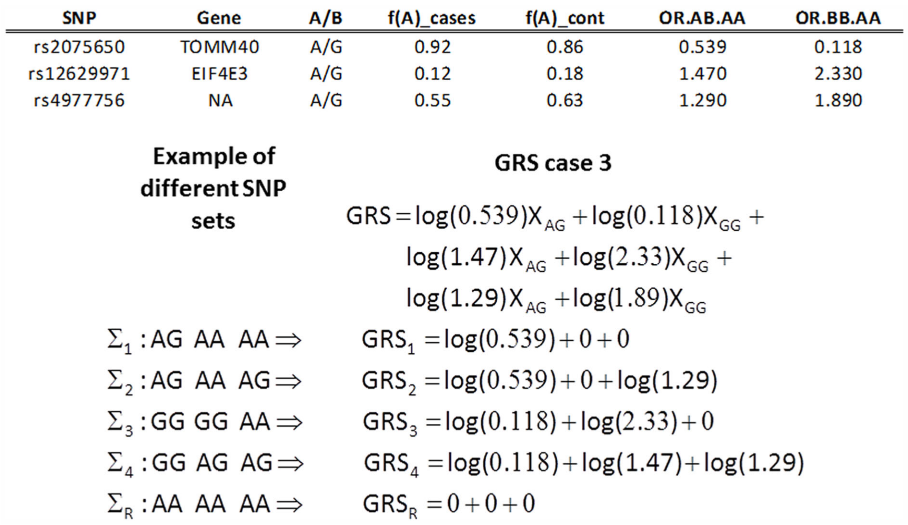
$$\mathrm{Sc}(\Sigma) = p(T|\Sigma) = \frac{p(T) \prod_{i=1}^{k} p(S_i|T)}{p(T) \prod_{i=1}^{k} p(S_i|T) + (1 - p(T)) \prod_{i=1}^{k} p(S_i|notT)}$$

where $p(T)$ and $1 - p(T)$ are the prior probabilities of having the trait T or not. The conditional probabilities $p(S_i \mid T)$ and $p(S_i \mid \text{not } T)$ represent the distribution of the ith SNP genotype in subjects with and without the trait T. They are typically estimated assuming genotypic association (Sebastiani et al., 2012), but they could also be estimated using an additive genetic model. The formula is derived using Bayes' theorem and assuming that the SNPs are independent, conditionally on T (Hand, 2009). The usual Bayesian classification rule is to classify a subject with the most probable outcome

$$\text{if } \mathrm{Sc}(\Sigma) > 0.5 \Rightarrow \text{classify as } T.$$

This rule is based on a 0–1 loss that assigns the same weight to misclassification errors. A general loss function that weights differently sensitivity and specificity would lead to the classification rule:

$$\text{if } \mathrm{Sc}(\Sigma) > \frac{\lambda}{1 + \lambda} \Rightarrow \text{classify as } T \text{ for } \lambda > 0$$



SNP	Gene	A/B	f(A)_cases	f(A)_cont	OR.AB.AA	OR.BB.AA
rs2075650	TOMM40	A/G	0.92	0.86	0.539	0.118
rs12629971	EIF4E3	A/G	0.12	0.18	1.470	2.330
rs4977756	NA	A/G	0.55	0.63	1.290	1.890

Example of different SNP sets

GRS case 3

$$\mathrm{GRS} = \log(0.539)X_{AG} + \log(0.118)X_{GG} + \log(1.47)X_{AG} + \log(2.33)X_{GG} + \log(1.29)X_{AG} + \log(1.89)X_{GG}$$

$$\Sigma_1 : AG\ AA\ AA \Rightarrow \quad \mathrm{GRS}_1 = \log(0.539) + 0 + 0$$

$$\Sigma_2 : AG\ AA\ AG \Rightarrow \quad \mathrm{GRS}_2 = \log(0.539) + 0 + \log(1.29)$$

$$\Sigma_3 : GG\ GG\ AA \Rightarrow \quad \mathrm{GRS}_3 = \log(0.118) + \log(2.33) + 0$$

$$\Sigma_4 : GG\ AG\ AG \Rightarrow \quad \mathrm{GRS}_4 = \log(0.118) + \log(1.47) + \log(1.29)$$

$$\Sigma_R : AA\ AA\ AA \Rightarrow \quad \mathrm{GRS}_R = 0 + 0 + 0$$

FIGURE 2 | Example of GRS (case 3 in Table 1). The table on top reports the A/B alleles for the three SNPs, the frequencies of A allele in cases and controls, and the odds ratio for exceptional longevity in carriers of the AB allele relative to carriers of the AA allele (OR.AB.AA), and the odds ratio for exceptional longevity in carriers of the BB allele relative to carriers of the AA allele (OR.BB.AA). The bottom panel shows the calculations of the GRS with weights as in case 3. The profile Σ_R denotes the referent group.

that can also be written as:

$$\mathrm{Sc}(\Sigma) > \frac{\lambda}{1+\lambda} \Leftrightarrow \log\left(\frac{p(T|\Sigma)}{1-p(T|\Sigma)}\right) > \log(\lambda)$$

and simple algebra shows that this is equivalent to:

$$\log\left(\frac{p(T|\Sigma)}{1-p(T|\Sigma)}\right) = \log\left(\frac{p(T)\prod_{i=1}^{k} p(S_i|T)}{(1-p(T))\prod_{i=1}^{k} p(S_i|\text{not } T)}\right) =$$
$$\log\left(\frac{\prod_{i=1}^{k} p(T)p(S_i|T)}{\prod_{i=1}^{k}(1-p(T))p(S_i|\text{not } T)}\right) =$$
$$\log\left(\prod_{i=1}^{k}\frac{p(T|S_i)}{1-p(T|S_i)}\right) = \sum_{i=1}^{k}\log\left(\frac{p(T|S_i)}{1-p(T|S_i)}\right) > \log(\lambda)$$

As long as the log-odds ratios are calculated using the same genetic model, this classification rule is equivalent to the classification rule based on the GRS (either case 2 or 3)

$$\text{if } \sum_{i=1}^{k}\log\frac{p(T|S_i)/(1-p(T|S_i))}{p(T|S_i=\mathrm{AA})/(1-p(T|S_i=\mathrm{AA}))} > \frac{\tau-\gamma_0}{\gamma_1}$$
$$\Rightarrow \text{classify as } T$$

by setting the threshold

$$\tau = \gamma_0 - \gamma_1\sum_{i=1}^{k}\log\left(\frac{p(T|S_i=\mathrm{AA})}{1-p(T|S_i=\mathrm{AA})}\right) + \gamma_1\log(\lambda)$$

We state this relation formally.

Property 3: Equivalence of classification rules based on the GRS and the NBC

The classification rules based on a logistic model of a GRS(case 2 or 3) and a NBC are equivalent when the same genetic models are used to link individual SNPs to the trait.

The details of the algebraic manipulations are in Section "Appendix." □

Note that the equivalence between the classification rules based on a NBC and a logistic regression model with a GRS as in case 2 or 3 is a simple consequence of the fact that both models base the prediction on a weighted average of ORs of the individual SNPs. This equivalence is independent of the choice of the prior for T because different prior distributions will lead to equivalent classification rules but with different classification thresholds. Also, the equivalence of classification rules based on GRS and NBC implies that when alternative classifiers are compared by the area under the receiving operator curve they must reach the same value. This is shown in the next example.

Example. To demonstrate the connection between the NBC and the GRS in case 3, we performed a simple simulation. We simulated a dataset with 3000 cases and 3000 controls, and genotype data from 75 causal SNP and 500,000 null SNPs. For the null SNPs, we randomly selected frequencies of the minor allele (p) from a uniform (0.05, 0.5) distribution and genotype frequencies were generated assuming Hardy–Weinberg equilibrium $[p^2, 2p(1-p), (1-p)^2]$. The causal SNPs were simulated with ORs of 1.2, 1.3, 1.4, 1.5, and 1.6 and minor allele frequencies (MAFs) of 0.1, 0.2, 0.3, 0.4, and 0.5. A causal SNP was simulated for each combination of the above ORs and MAFs (25 combinations) under an additive, recessive and dominant mode of inheritance (25 combinations × 3 modes of inheritance = 75 SNPs). The genotype frequencies in controls were generated to follow Hardy–Weinberg equilibrium $[p^2,2p(1-p),(1-p)^2]$. The genotype frequencies in cases for the additive, recessive, and dominant models were $[p^2,2\text{OR}p(1-p),\text{OR}^2(1-p)^2]$, $[p^2,2p(1-p),\text{OR}(1-p)^2]$ and $[p^2,2\text{OR}p(1-p),\text{OR}(1-p)^2]$, respectively. For the cases, the genotype frequencies were divided by the sum of the frequencies so that the frequencies add up to 1. Using the genotype frequencies for each SNP, we simulated a discovery set of 3000 cases and 3000 controls and a replication set with the same sample sizes.

The data in the discovery set were analyzed to generate genetic risk models based on GRS and NBCs in the following way. A Bayesian genome-wide association study was performed on the discovery set and SNPs were ordered according to the posterior probability for the genotypic association to build nested NBCs with increasing number of SNPs as in Sebastiani et al. (2012). To obtain the weights for the three GRSs, we ran two logistic regression models for each SNP, using an additive mode of inheritance and a genotypic mode of inheritance. The results of these analyses were used to detect the risk alleles of SNPs for nested GRS as in case 1; and to estimate the weights of GRS as in cases 2 and 3. Using SNPs ordered by the posterior probability for the genotypic association, we then built three sets of classification models based on logistic regression and the three different GRS, with increasing number of SNPs. The prediction models were tested on the replication set to avoid issues of over-fitting. The simulation described above was repeated five times and the mean AUC across the replicates was used to assess accuracy.

Figure 3 (left panel) shows the mean AUC across five replicates for the NBCs and logistic regression models for different GRSs, with increasing number of SNPs. As expected based on our mathematical calculations, the AUCs of the genetic risk models based on the NBCs and the GRSs with a genotypic weights are identical (**Figure 3**, left panel), and the predicted probabilities are almost identical (**Figure 3**, right panel). The weighted and unweighted GRS using an additive mode of inheritance have lower AUCs demonstrating the loss of accuracy with assuming additivity when some of the SNPs do not follow an additive mode of inheritance. Of course if all SNPs do in fact follow an additive model of the inheritance, the genotypic and additive prediction models would perform similarly. The trend of the AUC shows that accuracy keeps increasing as true positive SNPs are included in the model, and then declines when each classification model starts including false positive SNPs. The decline is more evident for the case 1 GRS, while both weighted GRS based on additive or genotypic associations appear to be more robust.

DISCUSSION

One of the selling points of genome-wide association studies was to discover genetic variants that are associated with increased susceptibility for disease and could be used for personalized diagnosis and prognosis. Initial results published for example in Meigs et al. (2008) and Paynter et al. (2010) however showed that genetic data added limited predicted values to well established risk factors of Type II diabetes and cardiovascular disease. These initial studies

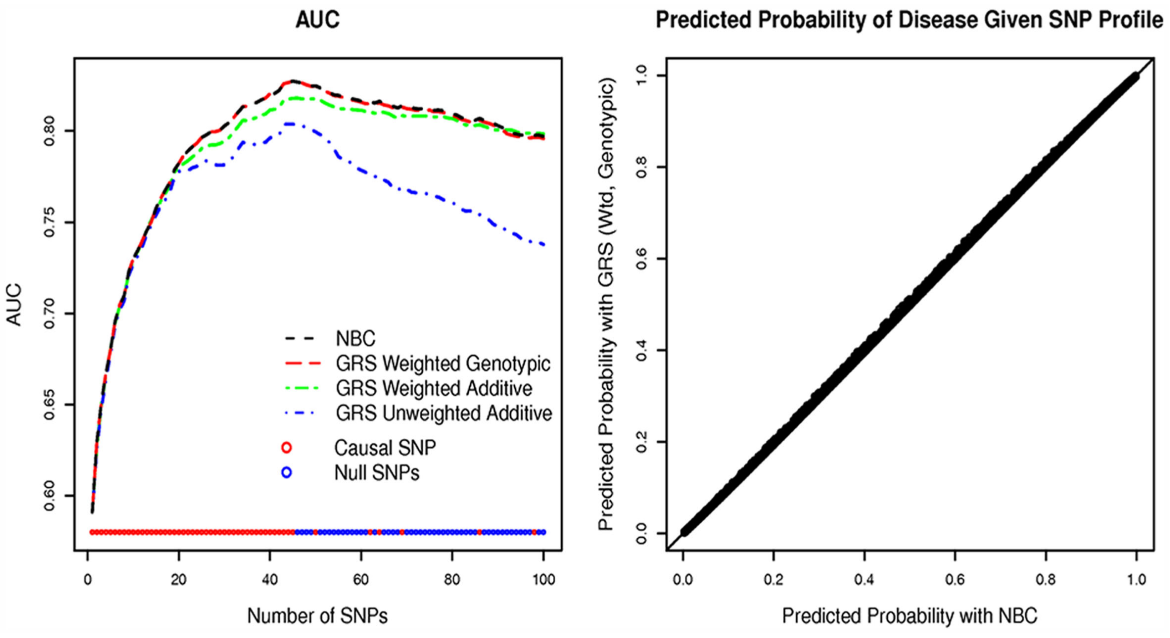


FIGURE 3 | Results of simulation for replication set. The left hand plot graphs the mean area under the ROC (AUC) versus the number of SNPs in the prediction model. The colored lines refer to the AUC of the NBC (black), the unweighted GRS from an additive model (case 1, blue), the weighted GRS from an additive model (case 2, green), and the weighted GRS from a genotypic model (case 3, red). The maximum AUC occurs at 45 SNPs. The right hand plot graphs the probability of the trait *T* given the weighted GRS (genotypic model) on the *y*-axis versus the probability of disease given the SNP profile estimated by NBC on the *x*-axis for a model containing 45 SNPs for one of the replicates. The right hand plot is similar across the five replicates.

limited the attention to those SNPs that reached genome-wide significance and their effect was summarized into a GRS. Since then, a growing body of literature has shown the increased value of deeper mining of genome-wide association studies but inclusion of large number of SNPs in genetic risk model has continued to resort on GRSs (Cui, 2009; Goddard et al., 2009; Kooperberg et al., 2009; Purcell et al., 2009; Yang et al., 2010; Chen et al., 2011; Chibnik et al., 2011), while machine learning type methods continue to be rare regardless of some successful applications (Wei et al., 2009; Okser et al., 2010; Kang et al., 2011b; Sebastiani et al., 2012).

Our study shows that risk prediction based on a GRS is mathematically equivalent to risk prediction based on a NBC, when the same SNPs with the same mode of inheritance are used in the models. The equivalence is based on the fact that both models essentially base the prediction on a weighted average of ORs of the individual SNPs. While this equivalence establishes the validity of methods based on the NBC for genetic risk prediction and we hope will contribute to make this approach more popular in this field, it also shows that contrary to what stated in Okser et al. (2010) a NBC does not include interactions of SNPs but only additive genetic effects. However, the directed graphical model underlying a NBC can be extended to more general structures to include interactions between genes and/or environmental risk factors by maintaining the computations scalable to genome-wide genotype data and even whole genome sequence data (Sebastiani and Perls, 2008).

Figure 4 shows some ways to extend NBCs for risk prediction to include population ancestry, as well as genetic and non-genetic effects that may be missed by test for marginal associations. **Figure 4A** describes a directed acyclic graph (DAG) with one parent node (T) and two children nodes (X_1 and X_2) that may represent SNPs. The DAG describes the conditional independence of X_1 and X_2 given T. This type of DAG with one root node and multiple conditionally independent children represents a NBC (Sebastiani and Abad-Grau, 2007). The DAG in **Figure 4B** extends the NBC in **Figure 4A** with an additional node X_3 that is marginally independent of T, but conditionally dependent on T given X_2. In the context of genetic risk modeling, the node X_3 could represent a non-genetic risk factor that is associated with a trait T only in specific genetic backgrounds (the node X_2). The DAG in **Figure 4C** includes an additional node X_4 that is conditionally independent of all other nodes given X_1. This additional node may represent a gene × gene interaction that is induced by linkage disequilibrium. Note that both DAGS in **Figures 4B,C** would give the same classification score for T, because of the independence of T from X_4 given X_1. So, the DAG in **Figure 4C** would be useful for a better explanation of the biology rather than improving genetic risk prediction. Finally, the DAG in **Figure 4D** extends the DAG in **Figure 4B** by adding a link from T to X_3. The inclusion of this link makes the node X_3 marginally dependent of T and interaction between X_2 and X_3 changes the classification score compared to the DAG in **Figure 4B**.

In addition, and most importantly, the fact that all variables in a DAG are random provides a sound framework for marginal and conditional inference. For example, a genetic risk model based on a DAG can be used for predicting the outcome of a subject by marginalizing out unobserved variables (Solovieff et al., 2011).

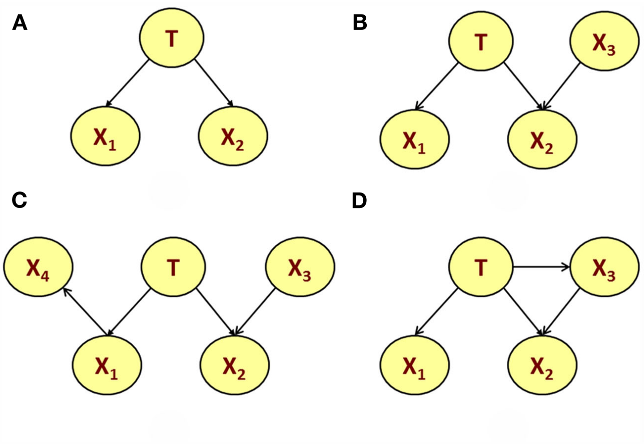


FIGURE 4 | Examples of directed acyclic graph (DAG). All nodes are random variables and the DAG represents Markov properties of marginal and conditional independence (Lauritzen and Sheehan, 2004). In particular, the global Markov property states that a node is independent of all other nodes in the DAG given its parent nodes, its children nodes and additional parents of its children (Lauritzen and Sheehan, 2004). In addition, two nodes are marginally independent when they have no directed joining paths after their children are dropped. Therefore, the nodes X_1 and X_2 in the DAG in **(A)** are conditionally independent given T. The DAG in **(B)** adds the node X_3 to the NBC in **(A)**. This additional node is marginally independent of T but conditionally dependent on T given X_2. The DAG in **(C)** includes an additional node X_4 that is conditionally independent of all other nodes given X_1. Finally, the DAG in **(D)** extends the DAG in **(B)** by adding a link from T to X_3 so that X_3 and T are marginally dependent.

Our analysis is limited to binary outcomes, but we expect that similar results hold when the outcome to be predicted is a quantitative trait that follows a normal distribution. Furthermore, our analysis shows that linear transformations of a GRS do not impact predictive accuracy, and similarly, that the predictive accuracy of a NBC cannot be changed by a choice of prior for T. Improving the accuracy can be accomplished by selection of the most predictive SNP and by choosing alternative weights to calculate the GRS. There is no obvious similar choice for a NBC. However, a closely related approach that we used in Sebastiani et al. (2012) to improve the predictive accuracy is to use ensemble of nested NBCs. Finally, the machine learning community has developed many feature selection algorithms for building classifiers (Hastie et al., 2009) that, by the equivalence proved in this paper, may prove to be useful to generate better genetic risk models.

ACKNOWLEDGMENTS

Research supported with funds from NIH/NHLBI R01 HL089655-03 and R21HL114237 (Paola Sebastiani). We thank Dr. Maria Abad Grau for stimulating discussion during her visit at Boston University in August 2011 that triggered the writing of this manuscript.

REFERENCES

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* 7, 781–791.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Van Eerdewegh, P. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.* 28, 171–182.

Chen, H., Poon, A., Yeung, C., Helms, C., Pons, J., Bowcock, A. M., Kwok, P. Y., and Liao, W. (2011). A genetic risk score combining ten psoriasis risk loci improves disease prediction. *PLoS ONE* 6, e19454. doi:10.1371/journal.pone.0019454

Chibnik, L. B., Keenan, B. T., Cui, J., Liao, K. P., Costenbader, K. H., Plenge, R. M., and Karlson, E. W. (2011). Genetic risk score predicting risk of rheumatoid arthritis phenotypes and age of symptom onset. *PLoS ONE* 6, e24380. doi:10.1371/journal.pone.0024380

Cui, J. (2009). Overview of risk prediction models in cardiovascular disease research. *Ann Epidemiol* 19, 711–717.

Goddard, M. E., Wray, N. R., Verbyla, K., and Visscher, P. M. (2009). Estimating effects and making predictions from genome-wide marker data. *Stat. Sci.* 24, 517–529.

Hand, D. J. (2009). "Naive Bayes," in *The Top Ten Algorithms in Data Mining*, eds X. Wu and V. Kumar (London: Chapman and Hall), 163–178.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Jewell, R. (2003). *Statistics for Epidemiology.* Boca Raton: CRC/Chapman and Hall.

Jiang, X., Barmada, M. M., Cooper, G. F., and Becich, M. J. (2011). A Bayesian method for evaluating and discovering disease loci associations. *PLoS ONE* 6, e22075. doi:10.1371/journal.pone.0022075

Kang, J., Zheng, W., Li, L., Lee, J., Yan, X., and Zhao, H. (2011a). Use of Bayesian networks to dissect the complexity of genetic disease: application to the Genetic Analysis Workshop 17 simulated data. *BMC Proc.* 5(Suppl. 9), S37. doi:10.1186/1753-6561-5-S9-S37

Kang, J., Kugathasan, S., Georges, M., Zhao, H., and Cho, J. H. (2011b). Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum. Mol. Genet.* 20, 2435–2442.

Kooperberg, C., LeBlanc, M., and Obenchain, V. (2009). Risk prediction using genome-wide association studies. *Genet. Epidemiol.* 34, 643–652.

Kraft, P., and Hunter, D. J. (2009). Genetic risk prediction – are we there yet? *N. Engl. J. Med.* 360, 1701–1703.

Lauritzen, S. L., and Sheehan, N. A. (2004). Graphical models for genetic analysis. *Stat. Sci.* 18, 489–514.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models.* London: Chapman and Hall.

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions: a review. *Appl. Bioinformatics* 5, 77–88.

Meigs, J. B., Shrader, P., Sullivan, L. M., McAteer, J. B., Fox, C. S., Dupuis, J., Manning, A. K., Florez, J. C., Wilson, P. W., D'Agostino, R. B. Sr., and Cupples, L. A. (2008). Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* 359, 2208–2219.

Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N., and White, B. C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.* 241, 252–261.

Okser, S., Lehtimaki, T., Elo, L. L., Mononen, N., Peltonen, N., Kahonen, M., Juonala, M., Fan, Y. M., Hernesniemi, J. A., Laitinen, T., Lyytikainen, L. P., Rontu, R., Eklund, C., Hutri-Kahonen, N., Taittonen, L., Hurme, M., Viikari, J. S., Raitakari, O. T., and Aittokallio, T. (2010). Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young Finns study. *PLoS Genet.* 6, e1001146. doi:10.1371/journal.pgen.1001146

Paynter, N. P., Chasman, D. I., Pare, G., Buring, J. E., Cook, N. R., Miletich, J. P., and Ridker, P. M. (2010). Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA* 303, 631–637.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752.

Rodin, A. S., and Boerwinkle, E. (2005). Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels). *Bioinformatics* 21, 3273–3278.

Sebastiani, P., and Abad-Grau, M. (2007). "Bayesian networks for genetic analysis," in *Bioinformatics: An Engineering Case-Based Approach*, eds G. Alterovitz and M. F. Ramoni (Cambridge, MA: Artech House), 205–228.

Sebastiani, P., and Perls, T. T. (2008). "Complex genetic models," in *Bayesian Networks*, eds O. Pourret, P. Naïm, and B. Marcot (Chichester: John Wiley & Sons), 53–72.

Sebastiani, P., and Perls, T. T. (2010). Prediction models that include genetic data. *Circ. Cardiovasc. Genet.* 3, 1–2.

Sebastiani, P., Ramoni, M. F., Nolan, V., Baldwin, C. T., and Steinberg, M. H. (2005). Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat. Genet.* 37, 435–440.

Sebastiani, P., Solovieff, N., DeWan, A., Walsh, K., Puca, A., Hartley, S. W., Melista, E., Andersen, S., Dworkis, D. A., Wilk, J. B., Myers, R. H., Steinberg, M. H., Montano, M., Baldwin, C. T., Hoh, J., and Perls, T. T. (2012). Genetic signatures of exceptional longevity in humans. *PLoS ONE* 7, e29848. doi:10.1371/journal.pone.0029848

Solovieff, N., Baldwin, C. T., Steinberg, M. H., Perls, T. T., and Sebastiani, P. (2011). "Incorporating genetic ancestry into risk prediction models," in *The 12th International Congress of Human Genetics and the American Society of Human Genetics 61st Annual Meeting*, Montreal.

Stengard, J. H., Dyson, G., Frikke-Schmidt, R., Tybjaerg-Hansen, A., Nordestgaard, B. G., and Sing, C. F. (2010). Context-dependent associations between variation in risk of ischemic heart disease and variation in the 5′ promoter region of the apolipoprotein E gene in Danish women. *Circ. Cardiovasc. Genet.* 3, 22–30.

Wei, Z., Wang, K., Qu, H. Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J. T., Chiavacci, R., Stanley, C., Monos, D., Grant, S. F., Polychronakos, C., and 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. doi: 10.1371/journal.pgen.1000678

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* New York: John Wiley & Sons.

Wu, C., Walsh, K., DeWan, A., Hoh, J., and Wang, Z. (2011). Disease risk prediction with rare and common variants. *BMC Proc.* 5(Suppl. 9), S61. doi:10.1186/1753-6561-5-S9-S61

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., Hill, W. G., Landi, M. T., Alonso, A., Lettre, G., Lin, P., Ling, H., Lowe, W., Mathias, R. A., Melbye, M., Pugh, E., Cornelis, M. C., Weir, B. S., Goddard, M. E., and Visscher, P. M. (2010). Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* 43, 519–525.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 24 December 2011; accepted: 12 February 2012; published online: 29 February 2012.

Citation: Sebastiani P, Solovieff N and Sun JX (2012) Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all! Front. Gene. ***3****:26. doi: 10.3389/fgene.2012.00026*

This article was submitted to Frontiers in Statistical Genetics and Methodology, a specialty of Frontiers in Genetics.

APPENDIX

DERIVATION OF PROPERTY 3

$$\sum_{i=1}^{k} \log \frac{p(T|S_i)/(1-p(T|S_i)}{p(T|S_i=\mathrm{AA})/(1-p(T|S_i=\mathrm{AA}))} > \frac{\tau-\gamma_0}{\gamma_1} \Rightarrow \text{classify as } T$$

if and only if

$$\sum_{i=1}^{k} \log\left(\frac{p(T|S_i)}{1-p(T|S_i)}\right) - \sum_{i=1}^{k} \log\left(\frac{p(T|S_i=\mathrm{AA})}{1-p(T|S_i=\mathrm{AA})}\right) > \frac{\tau-\gamma_0}{\gamma_1} \Rightarrow \text{classify as } T$$

if and only if

$$\sum_{i=1}^{k} \log\left(\frac{p(T|S_i)}{1-p(T|S_i)}\right) > \frac{\tau-\gamma_0}{\gamma_1} + \sum_{i=1}^{k} \log\left(\frac{p(T|S_i=\mathrm{AA})}{1-p(T|S_i=\mathrm{AA})}\right) \Rightarrow \text{classify as } T$$

if and only if

$$\sum_{i=1}^{k} \log\left(\frac{p(T|S_i)}{1-p(T|S_i)}\right) > \log(\lambda)$$

where $\log(\lambda) = \dfrac{\tau-\gamma_0}{\gamma_1} + \displaystyle\sum_{i=1}^{k} \log\left(\frac{p(T|S_i=\mathrm{AA})}{1-p(T|S_i=\mathrm{AA})}\right)$

and $\tau = \gamma_0 - \gamma_1 \displaystyle\sum_{i=1}^{k} \log\left(\frac{p(T|S_i=\mathrm{AA})}{1-p(T|S_i=\mathrm{AA})}\right) + \gamma_1 \log(\lambda)$