Int J Mol Epidemiol Genet 2015;6(1):1-8
www.ijmeg.org /ISSN:1948-1756/IJMEG0012842

Original Article

Evaluation of genetic risk scores for prediction of dichotomous outcomes

Wonsuk Yoo[1], Selina A Smith[1], Steven S Coughlin[2]

[1]*Institute of Public and Preventive Health, and Department of Dental Medicine, Georgia Regents University, Augusta, GA;* [2]*Department of Community Health and Sustainability, Division of Public Health, University of Massachusetts, Lowell, MA*

Received July 13, 2015; Accepted August 23, 2015; Epub September 9, 2015; Published September 15, 2015

Abstract: Substantial uncertainty exists as to whether combining multiple disease-associated single nucleotide polymorphisms (SNPs) into a genotype risk score (GRS) can improve the ability to predict the risk of disease in a clinically relevant way. We calculated the ability of a simple count GRS to predict the risk of a dichotomous outcome under both multiplicative and additive models of combined effects. We then compared the results of these simulations with the observed results of published GRS measured within multiple epidemiologic cohorts. If the combined effect of each disease-associated SNP included in a GRS is multiplicative on the risk scale, then a count GRS score should be useful for risk prediction with as few as 10-20 SNPs. Adding additional SNPs to the GRS under this model dramatically improves risk prediction. By contrast, if the combined effect of each SNP included in a GRS is linearly additive on the risk scale, a simple count GRS is unlikely to provide clinically useful risk prediction. Adding additional SNPs to the GRS under this model does not improve risk prediction. The combined effect of SNPs included in several published GRS measured in several well-phenotyped epidemiologic cohort studies appears to be more consistent with a linearly additive effect. A simple count GRS is unlikely to be clinically useful for predicting the risk of a dichotomous outcome. Alternative methods for constructing GRS that attempt to identify and include SNPs that demonstrate multiplicative gene-gene or gene-environment interactive effects are needed.

Keywords: Genotype risk score (GRS), risk prediction, multiple disease-associated single nucleotide polymorphisms, simple count GRS, multiplicative or additive on risk scale, simulations, dichotomous outcomes

Introduction

Multiple single nucleotide polymorphisms (SNPs) have been associated with complex diseases such as obesity, cancer, diabetes, or cardiovascular diseases. Genetic polymorphisms associated with such complex diseases have been investigated in studies generally focused on a priori selected candidate genes. Advances in genomic technologies have made it possible to genotype and evaluate many SNPs throughout the human genome to identify novel susceptibility genes. The genetic risk score (GRS) approach has frequently been used to aggregate the contribution of multiple SNPS (combining genetic information) and to test for improved performance in predicting incidence of disease. However, conflicting results have been obtained. Several studies showed that information about multiple SNPs combined into a GRS was associated with complex diseases such as obesity, type 2 diabetes, coronary heart disease, etc [1-3]. However, other studies have shown that the usage of GRS methods did not improve risk prediction [4-7].

Recent simulation studies have demonstrated some interesting features of genetic profiles that explain why the predictive value of a larger number of multiple weak susceptibility variants may be difficult to improve [8, 9]. Most simulation studies assume multiplicative risk effects and suggest that 10 or 20 SNPs with marginal effects should be useful in risk prediction [8-10]. However, published results of the predictive ability of genotype risk score derived from epidemiologic data have been disappointing. This simulation study aims to further evaluate the benefit of analyzing multiple genetic polymorphisms with a small marginal risk effect using genetic risk scores for risk prediction of complex diseases.

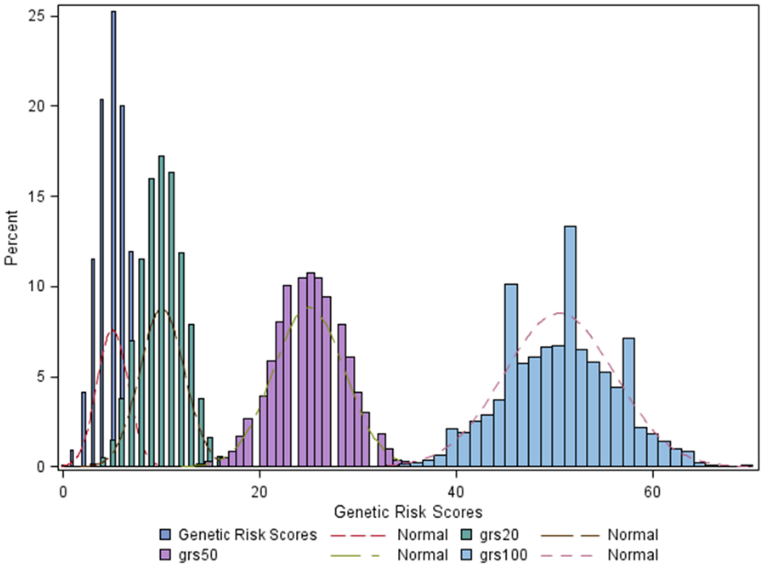


Figure 1. The distribution of GRSs for four different genetic profiles of 10 SNPs, 20 SNPs, 50 SNPs, and 100 SNPs.

Methods

Modeling strategy

The modeling procedure has been published in detail elsewhere [11, 12], but is briefly summarized here. The strategy for modeling has three steps: (1) modeling the genetic profiles of all subjects, (2) calculating the disease risks associated with the genetic profiles, and (3) defining the disease status of all subjects. In order to construct the genetic profiles, we assumed that all genotypes and allele proportions were in Hardy-Weinberg equilibrium, and genes were generated to be independent (no linkage disequilibrium). We constructed the genetic profiles by randomly assigning the genotypes of each genetic variant to all individuals, so that the genotype distributions are in line with the specified genotype frequencies. Our interest is to determine the risk of disease for each subject associated with GRS. The cumulative genetic information is obtained by summing the number of risk alleles for each individual in order to evaluate the aggregated genetic effect on risk prediction [2, 4, 13]. We considered two types of risk effects models (“additive effects model” and “multiplicative effects model”) in determining each individual’s risk of disease. The additive effects model assume that the combined effect of each SNP included in a GRS is linearly additive on the risk scale while the multiplicative effects model indicate that the combined effect is log-linearly additive (multiplicative) on the risk scale. For assigning disease status, we assumed that subjects with high disease risks are more likely to be assigned to the group that will develop disease than those with lower risk. A disease status was generated by comparing the disease risk of each subject to a randomly drawn value between 0 and 1 from a uniform distribution.

Simulation setting

We simulated four different data sets of SNPs, 10, 20, 50, and 100, which include information on genetic profiles, disease rates and disease status for 100,000 subjects for both additive risk effects and multiplicative risk effects models. Each SNP assumes 50% frequency of heterogeneity and 25% increased relative risk (RR) for each additional genetic variant of dichotomous disease (RR_{GRS}=1.25). For both effects models, GRS_i and $mGRS$ are defined to be the number of risk alleles and the mean value of GRS for *i*th subject, respectively. RR_i is the relative risk for each subject while RR_{mGRS} is the relative risk associated with the mean of GRS. For additive risk score model, the relative risks for each subject are calculated by

$$RR_i = 1 + (RR_{GRS} - 1) \times GRS_i$$

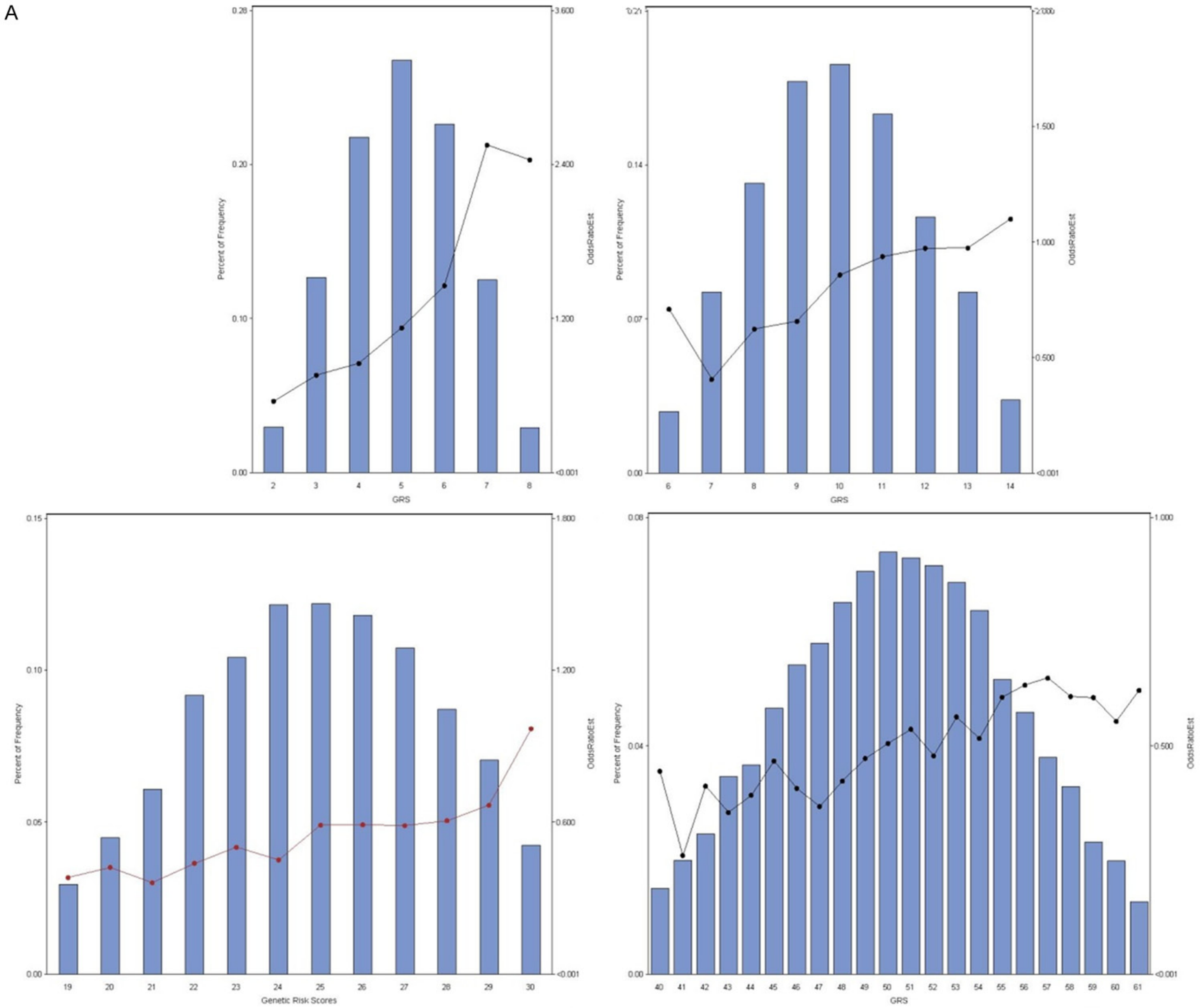
and then the disease risk rate for each subject, X_i, is calculated as follows:

$$X_i = \left(\frac{1 + (RR_{GRS} - 1) \times GRS_i}{1 + (RR_{GRS} - 1) \times mGRS}\right) \times X_{pop}$$

where X_{pop} is the assumed population disease rate (20%). The disease rate under multiplicative effects models is calculated by

$$X_i = \left(\frac{RR^{GRS_i}}{RR^{mGRS}}\right) \times X_{pop}$$

The disease rate is the ratio of individual’s relative risk associated with GRS of each subject over the relative risk associated with the mean value of GRS.


A
0.28
0.20
0.10
0.00
Percent of Frequency
3.600
2.400
1.200
<0.001
OddsRatioEst
2
3
4
5
6
7
8
GRS
0.21
0.14
0.07
0.00
Percent of Frequency
2.000
1.500
1.000
0.500
<0.001
OddsRatioEst
6
7
8
9
10
11
12
13
14
GRS
0.15
0.10
0.05
0.00
Percent of Frequency
1.800
1.200
0.600
<0.001
OddsRatioEst
19
20
21
22
23
24
25
26
27
28
29
30
Genetic Risk Scores
0.08
0.04
0.00
Percent of Frequency
1.000
0.500
<0.001
OddsRatioEst
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
GRS

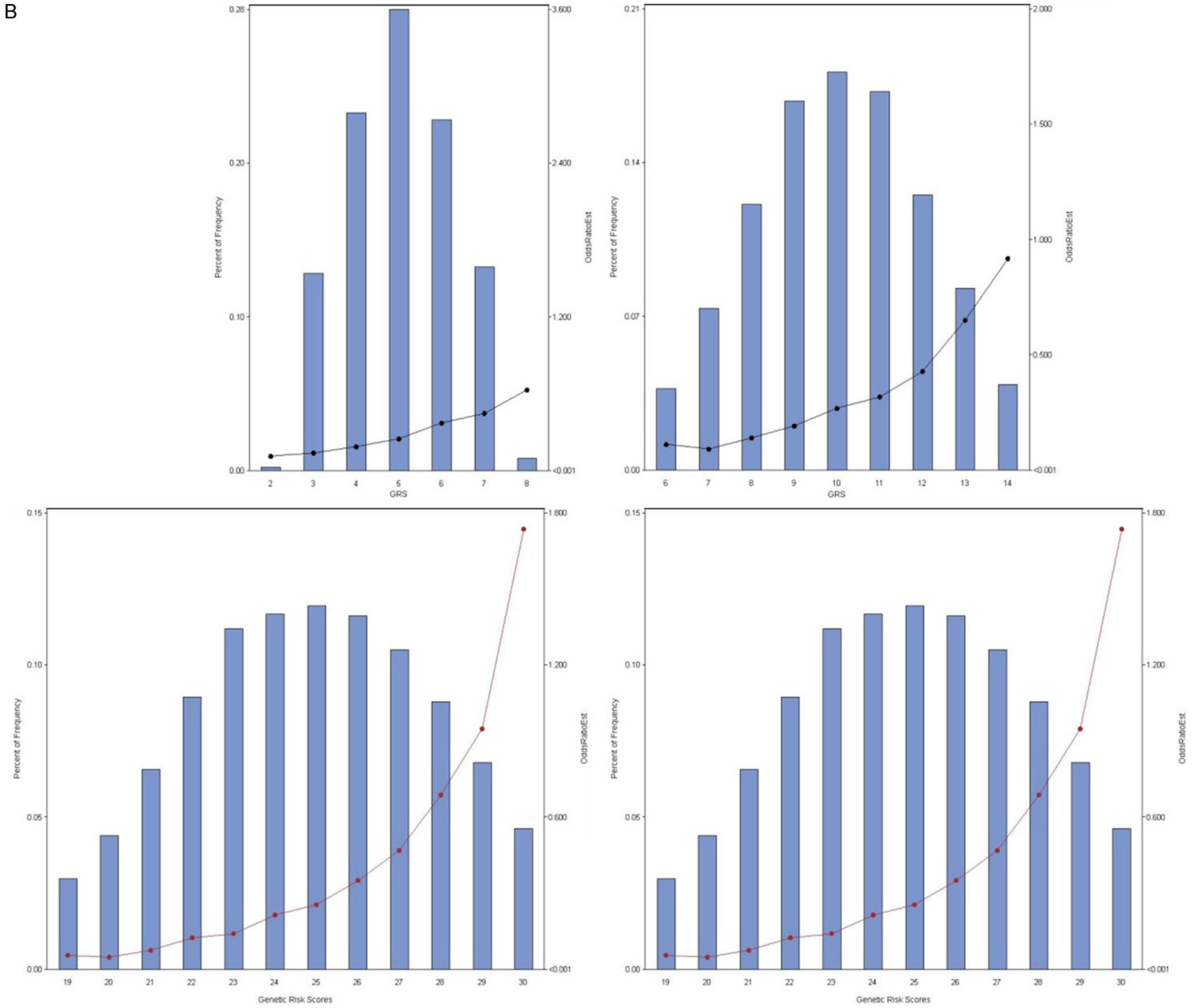

B
0.28
0.20
0.10
0.00
Percent of Frequency
3.600
2.400
1.200
<0.001
OddsRatioEst
2
3
4
5
6
7
8
GRS
0.21
0.14
0.07
0.00
Percent of Frequency
2.000
1.500
1.000
0.500
<0.001
OddsRatioEst
6
7
8
9
10
11
12
13
14
GRS
0.15
0.10
0.05
0.00
Percent of Frequency
1.800
1.200
0.600
<0.001
OddsRatioEst
19
20
21
22
23
24
25
26
27
28
29
30
Genetic Risk Scores
0.15
0.10
0.05
0.00
Percent of Frequency
1.800
1.200
0.600
<0.001
OddsRatioEst
19
20
21
22
23
24
25
26
27
28
29
30
Genetic Risk Scores

Figure 2. Plot of GRS frequencies and disease rates over GRS. (A) for additive risk effects models, (B) for multiplicative risk effects models.

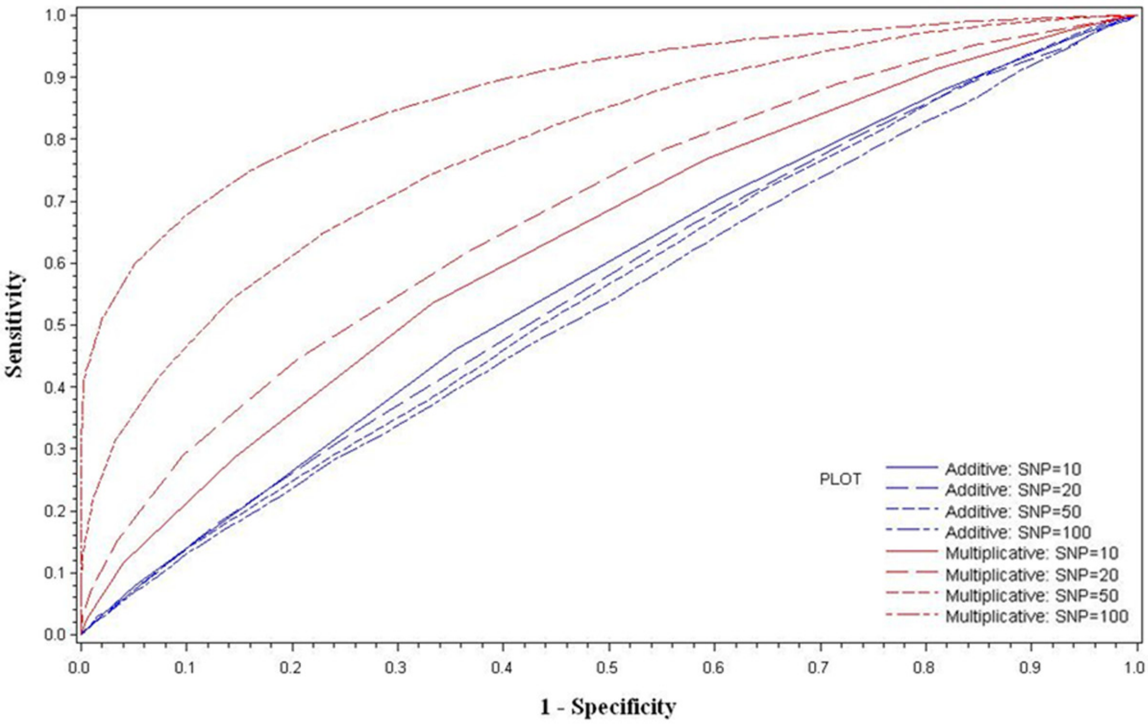


Figure 3. The ROC curves of four different genotype profiles for additive and multiplicative effects models.

Results

Genetic profiles of 10 SNPs, 20 SNPs, 50 SNPs and 100 SNPs with genotype frequency of 50% resulted in a mean number of risk genotype in the simulation data of 5, 10, 25 and 50, respectively. The genetic risk score was created easily by summing the risk alleles from each genetic variant. As seen in **Figure 1**, the observed range of the number of risk genotypes was 0 to 10 when the data has 10 SNPs, 2 to 18 for 20 SNPs, 12 to 38 for 50 SNPs, and 33 to 70 for 100 SNPs. All genetic variants involved in the genetic profiles had the same relative risk and risk genotype frequencies. The x-axis indicates the number of GRS in the genetic profiles, and the y-axis indicates the percentage of subjects in the simulated population. We investigated the disease rates over genetic risk scores. As expected, the disease rates rise as the GRS increase. The increment is somewhat larger in the multiplicative risk models than additive risk models (**Figure 2**). The odds ratios per GRS were investigated for four different genetic profiles and two effects models of additive and multiplicative risks. The odds ratios increased in the multiplicative effects as the size of genetic profiles increased while the odds ratios slightly decreased in the additive effects model. We also examined the relative risks of quintiles of GRS for both risk models. A distinct trend was observed in the relative risks over quintiles of GRS for the multiplicative effects model. That is, with advancing quintiles of GRS, the relative risks increased. For comparison with risk effects model, we found that multiplicative effects models show higher relative risks compared to additive effects models. The marked increase in the relative risks was shown in the multiplicative effects model as the number of SNPs increase. For the multiplicative effects models, fifth quintile, the relative was 2.3 for 10 SNPs and 15.5 for 100 SNPs. However, the relative risks for quintiles slightly decreased as the number of SNPs increased in the additive effects model. The area under ROC curve (AUC) of genetic profiles ranged from 0.6353 for 10 SNPs, 0.6804 for 20 SNPs, 0.7867 for 50

Table 1. A comparison table between additive GRS and multiplicative GRS, including odds ratios of GRS, RRs of quintiles for GRS, and areas under the curves (AUC) for all four different genotype profiles in both additive and multiplicative effects models

GRS type	GRS statistics	Generated data type (Number of SNPs)			
		SNP = 10	SNP = 20	SNP = 50	SNP = 100
Additive	OR per GRS	1.174	1.093	1.054	1.021
	(95% CI)	(1.138, 1.212)	(1.070, 1.117)	(1.040, 1.068)	(1.012, 1.030)
	RR of Q5/Q1	1.5540	1.5842	1.3488	1.2133
	RR of Q4/Q1	1.4748	1.3465	1.2099	1.1136
	RR of Q3/Q1	1.3201	1.3003	1.2037	1.0776
	RR of Q2/Q1	1.1978	1.1584	1.1019	1.0166
	AUC	0.5706	0.5579	0.5502	0.5317
	(95% CI)	(0.557, 0.584)	(0.544, 0.572)	(0.536, 0.564)	(0.518, 0.546)
Multiplicative	OR per GRS	1.381	1.370	1.444	1.481
	(95% CI)	(1.338, 1.426)	(1.338, 1.402)	(1.419, 1.470)	(1.458, 1.505)
	RR of Q5/Q1	2.3004	3.6599	7.4452	15.5278
	RR of Q4/Q1	1.9313	2.3655	4.1164	7.6389
	RR of Q3/Q1	1.5751	2.0965	2.7192	4.2037
	RR of Q2/Q1	1.3262	1.4873	1.7877	2.0370
	AUC	0.6354	0.6804	0.7867	0.8787
	(95% CI)	(0.623, 0.648)	(0.668, 0.693)	(0.777, 0.797)	(0.872, 0.886)

SNPs, and 0.8787 for 100 SNPs in the multiplicative effects models while the AUC went from 0.5706 for 10 SNPs to 0.5317 for 100 SNPs, which indicates a slight decrease as the number of SNPs increases. **Figure 3** shows the ROC curves for four different genotype profiles for both risk effects models. **Table 1** includes the odds ratios of GRS, relative risk associated with GRS, and AUC for each genetic profile. Based on our results, we conclude that: (1) GRS under the additive model is *not* useful for predicting dichotomous outcomes. Adding additional SNPs to the GRS provides very little additional information and therefore is also not likely to be clinically useful, (2) By contrast, GRS under the multiplicative model can be useful for predicting dichotomous outcomes. Adding additional SNPs to the multiplicative GRS *does* provide additional information. Good prediction of dichotomous outcomes can be obtained with as few as 10-20 SNPS, each with small effects (RR < 1.2) under the multiplicative model (consistent with other simulation studies), (3) Available GRS data in published cohorts is much more consistent with an additive model than a multiplicative model. This is similar to the combined effects of other non-genetic risk factors (e.g. the combined effect of elevated LDL cholesterol and hypertension is very close to linear additive effects, instead of multiplicative). It is supported by three lines of evidence: Logistic Model RR per increase in GRS, Area under ROC curve, and RR of increasing quantiles (e.g. quintiles), and (4) Therefore, for GRS to be clinically useful for predicting dichotomous outcomes, alternative non-parametric statistical models are needed to identify a group of SNPS whose combined effect is closer to multiplicative than additive (adding all SNPs may create too much noise and obscure the combined effects of the more important SNPs, i.e. those whose combined effects are multiplicative). This may be a preferred strategy and provide more powerful predictive information than simply including all disease-associated SNPS in a simple count GRS.

Discussion

Epidemiologic studies that have used genetic risk scores for cardiovascular disease have found some evidence of increased prediction [5, 14]. There is debate over whether variants of a relatively small number of genes, each with weak or modest individual effects, account for a large proportion of common diseases in the population, or whether a large number of rare variants with large effects underlie genetic susceptibility to these diseases. It is not clear how many genes are necessary to account for an appreciable population-attributable fraction of these diseases. Yang and his colleague [9] estimated the number of disease susceptibility genes needed to account for varying population attributable fractions of a common complex disease. They concluded that only less than 20 genes are usually needed to explain 50% of the burden of a disease in the population if the predisposing genotypes are common (≥ 25%), even if the individual risk ratios are relatively small (RR = 1.2-1.5). Our results are consistent with Yang et al. (2005).

During the last decades, epidemiological and laboratory studies have supported solid evidence that significant gene-gene as well as gene-environment interactions underlie chronic complex diseases. Thus, in the epidemiologic literature on complex chronic diseases (cancer, diabetes, cardiovascular diseases, etc), both gene-gene and gene-environmental interaction figure prominently. However, although there is a clear linkage between lifestyle and genetic background, a thorough understanding of the underlying mechanisms and how these complex and chronic diseases are triggered and progress is only just beginning to emerge. That is, the complex interplay between genes and environment in chronic diseases is generally not well understood. Behavioral and environmental factors such as cigarette smoking and alcohol consumption would be useful to include in risk prediction. If research aims to reduce the burden of complex disease, research priorities should include the identification and development of novel biomarkers, providing an easy in-vitro diagnostic approach for phenotype classification of the patients. An important approach toward this goal will be the integration of omics data, requiring huge investments in bioinformatics and systems biology [15].

Several empirical studies have considered whether multiple genetic variants will afford better risk prediction to identify individuals that are at higher risk. Most of those studies assumed multiplicative gene effects on risk scale (additive gene action on the log risk scale) and those studies found that risk alleles underlying complex genetic diseases have small marginal effects, with most genotype relative risks in the range of 1.1 to 2.0. Based on empirical studies, the additive effects models offers a better fit in risk prediction modeling using multiple polymorphisms than multiplicative effect models [4-8, 16, 17]. Most epidemiologic studies have used a parametric model of logistic regression models for risk prediction. However, since the logistic models are in nature a log additive (multiplicative) over the risk scale, they do not fit the multigenic studies that might better fit linearly additive effects over the risk. Alternative methods for constructing GRS that attempt to identify and include SNPs that demonstrate multiplicative gene-gene or gene-environment interactive effects are needed. Recursive partitioning methods can be a potential alternative to overcome the limitation of current logistic models to identify risk prediction [18-21].

Acknowledgements

We thank Dr. Brian A. Ference for valuable comments and discussions. This research was partially supported by 5R01CA166785, U54-MD008149, and U54 CA118638.

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Wonsuk Yoo, Institute of Public and Preventive Health, and Department of Dental Medicine, Georgia Regents University, Augusta, GA. E-mail: wyoo@gru.edu

References

[1] Gunjaca G, Boban M, Pehlic M, Zemunik T, Budimir D, Kolcic I, Lauc G, Rudan I and Polasek O. Predictive value of 8 genetic loci for serum uric acid concentration. Croat Med J 2010; 51: 23-31.

[2] Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, Kane JP, Pankow JS, Devlin JJ, Willerson JT and Boerwinkle E. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. Am J Epidemiol 2007; 166: 28-35.

[3] Piccolo SR, Abo RP, Allen-Brady K, Camp NJ, Knight S, Anderson JL and Horne BD. Evaluation of genetic risk scores for lipid levels using genome-wide markers in the Framingham Heart Study. BMC Proc 2009; 3 Suppl 7: S46.

[4] Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, Cai T, Hunter DJ and Hu FB. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med 2009; 150: 541-550.

[5] Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, Hirschhorn JN, Berglund G, Hedblad B, Groop L, Altshuler DM, Newton-Cheh C and Orho-Melander M. Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med 2008; 358: 1240-1249.

[6] Paynter NP, Chasman DI, Pare G, Buring JE, Cook NR, Miletich JP and Ridker PM. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA 2010; 303: 631-637.

[7] Talmud PJ, Hingorani AD, Cooper JA, Marmot MG, Brunner EJ, Kumari M, Kivimaki M and Humphries SE. Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study. BMJ 2010; 340: b4838.

[8] Wray NR, Goddard ME and Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 2007; 17: 1520-1528.

[9] Yang Q, Khoury MJ, Friedman J, Little J and Flanders WD. How many genes underlie the occurrence of common complex diseases in the population? Int J Epidemiol 2005; 34: 1129-1137.

[10] Drenos F, Whittaker JC and Humphries SE. The use of meta-analysis risk estimates for candidate genes in combination to predict coronary heart disease risk. Ann Hum Genet 2007; 71: 611-619.

[11] Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW and van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 2006; 8: 395-400.

[12] van der Net JB, Janssens AC, Sijbrands EJ and Steyerberg EW. Value of genetic profiling for the prediction of coronary heart disease. Am Heart J 2009; 158: 105-110.

[13] Horne BD, Anderson JL, Carlquist JF, Muhlestein JB, Renlund DG, Bair TL, Pearson RR and Camp NJ. Generating genetic risk scores from intermediate phenotypes for use in association studies of clinically significant endpoints. Ann Hum Genet 2005; 69: 176-186.

[14] Ioannidis JP. Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. Circ Cardiovasc Genet 2009; 2: 7-15.

[15] Coughlin SS. “Test, Listen, Cure” (TLC) Hepatitis C Community Awareness Campaign. JMIR Res Protoc 2015; 4: e13.

[16] He M, Cornelis MC, Franks PW, Zhang C, Hu FB and Qi L. Obesity genotype score and cardiovascular risk in women with type 2 diabetes mellitus. Arterioscler Thromb Vasc Biol 2010; 30: 327-332.

[17] Qi L, Cornelis MC, Zhang C, van Dam RM and Hu FB. Genetic predisposition, Western dietary pattern, and the risk of type 2 diabetes in men. Am J Clin Nutr 2009; 89: 1453-1458.

[18] Breiman L, Friedman J, Olshen, R, Stone C. Classification and Regression Trees. Wadsworth, 1984.

[19] Breiman L. Random Forests. Machine Learning 2001; 45.

[20] Ruczinski I KC, LeBlanc M. Logic regression. Journal of Computational and graphical Statistics 2003; 12: 475.

[21] Yoo W, Ference BA, Cote ML and Schwartz A. A Comparison of Logistic Regression, Logic Regression, Classification Tree, and Random Forests to Identify Effective Gene-Gene and Gene-Environmental Interactions. Int J Appl Sci Technol 2012; 2: 268.