European Journal of Human Genetics (2012) 20, 1270–1274

www.nature.com/ejhg

ARTICLE

Analytical and simulation methods for estimating the potential predictive ability of genetic profiling: a comparison of methods and results

Suman Kundu[1], Lennart C Karssen[1] and A Cecile JW Janssens*[,1]

Various modeling methods have been proposed to estimate the potential predictive ability of polygenic risk variants that predispose to various common diseases. However, it is unknown whether differences between them affect their conclusions on predictive ability. We reviewed input parameters, assumptions and output of the five most common methods and compared their estimates of the area under the receiver operating characteristic (ROC) curve (AUC) using hypothetical data representing effect sizes and frequencies of genetic variants, population disease risk and number of variants. To assess the accuracy of the estimated AUCs, we aimed to reproduce the AUCs of published empirical studies. All methods assumed that the combined effect of genetic variants on disease risk followed a multiplicative risk model of independent genetic effects, but they either assumed per allele, per genotype or dominant/recessive effects for the genetic variants. Modeling strategy and input parameters differed. Methods used simulation analysis or analytical formulas with effect sizes quantified by odds ratios (ORs) or relative risks. Estimated AUC values were similar for lower ORs (<1.2). When AUCs were larger (>0.7) due to variants with strong effects, differences in estimated AUCs between methods increased. The simulation methods accurately reproduced the AUC values of empirical studies, but the analytical methods did not. We conclude that despite differences in input parameters, the modeling methods estimate similar AUC for realistic values of the ORs. When one or more variants have stronger effects and AUC values are higher, the simulation methods tend to be more accurate.
European Journal of Human Genetics (2012) **20,** 1270–1274; doi:10.1038/ejhg.2012.89; published online 30 May 2012

Keywords: risk prediction; modeling; discriminative accuracy; AUC; complex disease

INTRODUCTION

The success of genome-wide association studies has fueled interest in genetic risk prediction of multifactorial diseases, such as type 2 diabetes, cardiovascular disease and non-familial cancers. The known contribution of genetic variants to the prediction of most diseases is still limited,[1–2] as the variants identified to date together explain only a small part of the heritability.[3] Further research is needed to find out the extent to which genetic variants can improve the prediction of multifactorial diseases.

To investigate the potential predictive ability of genetic risk models, researchers are using modeling studies to quantify the area under the receiver operating characteristic (ROC) curve (AUC) as a measure of discriminative accuracy.[4–8] These studies have demonstrated that hundreds of genetic variants are required to obtain an AUC of 0.70 when their effect sizes are small (odds ratio (OR) <1.2),[5] and that the upper limit of the AUC is determined by the heritability of the disease and the population disease risk.[5,9] For example, when the heritability of the disease is 10% and the population disease risk is 20%, the maximum AUC value that can be obtained by genetic risk models will be around 0.80.[5]

The modeling methods published to date have similarities and differences in terms of input parameters, underlying assumptions and output produced. For example, all methods assume multiplicative joint effects of genetic variants, but to express the effect sizes of the variants some method use relative risks (RRs), whereas others use ORs as input data. These differences may impact the AUC and lead to different inferences about the predictive ability of genetic risk models, but this impact is not obvious as AUC is known to be an insensitive metric, unable to detect the contribution of significant risk factors.[10] As it is unknown whether these differences between the modeling strategies affect conclusions on the predictive ability, we reviewed published modeling methods that intend to estimate the potential predictive ability of genetic risk models. We compared the input parameters, underlying assumptions and output, and investigated the agreement of estimated AUCs between the methods in several hypothetical scenarios. We also assessed the accuracy of estimated AUCs by attempting to reproduce the AUC values reported in several published empirical studies.

METHODS

Analytical and simulation methods

We compared the five published methods that aim to investigate the predictive ability of genetic risk models by quantifying the AUC.[4–8] The methods are referred in this paper by the name of the first author. Three methods use analytical formulas and two use simulations to obtain the AUC. First, the analytic method by Lu[7] calculates the frequencies and likelihood ratios of all genotype combinations separately for cases and controls from the population disease risk, and the RRs and frequencies of all genetic variants. The AUC values are subsequently obtained from the distribution of likelihood ratios in

[1]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands
*Correspondence: Dr ACJW Janssens, Department of Epidemiology, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands. Tel: +31 10 7044214; Fax: +31 10 7044657; E-mail: a.janssens@erasmusmc.nl
Received 25 January 2012; revised 29 March 2012; accepted 18 April 2012; published online 30 May 2012

cases and controls. Second, the analytic method by Moonesinghe[4] obtains the AUC using a formula that requires RRs and frequencies for dominant or recessive effects of the variants. This method approximates the distributions of the number of risk genotypes for cases and controls by normal distributions that are subsequently used to obtain the AUC value. Third, the analytical method by Gail[8] computes the RRs of all possible genotype combinations for the entire population and for cases, and uses these distributions to obtain the AUC. The simulation methods by Pepe[6] and Janssens[5] both first construct genotype data for individuals of a hypothetical population according to the frequencies of the genetic variants. Based on these data and the ORs, they estimate the disease risk, which is then used to obtain the disease status for each individual in the hypothetical data set. Using the estimated disease risks and disease status, the methods finally calculate the AUC value. These two methods differ in how the genetic effects of the variants are considered. The method by Pepe requires per allele frequencies and ORs to construct individual genotype data, and estimates disease risks using a logistic regression equation, whereas the method by Janssens can use per genotype, per allele or dominant/recessive effect of the risk allele to construct genotype data, and estimates disease risks using Bayes' theorem.

We documented the modeling strategy, input parameters, assumptions and output. To ensure that all these items were assessed for all methods, a checklist of the documented items was made and the five methods were reviewed again. If an item was not explicitly mentioned, deductive reasoning was used to document it. For example, if a method constructed the combined effect of all genetic variants by multiplying the effects of each single variant, we recorded that the method assumed independent genetic effects. Data extraction was done by two researchers (SK, LCK) independently and discrepancies were discussed with a third researcher (ACJWJ).

Table 1 presents an overview of the modeling strategy, input parameters, assumptions and output of the methods. To obtain AUC values, the methods use different input parameters. All methods require effect estimates and frequencies of the genetic variants included, but the effect sizes of genetic variants have to be entered differently. The method by Lu can handle ORs and RRs, whereas the methods by Pepe, Janssens and Gail require ORs and the methods by Moonesinghe requires RRs. All but two methods require an estimate of the population risks, and the simulation models additionally need a specification of the population size.

All methods assume that (i) the combined effect of the genetic variants on disease risk follows a multiplicative (ie, log-additive) risk model; (ii) genetic variants inherit independently, that is, no linkage disequilibrium between the variants; (iii) genetic variants have independent effects on the disease risk, which indicates no interaction among variants. Furthermore, if methods need to convert allele frequencies into genotype frequencies, they additionally assume that all genotypes and allele frequencies are in Hardy–Weinberg Equilibrium. Two methods assume that the disease is rare. Finally, the methods differ in how genetic variants need to be included. Two methods assume per allele (additive) effects of the risk allele, one assumes that the effects vary between genotypes and one assumes dominant or recessive effects of the risk alleles. The fifth method does not make any assumptions about the genetic effects and allows these to vary between the variants considered.

We had selected methods that obtain the AUC as a measure of predictive ability, but most methods can obtain other predictive measures of (genetic) risk models as well. Moonesinghe's method provides a formula to specifically calculate the AUC, but all other methods can be used to obtain other plots and metrics as well, such as risk distributions and predictiveness curves. The simulation methods can be used to compare risk models by, for example, reclassification measures.

Data analysis and data generation

To investigate the agreement in estimated AUCs, we applied the five methods in various hypothetical scenarios. Scenarios were defined as any combination of (i) the number of genetic variants included, chosen to be 10 or 50; (ii) the OR of the risk allele, set to 1.1, 1.4 or 2.0; (iii) the risk allele frequency, set to 0.05 or 0.25; and (iv) the disease risk in the population, set to 1 or 25%, as listed in Table 2. In these hypothetical scenarios, we assumed that all genetic variants had the same risk allele frequencies and ORs.

Table 1 Overview of input parameters, assumptions and output of the modeling methods

First author[Ref]	*Pepe*[6]	*Janssens*[5]	*Lu*[7]	*Moonesinghe*[4]	*Gail*[8]
Modeling strategy	Simulation	Simulation	Analytic formula	Analytic formula	Analytic formula
Input parameters					
Population risk	Yes	Yes	Yes	No	No[a]
Variant effect estimate (RR/ OR)	Yes (OR)	Yes (OR)	Yes (RR/OR)	Yes (RR)	Yes (OR)
Variant frequency	Yes	Yes	Yes	Yes	Yes
Population size	Yes	Yes	No	No	No
Assumptions					
Multiplicative risk model	Yes	Yes	Yes	Yes	Yes
No linkage disequilibrium	Yes	Yes	Yes	Yes	Yes
Independent effects	Yes	Yes	Yes	Yes	Yes
HWE[b]	Yes	Yes[c]	No	No	Yes
Rare disease	No	No	No	Yes	Yes
Genetic effects	Per allele	All	Per genotype	Dominant or recessive	Per allele
Output parameters					
Create data set	Yes	Yes	No	No	No
AUC	Yes	Yes	Yes	Yes	Yes
Other output possible[d]	Risk distribution curves	Risk distribution curves	Risk distribution curves	No	Risk distribution curves
	Predictiveness curve	Predictiveness curve	Predictiveness curve		Predictiveness curve
	Reclassification measures	Reclassification measures			

Abbreviations: AUC, area under the receiver operating characteristic curve; HWE, Hardy–Weinberg equilibrium; OR, odds ratio; RR, relative risk.
[a]The formula includes a constant *k*, indicating the absolute risk of disease or incidence for individuals who have the lowest RR, which can be obtained from the population risk. Yet, the exact value of *k* is not required to obtain AUC as AUC is independent of disease risk.
[b]HWE is assumed for the distribution of genotypes in the total population.
[c]HWE is only assumed when genetic effects are entered as per allele effects.
[d]Output that can be obtained using the method, but which was not necessarily proposed in the original paper.

Table 2 Estimated area under the ROC curve for hypothetical values of the input parameters

				Estimated AUC					
				Simulation		*Analytical*			
Odds ratio	*Risk allele frequency (%)*	*Population disease risk (%)*	*Number of genes*	*Pepe*	*Janssens*	*Lu*	*Moonesinghe*		*Gail*
							Dominant	*Recessive*	
1.1	5	1	10	0.53	0.53	0.52	0.53	0.51	0.52
			50	0.56	0.56	—[a]	0.56	0.52	—[a]
		25	10	0.53	0.53	0.52	0.52	0.51	0.52
			50	0.56	0.56	—[a]	0.54	0.51	—[a]
	25	1	10	0.55	0.55	0.55	0.55	0.53	0.55
			50	0.62	0.61	—[a]	0.61	0.57	—[a]
		25	10	0.55	0.55	0.55	0.53	0.52	0.55
			50	0.61	0.61	—[a]	0.58	0.56	—[a]
1.4	5	1	10	0.60	0.60	0.59	0.60	0.53	0.59
			50	0.71	0.71	—[a]	0.71	0.57	—[a]
		25	10	0.59	0.59	0.59	0.57	0.52	0.59
			50	0.69	0.69	—[a]	0.66	0.55	—[a]
	25	1	10	0.68	0.68	0.68	0.67	0.62	0.68
			50	0.85	0.85	—[a]	0.83	0.76	—[a]
		25	10	0.67	0.67	0.67	0.63	0.59	0.68
			50	0.81	0.81	—[a]	0.77	0.70	—[a]
2.0	5	1	10	0.70	0.70	0.70	0.70	0.58	0.70
			50	0.87	0.88	—[a]	0.89	0.67	—[a]
		25	10	0.68	0.68	0.66	0.65	0.55	0.70
			50	0.82	0.82	—[a]	0.81	0.62	—[a]
	25	1	10	0.84	0.84	0.84	0.82	0.75	0.84
			50	0.97	0.97	—[a]	0.98	0.94	—[a]
		25	10	0.80	0.80	0.60	0.76	0.69	0.84
			50	0.93	0.93	—[a]	0.94	0.87	—[a]

Abbreviations: AUC, area under receiver operating characteristic curve; ROC, receiver operating characteristic.
[a]AUC cannot be obtained because of insufficient computer memory.

To assess the accuracy of the estimated AUCs, we investigated whether the methods could accurately reproduce AUCs of published empirical studies (Table 3). We selected studies that assessed the AUC of genetic risk models and reported the ORs and frequencies of genetic variants included in the model. Population disease risks were taken from the empirical studies listed in Table 3 or from other epidemiological studies if they were not listed in the original paper. As random factors, such as rounding of values and random deviations from Hardy–Weinberg Equilibrium, may have impacted the empirical AUC, we conclude that the methods accurately reproduce the empirical studies when the predicted AUC is similar to the empirical AUC, but not necessarily exactly the same.

As the methods differ in how genetic variants need to be entered in the method, as per allele, per genotype or dominant/recessive effects of the risk alleles, transformations were needed when the specified and required frequencies and risk estimates did not match. Specified values of the frequencies, ORs and population risk were used to construct a (3×2) genotype by disease status contingency table, from which all required frequencies and risk ratios (OR/RR) were calculated. Hardy–Weinberg Equilibrium was assumed to obtain genotype frequencies when allele frequencies were specified.

For the simulation methods, genetic variants and disease status were constructed for 100 000 individuals and all simulations were repeated 100 times to obtain robust estimates of the AUC. Presented AUC estimates are averages of the 100 runs. All analyses were performed using software written in the R language (version 2.12.1).[11] Extensive details together with the mathematical explanation of the five methods and the source codes or references to the source codes are provided in the Supplemental Material.

RESULTS

Table 2 shows the estimated AUC values obtained by the five methods for the hypothetical scenarios. As expected, higher risk allele frequencies, higher ORs and larger number of genetic variants yielded higher AUC values for all methods. The differences in AUC between the methods were larger when the AUC values were higher; for example, when higher ORs or more genetic variants were considered. The AUC values calculated using the simulation methods were identical up to two decimals in most scenarios. The analytical method of Moonesinghe consistently produced lower AUC values than the simulation methods, particularly when recessive effects of the variants were assumed. The same results were observed when the risk allele frequency was 75%, with the exception that the recessive model estimated higher AUC values than the dominant model (data not shown). The analytical method of Lu yielded lower AUC estimates than the simulation methods when AUCs were higher (>0.80). The analytical method of Gail obtained similar AUC values as the simulation methods when the disease risk was 5%, but overestimated the AUC when the disease risk was 25%. Both the methods by Lu and Gail were unable to compute the AUC when the number of genetic variants was 50.

Table 3 presents the estimated AUCs for the scenarios that used the frequencies and ORs of genetic variants and population risks obtained from published empirical studies. The estimated AUCs using the simulation methods and the analytic methods of Gail and Lu were

Table 3 Estimated area under the ROC curve for input parameters from published empirical studies

						Estimated AUC					
						Simulation		Analytical			
Genetic effects	*First author*[Ref]	*Disease outcome*	*Population disease risk (%)*[Ref]	*Number of genetic variants*	*Published AUC*	*Pepe per allele*	*Janssens All*	*Lu per genotype*	*Moonesinghe Dominant*	*Moonesinghe Recessive*	*Gail per allele*
Per genotype	Van Hoek[14]	Type 2 diabetes	20[14]	18	0.60	0.59	0.59	—[a]	0.56	0.57	—[a]
Per allele	Mealiffe[15]	Breast cancer	12.15[16]	7	0.59	0.59	0.59	0.59	0.57	0.56	0.59
	Helfand[17]	Prostate cancer	16.22[16]	9	0.66	0.68	0.68	0.68	0.63	0.62	0.68
	Hu[18]	Type 2 diabetes	20[14]	11	0.62	0.63	0.63	0.63	0.58	0.59	0.63
	Lin[19]	Type 2 diabetes	20[14]	15	0.59	0.61	0.61	—[a]	0.58	0.57	—[a]
	Takahashi[20]	Osteoarthritis (knee)	27.8[12]	3	0.55	0.56	0.56	0.56	0.53	0.53	0.56
	Qi[22]	Type 2 diabetes	20[14]	17	0.62	0.64	0.64	—[a]	0.60	0.58	—[a]
Combination of per genotype and dominant/ recessive	Seddon[23]	Age-related macular degeneration	6.5[24]	14	0.82	—[b]	0.81	0.80	0.77	0.72	—[b]

Abbreviations: AUC, area under receiver operating characteristic curve; ROC, receiver operating characteristic.
[a]AUC cannot be obtained because of insufficient computer memory.
[b]Per allele odds ratios and frequencies cannot be reconstructed from the dominant/recessive data.

always consistent with those of the empirical studies, but the analytical method of Moonesinghe underestimated all empirical AUCs. When the number of genetic variants was 15 or higher, the analytical methods by Gail and Lu were unable to compute the AUC because of computer memory limitations.

DISCUSSION

This paper provides a review of the five methods that have been proposed to investigate the potential predictive ability of genetic risk models by quantifying the AUC that can be expected. The five modeling methods use the same main assumptions, but they differed with regard to the modeling strategy. Estimates of the AUC differed between the methods when one or more variants had stronger effects and absolute AUC values were higher. The two simulation methods always obtained the same estimates and both accurately reproduced the AUCs of published empirical studies.

Modeling studies are used to estimate the potential predictive ability of genetic risk models on the basis of hypothetical epidemiological data. When the modeling is based on published ORs and frequencies rather than on hypothetical values for variants, some methods may be more flexible than others. If the coding of genetic variants differs between what is assumed in the method and what is published in the literature, transformations are needed. These transformations, such as converting the data into dominant/recessive effects of the risk alleles, may not be valid in reality and in our examples, and may explain the differences in estimated and published AUC values when transformations were applied, for example, for the method by Moonesinghe (Table 3). These transformations may have contributed to the differences in AUC values between the methods.

Although the methods share similar assumptions, they differ in the way the AUC is obtained. Some details in the calculation can be considered as limitations of the methods. For example, the analytic methods by Gail and Lu are not able to obtain the AUC for larger number of genetic variants because they calculate the frequencies of all possible combinations of the genotypes. As the number of combinations grows exponentially with increasing number of variants, at one point these methods reach the limits of computer memory. Using a computer with a 2.33-GHz processor and 2-GB RAM, we observed that the AUC could not be computed when the number of variants exceeded 14. When the number of genetic variants is larger, Gail's method can still be used by assuming a log-normal distribution of RRs for the genotype combinations in the population. Another example is that most methods assume the variants to have either per allele, per genotype or dominant/recessive effects, rather than allowing the effects to differ between them. Most empirical risk prediction studies these days consider weighted risk allele counts (weighted risk scores) when the number of variants is large, which is similar to the assumption of per allele effects. Assuming per genotype effects is more flexible, as it simultaneously expresses the per allele effects or dominant/recessive effects of the variants. Yet, solely assuming dominant/recessive effects of risk variants may not adequately express allelic effects and hence explain why the method of Moonesinghe underestimated the AUC values when assuming recessive effects. Even though AUC is known as an insensitive metric,[10] these differences in assumptions about the genetic effects had substantial impact on the observed AUC value.

We reviewed five methods that estimate the AUC of prediction models. There are two other modeling approaches for the predictive ability of genetic risk models that we did not evaluate because they do not estimate AUC based on published epidemiological data of genetic variants, that is, on ORs and frequencies. First, in a theoretical paper on the predictive ability of multiple genetic variants, Pharoah *et al*[12] described how genetic profiling yields a distribution of risk that can be useful for selecting high-risk groups in disease prevention. Second, Wray *et al*[13] described three different models for genetic risk prediction that assume different underlying distribution of the disease risk in the population. The methods by Pharoah *et al* and Wray *et al* use the same assumptions as the five discussed methods, including a multiplicative risk model for joint effects and independent effects of genetic variants.

All methods are methodologically simple and use assumptions that are generally reasonable. They assume that the combined effect of the genetic variants on disease risk follows a multiplicative risk model with independent effects (ie, no statistical interaction terms are included in the model) and that genetic variants inherit independently. Inclusion of gene–gene and gene–environment interactions may further improve the predictive ability of the methods. Although all five methods might be improved by including these extensions, their performance so far seems adequate given current understanding of the joint contribution of genetic variants to the disease risk. Currently many empirical studies calculate weighted risk scores where the differences in the effects between risk alleles are acknowledged. Of the modeling studies, some explicitly obtain weighted risks scores,[5,6,8] whereas others consider different effect sizes for risk alleles in other ways.[4,7]

In conclusion, the five most commonly used methods for quantifying the AUC of genetic risk prediction models have similar assumptions, but differ with regard to the input parameters required and the AUC values estimated. The simulation methods yielded consistent AUC estimates and both accurately replicated published empirical AUC values. The simulation studies provide valuable insight into the potential predictive ability of genetic risk prediction.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Vidi grant from the Netherlands Organization for Scientific Research (NWO), the Young Investigator grant from the Erasmus University Medical Center Rotterdam and by the Center for Medical Systems Biology within the framework of the Netherlands Genomics Initiative.

1 Janssens AC, van Duijn CM: Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet* 2008; **17**: R166–R173.
2 Hirschhorn JN, Gajdos ZK: Genome-wide association studies: results from the first few years and potential implications for clinical medicine. *Annu Rev Med* 2011; **62**: 11–24.
3 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
4 Moonesinghe R, Liu T, Khoury MJ: Evaluation of the discriminative accuracy of genomic profiling in the prediction of common complex diseases. *Eur J Hum Genet* 2010; **18**: 485–489.
5 Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM: Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med* 2006; **8**: 395–400.
6 Pepe MS, Gu JW, Morris DE: The potential of genes and other markers to inform about risk. *Cancer Epidemiol Biomarkers Prev* 2010; **19**: 655–665.
7 Lu Q, Elston RC: Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. *Am J Hum Genet* 2008; **82**: 641–651.
8 Gail MH: Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst* 2008; **100**: 1037–1041.
9 Wray NR, Yang J, Goddard ME, Visscher PM: The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* 2010; **6**: e1000864.
10 Cook NR: Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007; **115**: 928–935.
11 R Foundation for Statistical Computing; R Development Core Team: *R: A Language and Environment for Statistical Computing*, Vienna, Austria. http://www.R-project.org, 2011.
12 Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA: Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* 2002; **31**: 33–36.
13 Wray NR, Goddard ME: Multi-locus models of genetic risk of disease. *Genome Med* 2010; **2**: 10.
14 van Hoek M, Dehghan A, Witteman JC *et al*: Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes* 2008; **57**: 3122–3128.
15 Mealiffe ME, Stokowski RP, Rhees BK, Prentice RL, Pettinger M, Hinds DA: Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J Natl Cancer Inst* 2010; **102**: 1618–1627.
16 Howlader N, Noone AM, Krapcho M *et al*: *SEER Cancer Statistics Review, 1975–2008*. National Cancer Institute: Bethesda, MD. http://seer.cancer.gov/csr/1975_2008/ 2011.
17 Helfand BT, Fought AJ, Loeb S, Meeks JJ, Kan D, Catalona WJ: Genetic prostate cancer risk assessment: common variants in 9 genomic regions are associated with cumulative risk. *J Urol* 2010; **184**: 501–505.
18 Hu C, Zhang R, Wang C *et al*: PPARG, KCNJ11, CDKAL1, CDKN2A-CDKN2B, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 are associated with type 2 diabetes in a Chinese population. *PLoS One* 2009; **4**: e7643.
19 Lin X, Song K, Lim N *et al*: Risk prediction of prevalent diabetes in a Swiss population using a weighted genetic score–the CoLaus Study. *Diabetologia* 2009; **52**: 600–608.
20 Takahashi H, Nakajima M, Ozaki K, Tanaka T, Kamatani N, Ikegawa S: Prediction model for knee osteoarthritis based on genetic and clinical information. *Arthritis Res Ther* 2010; **12**: R187.
21 Jordan JM, Helmick CG, Renner JB *et al*: Prevalence of knee symptoms and radiographic and symptomatic knee osteoarthritis in African Americans and Caucasians: the Johnston County Osteoarthritis Project. *J Rheumatol* 2007; **34**: 172–180.
22 Qi Q, Li H, Wu Y *et al*: Combined effects of 17 common genetic variants on type 2 diabetes risk in a Han Chinese population. *Diabetologia* 2010; **53**: 2163–2166.
23 Seddon JM, Reynolds R, Maller J, Fagerness JA, Daly MJ, Rosner B: Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables. *Invest Ophthalmol Vis Sci* 2009; **50**: 2044–2053.
24 Klein R, Chou CF, Klein BE, Zhang X, Meuer SM, Saaddine JB: Prevalence of age-related macular degeneration in the US population. *Arch Ophthalmol* 2011; **129**: 75–80.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)