

Review

Genome-based prediction of common diseases: methodological considerations for future research

A Cecile JW Janssens* and Cornelia M van Duijn*

Address: *Department of Epidemiology, Erasmus University Medical Center Rotterdam, 3000 CA Rotterdam, The Netherlands.

Correspondence: A Cecile JW Janssens. Email: a.janssens@erasmusmc.nl.

Published: 18 February 2009

Genome Medicine 2009, **1**:N (doi:10.1186/gm20)

The electronic version of this article is the complete one and can be found online at <http://genomemedicine.com/content/1/2/20>

© 2009 BioMed Central Ltd

Abstract

The translation of emerging genomic knowledge into public health and clinical care is one of the major challenges for the coming decades. At the moment, genome-based prediction of common diseases, such as type 2 diabetes, coronary heart disease and cancer, is still not informative. Our understanding of the genetic basis of multifactorial diseases is improving, but the currently identified susceptibility variants contribute only marginally to the development of disease. At the same time, an increasing number of companies are offering personalized lifestyle and health recommendations on the basis of individual genetic profiles. This discrepancy between the limited predictive value and the commercial availability of genetic profiles highlights the need for a critical appraisal of the usefulness of genome-based applications in clinical and public health care. Anticipating the discovery of a large number of genetic variants in the near future, we need to prepare a framework for the design and analysis of studies aiming to evaluate the clinical validity and utility of genetic tests. In this article, we review recent studies on the predictive value of genetic profiling from a methodological perspective and address issues around the choice of the study population, the construction of genetic profiles, the measurement of the predictive value, calibration and validation of prediction models, and assessment of clinical utility. Careful consideration of these issues will contribute to the knowledge base that is needed to identify useful genome-based applications for implementation in clinical and public health practice.

Introduction

The past decade has seen rapid developments in our understanding of the genetic etiology of various common multifactorial diseases, such as age-related macular degeneration (AMD), type 1 and type 2 diabetes, cardiovascular diseases, Crohn's disease and various cancers [1]. Further developments in genomic research, such as the growing number of genome-wide association studies, the large-scale consortia that are pooling data from various studies, and the advances in statistical genomics and genotype technology, are drastically improving the chances of identifying common low risk variants and rare high risk variants. It is beyond doubt that

many more genetic susceptibility variants will be discovered in the next few years.

Expectations are high that increasing knowledge of the genetic bases of disease will eventually lead to personalized medicine, that is, to preventive and therapeutic interventions for complex diseases that are tailored to individuals on the basis of their genetic profiles [2,3]. Genome-based personalized medicine already exists for monogenic disorders. For example, female carriers of *BRCA1* or *BRCA2* mutations are offered biannual mammography screening or provided the opportunity of preventive surgery. Potential

Table 1**AUC and effect estimates of susceptibility variants for the prediction of three diseases**

Disease	Coronary heart disease		Systemic lupus erythematosus		Hypertriglyceridemia	
AUC*	0.55		0.67		0.80	
Reference	[35]		[28]		[12]	
Genes and effect estimates						
	<i>AGT</i>	1.28 (1.02, 1.61)	<i>HLA</i>	2.36 (2.11, 2.64) [‡]	<i>APOA5</i>	7.36 (3.98, 13.6)
	<i>ACE</i>	1.18 (0.97, 1.44)	<i>ITGAM</i>	1.62 (1.47, 1.78) [‡]	<i>APOA5</i>	5.57 (3.13, 9.90)
	<i>AGTRI</i>	1.21 (1.00, 1.45)	<i>IRF5/TNPO3</i>	1.54 (1.40, 1.70) [‡]	<i>TBL2</i>	2.81 (1.46, 5.24)
	<i>CYP11B2</i>	1.22 (1.01, 1.48)	<i>KIAA1542</i>	0.78 (0.73, 0.85)	<i>APOE</i>	2.14 (1.31, 3.49)
	<i>ADD1</i>	1.22 (1.01, 1.47)	<i>PXK</i>	1.25 (1.16, 1.35)	<i>GCKR</i>	2.11 (1.21, 3.67)
	<i>GNB3</i>	0.72 (0.52, 1.01)	rs10798269	0.82 (0.76, 0.88)	<i>GALNT2</i>	2.10 (1.15, 3.81)
					<i>TRIB1</i>	2.02 (1.24, 3.30)

*AUC, area under the receiver operating characteristic curve. Values are hazard ratios [35] or odds ratios with 95% confidence intervals. [‡]The original paper mentions several polymorphisms per gene and that one for each gene was included to assess the combined predictive value of six variants. The polymorphisms that had the highest odds ratios are reported here.

applications of genetic profiling in multifactorial diseases include tailoring of prevention programs to at-risk individuals, determining the starting age of participation in screening programs [4] and, when profiles predict treatment success, tailoring treatment modalities and starting doses.

As we have reviewed recently [5], the predictive value of genetic profiling is still limited at present, with a few promising exceptions. The area under the receiver operating characteristic curve (AUC) gives an assessment of the discriminative accuracy of a prediction model, that is, the degree to which the test results can discriminate between persons who will develop the disease and those who will not. AUC ranges from 0.50 (equal to tossing a coin) to 1.00 (perfect prediction). We found that the AUC was low for the genetic prediction of type 2 diabetes and coronary heart disease and high for the prediction of hypertriglyceridemia and AMD [5]. Table 1 illustrates that the high AUC of 0.80 for hypertriglyceridemia resulted from very strong individual genetic factors, with odds ratios ranging from 2.0 to 7.4, and the low AUC of 0.55 for coronary heart disease from genetic variants with low odds ratios. Note that the strongest genetic predictor by far for coronary heart disease had a weaker effect than the weakest predictor for hypertriglyceridemia. In order to achieve appreciable predictive value, genetic profiles need to include a few strong genetic risk factors or a large number of weak susceptibility variants [6].

Although the predictive value of genetic profiling is still limited, an increasing number of companies already offer personalized lifestyle health recommendations and nutritional supplements on the basis of clients' genetic profiles [7]. Despite the limited predictive value of genetic testing in multifactorial diseases, these commercial developments will

yield ongoing interest from consumers, from health care professionals confronted with questions from patients who underwent testing, and from policy makers who search for novel strategies to improve health care and population health. These developments ask for a solid evidence base for genomics applications. One of the major challenges for the coming decades will be to investigate the translation of this emerging genomic knowledge into public health and medical care [8,9].

In this article, we review recent studies on the predictive value of genetic profiling from a methodological perspective. We address five issues: the choice of the study population, the construction of genetic profiles, the measurement of the predictive value, calibration and validation of the predictive value, and finally assessment of the clinical utility of genetic profiles. These issues are illustrated using examples from recent studies on the predictive value of genetic profiling in common diseases. Methodological characteristics of these studies are listed in Table 2.

From gene discovery samples to the target populations

In the gene discovery phase, researchers often make use of highly selected series of patients and controls. Patients are selected for severe pathology, early onset and familial clustering of disease, and controls for the absence of pathology. This procedure substantially improves the statistical power of gene discovery research without creating any bias. But hyperselection of cases and controls can be a problem for evaluating the usefulness of genetic testing, as it typically leads to an overestimation of the effect sizes and, thus, to an overestimation of the predictive value. Effect sizes are inflated because frequencies of the risk genotypes are particularly increased in enriched patient populations and

Table 2

Methodological characteristics of recent studies on the prediction of complex diseases using multiple genes*

First author (year) and reference	Design	Cases	Controls†	n (cases/controls)‡	Variant selection§	Analyses	Evaluation	Calibration	Validation	Compared with clinical prediction
Cauchi (2008) [10]	Case-control	T2D	Normal glucose-tolerant individuals (screened)	4,232/4,595	From GWAS in same population	LR	Distribution RASs, OR for RASs, AUC	No	No	No
Harley (2008) [28]	Case-control	Women with SLE	Age-matched women without SLE	720/2,337	From GWAS in same population	LR	AUC	No	No	No
Humphries (2007) [15]	Prospective cohort	Coronary heart disease	Caucasian men	183/1,874	4 (out of 12) candidate genes	Cox PH, weighted risk score, risk score	Kaplan-Meier curves, AUC	No	No	Yes
Kathiresan (2008) [18]	Prospective cohort	Myocardial infarction, ischemic stroke and death from coronary heart disease	General population	2,383/3,994	9 (out of 11) candidate SNPs in 9 genes	Cox PH, RAS	Distribution RASs, incidence rates for RASs, Kaplan-Meier curves, AUC, reclassification	No	No	Only added value¶
Lango (2008) [11]	Case-control	T2D	Normoglycemic (screened)	2,309/2,598	18 established variants	LR, RAS	Distribution RASs, OR for RASs, AUC	No	No	Yes
Lyssenko (2005) [30]	Prospective cohort	T2D	Relatives and spouses	132/2,161	3 (out of 6) SNPs in 5 genes	Cox PH	HR for genotype combinations	No	No	No
Lyssenko (2008) [21]	Prospective cohort	T2D	Two cohorts: general population and non-diabetic relatives	2,201/16,630	11 (out of 16) established variants	LR, RAS	Distribution RASs, incidence rates for RASs, AUC, reclassification	No	No	Yes
Maller (2006) [13]	Case-control	Advanced AMD	Individuals without AMD or early AMD	1,238/934	5 (out of 1,536 tag SNPs) in candidate genes	LR	Relative risk for genotype combinations*	No	No	No
Meigs (2008) [22]	Prospective cohort	T2D	Offspring of general population cohort	255/2,122	18 established variants	LR, RAS	Distribution RASs, incidence rates for RASs AUC, reclassification	Yes	No	Yes
Morrison (2007) [16]	Prospective cohort	Coronary heart disease	General population	1,452/12,455	11 (out of 116) SNPs	Cox PH, RAS	Distribution RASs, AUC	No	Internal	Only added value
Podgoreanu (2006) [26]	Prospective cohort	Myocardial infarction	Patients undergoing elective cardiac surgery with cardio-pulmonary bypass	52/382	3 (out of 48) SNPs in 23 candidate genes	LR	AUC	No	No	Yes

Continued on the next page

Table 2 (continued)

Methodological characteristics of recent studies on the prediction of complex diseases using multiple genes*

First author (year) and reference	Design	Cases	Controls [†]	n (cases/controls) [‡]	Variant selection [§]	Analyses	Evaluation	Calibration	Validation	Compared with clinical prediction
Van der Net (2009) [17]	Prospective cohort	Coronary heart disease	FH patients	387/950	14 SNPs previously associated	Cox PH, RAS	Predicted risk for RASs, AUC	No	No	Yes
Van Hoek (2008) [14]	Prospective cohort	T2D	General population	1,287/5,221	18 established variants	Cox PH, LR, RAS	Predicted risks for RASs, OR of RASs, AUC	No	No	Yes
Vaxillaire (2008) [27]	Prospective cohort	T2D	General population	307/3,570	3 (out of 19) SNPs in 14 candidate genes	LR	OR of RAS, AUC	No	No	Yes
Wang (2008) [12]	Case-control	Severe hypertriglyceridemia	Normolipidemic controls	132/351	7 established variants	LR	AUC	Hosmer-Lemeshow goodness of fit	No	Only added value
Weedon (2006) [19]	Case-control	T2D	Individuals without T2D, including one normoglycemic subpopulation	2,409/3,668	3 established variants	LR, RAS	Distribution RASs, OR of RASs, AUC	No	No	No
Weersma (2008) [23]	Case-control	Chronic inflammatory bowel disease	Healthy controls	2,804/1,350	5 established genes	LR, RAS	OR of RASs	No	No	No
Yeh (2007) [25]	Case-control	Colorectal cancer	Healthy controls	727/736	3 (out of 10) established variants	LR	OR for genotype combinations [*]	No	No	No
Zheng (2008) [20]	Case-control	Prostate cancer	General population	2,893/1,781	5 (out of 16) SNPs in 5 candidate regions	LR, genotype score	Distribution genotype score, AUC, PAR	No	No	No

Abbreviations: AMD, age-related macular degeneration; AUC, area under the receiver operating characteristic curve, sometimes measured by the c-statistic; Cox PH, Cox proportional hazard regression analysis; FH, familial hypercholesterolemia; GWAS, genome-wide association studies; HR, hazard ratio; LR, logistic regression; OR, odds ratio; PAR, population attributable risk; RAS, risk allele score; SLE, systemic lupus erythematosus; SNP, single nucleotide polymorphism; T2D, type 2 diabetes. [†]For prospective cohort studies, this column describes the total population from which the cases were obtained. [‡]For prospective cohort studies, these numbers indicate the number of cases divided by the number of individuals who did not develop the disease during follow-up. [§]Numbers between parentheses indicate the total number of variants at the start of the analysis from which the most predictive variants were selected. Candidate means that the variants were selected from the literature, on the basis of association with disease risk in other studies. [¶]This study compared prediction from clinical risk factors with clinical risk factors plus genetic factors, but did not consider prediction from genetic factors alone. ^{}These studies did not intend to evaluate the predictive value, but investigated the combined effect of multiple variants on disease risk.

Table 3

Effect estimates of 18 established susceptibility variants on type 2 diabetes risk in two studies

Gene	Locus	Odds ratio (95% confidence interval)*	
		GoDARTS study [11]	Rotterdam study [14]
<i>TCF7L2</i>	rs7903146	1.36 (1.24, 1.48)	1.31 (1.19, 1.44)
<i>KCNJ11</i>	rs5219	1.25 (1.15, 1.36)	1.03 (0.93, 1.13)
<i>CDKN2A/2B</i>	rs10811661	1.21 (1.08, 1.35)	1.10 (0.98, 1.24)
<i>PPARG</i>	rs1801282	1.21 (1.07, 1.36)	1.09 (0.95, 1.24)
<i>ADAM30/NOTCH2</i>	rs2641348†	1.15 (1.01, 1.30)	1.01 (0.88, 1.17)
<i>CDKN2A/2B</i>	rs564398‡	1.13 (1.04, 1.22)	1.04 (0.95, 1.14)
<i>IGF2BP2</i>	rs4402960	1.12 (1.03, 1.22)	1.11 (1.01, 1.22)
<i>FTO</i>	rs8050136	1.11 (1.02, 1.20)	1.09 (0.99, 1.19)
<i>CDKAL1</i>	rs10946398§	1.11 (1.02, 1.21)	1.11 (1.02, 1.22)
<i>SLC30A8</i>	rs13266634	1.10 (1.01, 1.20)	1.13 (1.02, 1.24)
<i>TSPAN8/LGR5</i>	rs7961581¶	1.09 (1.00, 1.19)	1.09 (0.99, 1.20)
<i>CDC123</i>	rs12779790*	1.10 (0.99, 1.21)	0.95 (0.84, 1.06)
<i>WFS1</i>	rs10010131**	1.07 (0.99, 1.16)	1.12 (1.05, 1.27)
<i>TCF2</i>	rs757210††	1.07 (0.99, 1.16)	0.93 (0.85, 1.02)
<i>ADAMTS9</i>	rs4607103‡‡	1.05 (0.96, 1.16)	1.14 (1.03, 1.28)
<i>HHEX-IDE</i>	rs1111875	1.02 (0.94, 1.11)	1.06 (0.97, 1.15)
<i>THADA</i>	rs7578597	1.04 (0.90, 1.19)	1.10 (0.96, 1.27)
<i>JAZF1</i>	rs864745§§	1.00 (0.93, 1.09)	1.09 (1.00, 1.19)

*AUC (area under the receiver operating characteristic curve) was 0.60 for both studies. Values were obtained using logistic regression analyses. For several genes, the Rotterdam study [14] uses different single nucleotide polymorphisms from those listed: ‡‡rs4411878, r2 = 0.95; ¶rs11257622; r2 = 0.83; †rs1412829, r2 = 0.97; §§rs1635852, r2 = 0.97; †rs1493694; ¶rs1353362, r2 = 0.96; ††rs4430796, r2 = 0.61; **rs10012946, r2 = 1.00; §rs7754840, r2 = 1.00. Statistically significant associations are presented in bold and r2 is a measure of linkage disequilibrium.

particularly decreased in controls that have no pathology related to the disease of interest.

Table 2 shows that many studies on the predictive value of genetic profiling were conducted in hyperselected case-control series, comparing, for example, type 2 diabetes patients with normoglycemic individuals [10,11], patients with severe hypertriglyceridemia with normolipidemic controls [12], or patients with end-stage AMD with individuals who have no eye pathology [13]. By excluding individuals with modestly elevated glucose or lipid levels, these case-control series largely lose their relevance for investigating predictive potential in clinical practice, where persons with such levels are part of the population. Predicting progression to disease is most difficult in individuals with early symptoms or mild pathology, but prediction in this population is clinically highly relevant. One could argue that if the predictive value of genetic profiling is low in the samples used in these studies [10-13], it will be even poorer in unselected cohorts. Thus, hyperselected case-control studies can be useful to demonstrate that predictive genetic testing is not informative

and, given the commercial interest in genome-based applications, this is an important message to get across.

Another consideration is the use of case-control studies in general, as illustrated by the recent findings on type 2 diabetes. Lango *et al.* [11] investigated the predictive value of 18 polymorphisms in a case-control study, comparing patients with normoglycemic controls, and van Hoek *et al.* [14] looked at the same polymorphisms in a prospective cohort of individuals aged 55 years and older. In both studies [11,14], the AUC of the 18 polymorphisms was 0.60 and the improvement in AUC beyond prediction from age, sex and body mass index (BMI) was limited (Δ AUC = 0.02). But a more detailed analysis of the results reveals that even though the AUC was 0.60 in both studies, it was mainly contributed by different genetic variants in the two studies (Table 3). Moreover, the 0.02 improvement increased the AUC to 0.80 in the case-control study but to only 0.68 in the prospective cohort study. This difference is mostly explained by the difference in BMI. Mean BMI in the case-control study was 31.5 kg/m² in patients and 26.9 kg/m² in controls compared

with 28.0 kg/m² and 26.0 kg/m² in the prospective cohort study, indicating that BMI was a stronger predictor of type 2 diabetes in the case-control study.

Case-control studies tend to overestimate odds ratios and this may be related to selection bias (most likely the case in the example above) or information bias (patients may attribute a disease to a known risk factor and they over-report this exposure). An issue that is often ignored in gene discovery studies but that is extremely relevant in studies evaluating the predictive value is that of survival bias. If genes increase the risk of disease, they may also increase the risk of (early) mortality. Therefore, there are strong arguments that show the necessity that predictive testing in preventive medicine should be investigated in cohort studies consisting of individuals who do not have the disease of interest, and predictive testing for prognosis and therapy response should be evaluated prospectively in clinically relevant patient series.

There is no single golden standard by which study population and study design should be selected, other than that predictive genetic tests need to be evaluated in populations representative for their intended use. The choice of the target population is not arbitrary, but rather is a trade-off of the effectiveness, costs and harmful side effects of available interventions, among other factors. Table 2 shows three prospective cohort studies evaluating the prediction of coronary heart disease, one in Caucasian men of European ancestry aged 50-64 years [15], one in a general population of 45-64 years [16] and one in patients with familial hypercholesterolemia [17]. These different study populations assume different target populations for genetic profiling, and the predictive value will differ between these populations when disease risks, genotype frequencies and effect sizes are different.

Moving from risk variants to genomic profiles

When the predictive value of a limited number of variants is investigated in a large population-based study, disease risks can be calculated as the percentage of patients for each combination of genotypes. However, the number of genotype combinations increases exponentially with the number of variants tested. For example, combining 18 variants that have three possible genotypes, as did two studies of type 2 diabetes, theoretically yields 387,420,489 (3^{18}) unique profiles. To deal with such a large number, researchers adopt one of two approaches for the calculation of disease risks. First, they may calculate risk allele scores or genotype scores obtained by counting the number of risk alleles across all variants [10,11,14-23]. This approach assumes that the differences between the effects of the individual variants can be ignored, which may be a realistic assumption for multifactorial disorders given that the effect sizes are generally small [24]. Second, researchers may use logistic or Cox proportional hazards regression analyses for risk

prediction, which do account for differences in effects sizes between individual variants. Risk predictions from regression analysis can be regarded as weighted risk scores. Table 2 shows that all studies applied either logistic or Cox proportional hazards regression analyses, some in addition to the simpler risk allele scores.

In addition to the question of how to combine genetic variants into profiles, the question arises as to which of the variants to include. Several studies include variants that were already established risk factors (Table 2) [11,12,14,19, 21-23,25]. Others include polymorphisms from candidate genes or regions that have been associated with disease risk in at least one other study or that are likely to be functionally implicated (for example, [13,15,17,18,20,26,27]). And again others include polymorphisms identified in their own genome-wide association study [10,28]. Although the distinction between candidate and established variants is not crystal clear and findings from genome-wide association studies may be robust, we can expect that the predictive value of variants that are less convincingly established is less likely to be replicated in independent populations.

Another important issue in obtaining accurate estimates of the genetic predisposition at the individual level is how to handle gene-gene and gene-environment interaction in the prediction of common diseases. It is frequently argued that strong effects can be seen from the interaction of a gene with other genetic variants or environmental factors. Several studies reported in Table 2 investigated the presence of gene-gene interaction [19,23,26], but none included interaction effects in the regression models. The reported effect sizes for interaction terms were very modest, implying that the influence on the predictive value of risk profiles would have been limited [19]. When future studies give robust evidence for interaction, these interaction effects should be taken into account in the risk prediction.

Last but not least, we can anticipate improvement in the predictive value when we can identify the exact causal variants. Most variants that are included in the genetic profiles shown in Table 2 are derived directly from genome-wide association studies. There is a growing awareness that these might not be the causal variants and that the causal variants may have a very different allele distribution in patients and controls. It is anticipated that the causal variants will have stronger effects on disease risk. The large deep-sequencing efforts that are ongoing may shed light on this question.

Evaluation of the predictive value

The question of how well genetic profiles can predict disease can be answered by many different performance measures, which all are related but which highlight different features. Which measure is of interest depends on the question addressed. Individuals who undergo genetic testing will be

most interested in their absolute risks of disease conditional on their genetic profile. Only a few empirical studies have presented absolute risks [14,17,18], most likely because these cannot be calculated from case-control data without assumptions on the incidence of disease. Many other studies have reported used risk ratios (odds ratios, relative risks or hazard ratios), which each compare the risks of disease with a reference risk, namely that of individuals who carry no risk alleles. Here, also, evaluation studies diverge from gene discovery studies. Although the comparison with those with the lowest number of risk alleles is a valid approach and the recommended strategy in gene-discovery studies, it is less relevant in translational studies. Individuals who undergo genetic testing and receive their results are not interested in learning their risk compared with individuals who have an extremely low risk of disease [29] but rather compared with the average risk of disease, that is, the risk before testing, which for a common disease such as type 2 diabetes may be as high as 10%. Thus, comparing the risk or odds of disease with those with the average risk is more appropriate [29].

When deciding about whether or not to perform a test from a clinical perspective, physicians need to know to what extent a test can make a difference. This makes them more interested in the distribution of risk allele scores and, related to that, the distribution of risks and risk ratios. Many empirical studies do present distributions of risk allele scores [10,11,16,18-22], and several others do present risks associated with the risk allele scores but do not show their distribution (Table 2) [13,14,17,23,25,27,30]. These distributions are all different presentations of the discriminative accuracy of a test, generally measured as the AUC (see earlier). All but two [23,30] of the studies shown in Table 2 evaluated the AUC of genetic profiling. Despite reported shortcomings [31-33], AUC is very suitable as a first screening indication of predictive value. Further evaluation of clinical validity and utility is warranted only if a reasonable AUC is demonstrated at first. This further evaluation can include evaluation of absolute risks, reclassification [31], net reclassification improvement and integrated discrimination improvement [32]. The value of reclassification should not be overestimated, as illustrated by the study of Kathiresan and colleagues [18], who studied the addition of genetic factors to traditional risk factors for cardiovascular disease. In this study, adding genetic variants did not improve the AUC, but 26% of the individuals in the intermediate risk group (absolute risks 10-20%) were reclassified into the lower and higher risk groups. A closer look at the findings shows that the observed risk of those who were reclassified to the lower risk group was 8.2%, only slightly lower than the cut-off value of 10%, and the observed risk of those reclassified to the highest risk category was 14.7%, which was similar to the observed risk among those who remained in the intermediate category (14.5%). Reclassification may thus not lead to better classification when no improvement in AUC is seen.

Moving towards the calibration and validation of the predictive value

Prediction of complex diseases from risk allele scores or on regression models makes several assumptions. As discussed earlier, risk allele scores assume that the differences in effect sizes between the individual variants can be ignored and that there is no gene-gene interaction. Regression methods generally do not consider gene-gene or gene-environment interaction effects either. One way to test whether these assumptions are reasonable is to evaluate the concordance between observed and expected disease risks, a method that is called calibration. Although calibration is an essential step in the development of clinical prediction models, it was examined only in two of the studies reported in Table 2 [12,22]. Wang *et al.* [12] found that the prediction of hypertriglyceridemia from traditional and genetic factors showed good calibration, indicating that observed risks were reasonably predicted by a regression model that did not include interaction effects.

All prediction models perform best in the dataset from which they were obtained. Therefore, it is crucial to replicate the predictive value of genetic profiling in independent datasets. Validation investigates the extent to which genetic profiles have similar predictive value in independent datasets. In large-scale studies, prediction models usually are developed using part of the data and applied to predict the outcome of interest in the rest (internal validation). In addition, the prediction needs to be evaluated in an independent dataset (external validation) to demonstrate its value.

Table 2 shows that only one study performed internal validation for the selection of the markers [16], and none performed external validation. One might think that the two studies investigating the same 18 polymorphisms in type 2 diabetes are replication studies [11,14], but this is not true. Because the studies were published at the same time, they each developed their own prediction model. Given that these prediction models had only four variants in common, it is reasonable to expect that the AUC would be lower if the two research groups had validated their models on each others data. The lack of replication studies is, however, not so much a problem for studies that already show poor predictive value in the original population, as the predictive value typically becomes worse in the replication study. It is, however, important for studies that potentially show useful predictive value, such as the study on myocardial infarction following surgery (AUC 0.76) [26] and the studies in AMD and hypertriglyceridemia (both AUC 0.80) [12,13].

Moving from clinical validity to personalized medicine

Whether genetic testing is useful for public health or clinical practice depends on what the implications of the test results are. The usefulness depends on the availability of alternative strategies for disease prediction and the availability of

preventive or therapeutic interventions that can be targeted to genetic profiles, among other factors, such as effectiveness and cost-effectiveness of interventions, and patient preferences and attitudes.

An important consideration is whether genetic profiles yield substantially better predictive value than traditional risk factors. For instance, genes associated with cardiovascular disease may also be involved in intermediate outcomes, such as dyslipidemia or hypertension [18,34]. From a theoretical perspective, genetic factors will not remain significant when considering both genetic factors and intermediate outcomes in the prediction analysis [5]. Overall, genetic factors will not be better predictors of disease risk than intermediate factors, but their greater ease of assessment may be worth a slight reduction in the predictive value. However, it should be realized that even when genetic profiles predict disease equally as well as intermediate biomarkers, this does not mean that they are equally useful. Typical intermediate outcomes suggest the existence of early pathology and point to clear targets for intervention, such as weight loss or medication for lowering blood pressure or cholesterol, whereas targeted interventions are often not clear for genetic risks. An exception is intensive surveillance, which is useful to broader populations at risk, independent of the underlying pathology.

Another issue is the availability of specific interventions for specific genetic profiles. Very often the number of alternative therapeutic interventions is quite limited. Also, from a public health perspective it can be argued that most preventive strategies, such as weight control and smoking cessation, will have effects on multiple disease outcomes, making it unreasonable - and unethical - to specifically target these strategies to people on the basis of a genetic profile that increases their risk for a single disorder and to withhold it from others. That is, not only persons at increased genetic risk for diabetes should be advised to control their weight, but also persons at increased risk for other disorders such as arthritis, cardiovascular disease and cancer. Thus, a key question to answer is how we can justify personalization of preventive and therapeutic interventions.

Conclusions

Prediction studies so far have been rather simplistic in the sense that most were based on a small number of variants that by themselves explain only a fraction of the genetic variability, were conducted in non-representative cohorts, were neither calibrated nor validated and hardly investigated clinical utility. This should not be interpreted as shortcomings of these studies; questions concerning calibration, validation and clinical utility are relevant only for genetic profiles with promising discriminative values. On the basis of AUC, further evaluations could be worthwhile for AMD and hypertriglyceridemia [12,13], if only to find out whether their very high discriminative accuracy (AUC = 0.80) was

due to the hyperselected case-control design or to true strong genetic effects. Replication is warranted in independent population-based prospective cohort studies that include the whole range of the clinical spectrum.

Another important question is the level of predictive value that is to be targeted before implementing a genomic profile. The level aimed for depends on the intended application, particularly on the goal of testing, the medical, psychological and financial burden of the disease, the availability of (preventive) treatment and the adverse effects of false-positive and false-negative test results. The aim of genetic screening is often to select high-risk subjects for preventive treatment or intensified surveillance programs. High predictive value is needed for interventions that are invasive and irreversible, whereas lower predictive value may be sufficient for interventions such as adopting a healthy diet or increasing physical activity, which are beneficial and not harmful for a broader population.

A legitimate question is whether we should evaluate predictive genetic testing for common diseases at the moment. It is clear that our current knowledge of their genetic basis is insufficient, and will probably remain so for the next five years. However, the current interest from biotechnology companies that offer genetic profiling on the internet, and from customers who want to learn about their risks of disease, currently asks for empirical evidence. Whether future genetic profiles should only be offered commercially if the clinical utility has been proven beyond reasonable doubt (as is the case for medical tests and treatments) or can enter the market if proven not harmful (as expected for health products such as vitamins and anti-aging cosmetics) remains an open question. From a clinical and public health perspective, we need to build the knowledge base that is needed to identify useful genome-based applications for implementation in a clinical setting.

Abbreviations

AMD, age-related macular degeneration; AUC, area under the receiver operating characteristic curve; BMI, body mass index.

Acknowledgements

This study was supported by the Centre for Medical Systems Biology in the framework of the Netherlands Genomics Initiative. ACJWJ was sponsored by the VIDJ grant of the Netherlands Organisation for Scientific Research (NWO). We thank M van Hoek for providing the per-allele effect sizes for the results of her paper [14].

References

1. Manolio TA, Brooks LD, Collins FS: **A HapMap harvest of insights into the genetics of common disease.** *J Clin Invest* 2008, **118**:1590-1605.
2. Guttmacher AE, Collins FS: **Realizing the promise of genomics in biomedical research.** *JAMA* 2005, **294**:1399-1402.

3. Brand A, Brand H, Schulte in den Bäumen T: **The impact of genetics and genomics on public health.** *Eur J Hum Genet* 2008, **16**:5-13.
4. Pharoah PD, Antoniou AC, Easton DF, Ponder BA: **Polygenes, risk prediction, and targeted prevention of breast cancer.** *N Engl J Med* 2008, **358**:2796-2803.
5. Janssens ACJW, Van Duijn CM: **Genome-based prediction of common diseases: advances and prospects.** *Hum Mol Genet* 2008, **17**:R166-R173.
6. Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, Van Duijn CM: **Predictive testing for complex diseases using multiple genes: fact or fiction?** *Genet Med* 2006, **8**:395-400.
7. Janssens ACJW, Gwinn M, Bradley LA, Oostra BA, Van Duijn CM, Khoury MJ: **A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions.** *Am J Hum Genet* 2008, **82**:593-599.
8. Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L: **The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention?** *Genet Med* 2007, **9**:665-674.
9. Haga SB, Khoury MJ, Burke W: **Genomic profiling to promote a healthy lifestyle: not ready for prime time.** *Nat Genet* 2003, **34**:347-350.
10. Cauchi S, Meyre D, Durand E, Proença C, Marre M, Hadjadj S, Choquet H, De Graeve F, Gaget S, Allegaert F, Delplanque J, Permutt MA, Wasson J, Blech I, Charpentier G, Balkau B, Vergnaud AC, Czernichow S, Patsch W, Chikri M, Glaser B, Sladek R, Froguel P: **Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value.** *PLoS ONE* 2008, **3**: e2031.
11. Lango H: **UK Type 2 Diabetes Genetics Consortium, Palmer CN, Morris AD, Zeggini E, Hattersley AT, McCarthy MI, Frayling TM, Weedon MN: Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk.** *Diabetes* 2008, **57**:3129-3135.
12. Wang J, Ban MR, Zou GY, Cao H, Lin T, Kennedy BA, Anand S, Yusuf S, Huff MW, Pollex RL, Hegele RA: **Polygenic determinants of severe hypertriglyceridemia.** *Hum Mol Genet* 2008, **17**:2894-2899.
13. Maller J, George S, Purcell S, Fagerness J, Altschuler D, Daly MJ, Seddon JM: **Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration.** *Nat Genet* 2006, **38**:1055-1059.
14. van Hoek M, Dehghan A, Witteman JCM, Van Duijn CM, Uitterlinden AG, Oostra BA, Hofman A, Sijbrands EJ, Janssens AC: **Predicting type 2 diabetes based on polymorphisms from genome wide association studies: a population-based study.** *Diabetes* 2008, **57**:3122-3128.
15. Humphries SE, Cooper JA, Talmud PJ, Miller GJ: **Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men.** *Clin Chem* 2007, **53**:8-16.
16. Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, Kane JP, Pankow JS, Devlin JJ, Willerson JT, 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.
17. van der Net JB, Janssens ACJW, Defesche JC, Kastelein JJP, Sijbrands EJG, Steyerberg EW: **Usefulness of genetic polymorphisms and conventional risk factors to predict coronary heart disease in patients with familial hypercholesterolemia.** *Am J Cardiol* 2009, **103**: 375-380.
18. Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, Hirschhorn JN, Berglund G, Hedblad B, Groop L, Altschuler DM, Newton-Cheh C, Orho-Melander M: **Polymorphisms associated with cholesterol and risk of cardiovascular events.** *N Engl J Med* 2008, **358**:1240-1249.
19. Weedon MN, McCarthy MI, Hitman G, Walker M, Groves CJ, Zeggini E, Rayner NW, Shields B, Owen KR, Hattersley AT, Frayling TM: **Combining information from common type 2 diabetes risk polymorphisms improves disease prediction.** *PLoS Med* 2006, **3**:e374.
20. Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, Adami HO, Hsu FC, Zhu Y, Bälter K, Kader AK, Turner AR, Liu W, Bleecker ER, Meyers DA, Duggan D, Carpten JD, Chang BL, Isaacs WB, Xu J, Grönberg H: **Cumulative association of five genetic variants with prostate cancer.** *N Engl J Med* 2008, **358**:910-919.
21. Lysenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altschuler D, Nilsson P, Groop L: **Clinical risk factors, DNA variants, and the development of type 2 diabetes.** *N Engl J Med* 2008, **359**:2220-2232.
22. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, Cupples LA: **Genotype score in addition to common risk factors for prediction of type 2 diabetes.** *N Engl J Med* 2008, **359**:2208-2219.
23. Weersma RK, Stokkers PC, van Bodegraven AA, van Hogezaand RA, Verspaget HW, de Jong DJ, van der Woude CJ, Oldenburg B, Linskens R, van der Steege G, Festen EA, Hommes DW, Crusius JB, Wijmenga C, Nolte IM, Dijkstra G, Dutch Initiative on Crohn and Colitis (ICC): **Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort.** *Gut* 2009, **58**:338-395.
24. Janssens ACJW, Moonesinghe R, Yang Q, Steyerberg EW, Van Duijn CM, Khoury MJ: **The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases.** *Genet Med* 2007, **9**:528-535.
25. Yeh CC, Sung FC, Tang R, Chang-Chieh CR, Hsieh LL: **Association between polymorphisms of biotransformation and DNA-repair genes and risk of colorectal cancer in Taiwan.** *J Biomed Sci* 2007, **14**:183-193.
26. Podgoreanu MV, White WD, Morris RW, Mathew JP, Stafford-Smith M, Welsby JJ, Grocott HP, Milano CA, Newman MF, Schwinn DA; Perioperative Genetics and Safety Outcomes Study (PEGASUS) Investigative Team: **Inflammatory gene polymorphisms and risk of postoperative myocardial infarction after cardiac surgery.** *Circulation* 2006, **114**:1275-1281.
27. Vaxillaire M, Veslot J, Dina C, Proença C, Cauchi S, Charpentier G, Tichet J, Fumeron F, Marre M, Meyre D, Balkau B, Froguel P; DESIR Study Group: **Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study.** *Diabetes* 2008, **57**:244-254.
28. International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN), Harley JB, Alarcón-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, Moser KL, Tsao BP, Vyse TJ, Langefeld CD, Nath SK, Guthridge JM, Cobb BL, Mirel DB, Marion MC, Williams AH, Divers J, Wang W, Frank SG, Namjou B, Gabriel SB, Lee AT, Gregersen PK, Behrens TW, Taylor KE, Fernando M, Zidovetzki R, Gaffney PM, Edberg JC, Rioux JD, et al.: **Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXX, KIAA1542 and other loci.** *Nat Genet* 2008, **40**:204-210.
29. Janssens ACJW, Van Duijn CM: **Five genetic variants associated with prostate cancer.** *N Engl J Med* 2008, **358**:2739.
30. Lysenko V, Almgren P, Anevski D, Orho-Melander M, Sjogren M, Saloranta C, Tuomi T, Groop L; Botnia Study Group: **Genetic prediction of future type 2 diabetes.** *PLoS Med* 2005, **2**:e345.
31. Cook NR: **Use and misuse of the receiver operating characteristic curve in risk prediction.** *Circulation* 2007, **115**:928-935.
32. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS: **Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond.** *Stat Med* 2008, **27**:157-172.
33. Pepe MS, Janes HE: **Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer.** *J Natl Cancer Inst* 2008, **100**:978-979.
34. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, Stacey SN, Bergthorsson JT, Thorlacius S, Gudmundsson J, Jonsson T, Jakobsdottir M, Saemundsdottir J, Olafsdottir O, Gudmundsson LJ, Bjornsdottir G, Kristjansson K, Skuladottir H, Isaksson HJ, Gudbjartsson T, Jones GT, Mueller T, Gottsäter A, Flex A, Aben KK, de Veigt F, et al.: **A variant associated with nicotine dependence, lung cancer and peripheral arterial disease.** *Nature* 2008, **452**:638-642.
35. van der Net JB, van Etten J, Yazdanpanah M, Dallinga-Thie GM, Kastelein JJ, Defesche JC, Koopmans RP, Steyerberg EW, Sijbrands EJ: **Gene-load score of the renin-angiotensin-aldosterone system is associated with coronary heart disease in familial hypercholesterolaemia.** *Eur Heart J* 2008, **29**:1370-1376.