Human Molecular Genetics, 2008, Vol. 17, Review Issue 2 ***R166–R173***
doi:10.1093/hmg/ddn250

Genome-based prediction of common diseases: advances and prospects

A. Cecile J.W. Janssens[1,2,*] and Cornelia M. van Duijn[2]

[1]Department of Public Health and [2]Department of Epidemiology, Erasmus University Medical Center Rotterdam, Rotterdam, the Netherlands

Received July 9, 2008; Revised and Accepted August 16, 2008

Common diseases such as type 2 diabetes and coronary heart disease result from a complex interplay of genetic and environmental factors. Recent developments in genomics research have boosted progress in the discovery of susceptibility genes and fueled expectations about opportunities of genetic profiling for personalizing medicine. Personalized medicine requires a test that fairly accurately predicts disease risk, particularly when interventions are invasive, expensive or have major side effects. Recent studies on the prediction of common diseases based on multiple genetic variants alone or in addition to traditional disease risk factors showed limited predictive value so far, but all have investigated only a limited number of susceptibility variants. New gene discoveries from genome-wide association studies will certainly further improve the prediction of common diseases, but the question is whether this improvement is sufficient to enable personalized medicine. In this paper, we argue that new gene discoveries may not evidently improve the prediction of common diseases to a degree that it will change the management of individuals at increased risk. Substantial improvements may only be expected if we manage to understand the complete causal mechanisms of common diseases to a similar extent as we understand those of monogenic disorders. Genomics research will contribute to this understanding, but it is likely that the complexity of complex diseases may ultimately limit the opportunities for accurate prediction of disease in asymptomatic individuals as unraveling their complete causal pathways may be impossible.

INTRODUCTION

Genome-wide association studies are rapidly unraveling the role of genetic factors in the pathogeneses of common diseases (1). One of the major promises is that these advances will lead to personalized medicine, in which preventive and therapeutic interventions for complex diseases are tailored to individuals based on their *genetic profiles* (2,3). Personalized medicine already exists for monogenetic disorders such as Huntington disease, phenylketonuria (PKU) and hereditary forms of cancer, in which genetic testing is the basis for informing individuals about their future health status and for deciding upon specific, often radical interventions such as lifetime dietary restrictions and preventive surgery. Yet, the etiology of complex diseases is essentially different from that of monogenic diseases, and hence translating the new emerging genomic knowledge into public health and medical care is one of the major challenges for the next decades (4,5).

An essential prerequisite for personalized medicine to become feasible is a predictive test or prediction model that can discriminate between individuals who will develop the disease of interest and those who will not. The level of discrimination that is required in clinical care and public health applications depends, among other things, on the goal of testing, the burden of disease, the costs of disease, the availability of (preventive) treatment and the adverse effects of false-positive and false-negative test results. In this paper, we review recent studies that have examined the prediction of common diseases based on multiple genetic variants alone or in addition to traditional disease risk factors, and we discuss factors that determine the prospects of personalized medicine including new discoveries from genome-wide association studies, classical

*To whom correspondence should be addressed at: Department of Epidemiology, Erasmus University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, the Netherlands. Tel: +31 107044232; Fax: +31 107044657; Email: a.janssens@erasmusmc.nl

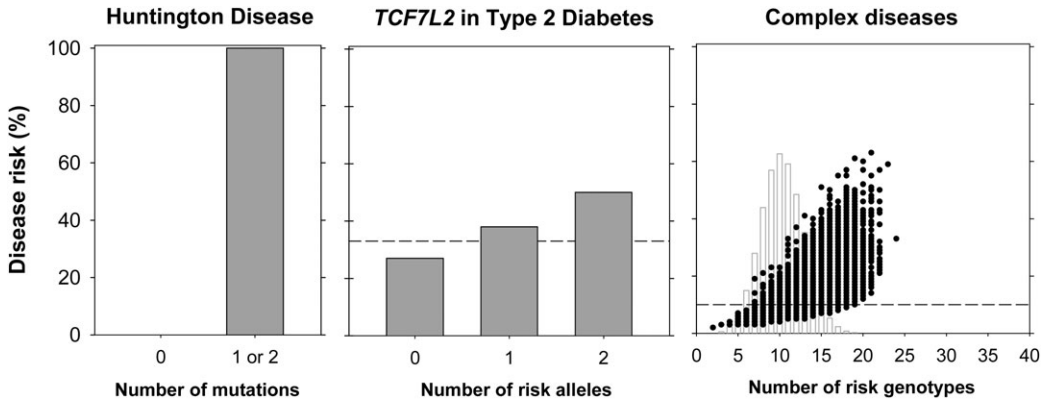

Figure 1. Disease risks associated with single genetic variants and genetic profiles. Disease risks for *TCF7L2* were based on odds ratios from a recent meta-analysis (38) using a population risk of type 2 diabetes of 33% (dashed line) (39). Disease risks for the complex diseases example were based on simulated data assuming a population risk of disease of 10% (dashed line), frequencies of the risk genotypes varying between 1 and 60% and odds ratios varying from 1.05 to 2.0. The bars in the scatterplot represent the frequency distribution of the number of risk genotypes. The example and the simulation strategy have been described previously (19,40).

knowledge of the heritability and the etiologic complexity of common diseases.

GENETIC TESTING IN COMMON DISEASES

The genetic origin of common complex or multifactorial diseases differs essentially from that of monogenic disorders. Monogenic disorders, such as Huntington disease, PKU and hereditary cancers, are completely or predominantly caused by DNA variations in one single gene, and hence, carriers of mutations typically have distinctly higher disease risks than non-carriers (Fig. 1A). These substantial differences in absolute risks of disease warrant differential preventive and therapeutic strategies for different genotype groups, if available. Complex diseases result from the joint effects of multiple genetic and environmental causes, with each factor having only a minor contribution to the occurrence of disease. Consequently, risks of disease differ only marginally between carriers and non-carriers of risk variants of one single susceptibility gene (Fig. 1B), and prediction of disease based on a single genetic variant is considered not informative (6,7). Genome-based prediction of complex diseases will imply the simultaneous testing at multiple genetic loci, known as genetic profiling.

The predictive value of genetic profiling has been investigated in a few empirical studies to date, and this number is steadily increasing. These studies, in type 2 diabetes, coronary heart disease, myocardial infarction and age-related macular degeneration (AMD) generally showed limited predictive value so far (Table 1), with the exception of the five susceptibility variants involved in AMD and the seven variants in hypertriglyceridemia (8–17). While the predictive value in these two studies may have been overestimated because they were performed in hyperselected populations not representative for clinical practice, comparing individuals with end-stage AMD with those without eye abnormalities (8,18) and individuals with hypertriglyceridemia with normolipidemic individuals (11), prediction of these disorders is deemed promising as the individual variants have very strong effects compared to what is generally seen in common diseases.

The genetic profiles that have been investigated empirically to date included only a small number of mostly weak susceptibility variants and they, therefore, are not yet useful for the application in clinical medicine or public health. Simulation studies have shown that the predictive value of a larger number of genes (up to hundreds) can theoretically attain the same level as that of traditional risk factors predicting cardiovascular disease (19,20), but it may not evidently become much better. There is a 'natural' limit to the predictive value of genetic profiling as common diseases that are only partly influenced by genetic factors that can never be perfectly predicted by genetic testing. In fact, the maximum predictive value in terms of the area under the receiver operating characteristic curve (AUC, see table legends for details) that can be attained for genetic profiling is determined by the heritability and the prevalence of the disease (19). For applications in health care, this means that genetic profiling may become useful for the identification of individuals at increased risk of disease to the same extent as traditional risk factors do, but that its predictive value may not be high enough for decisions about invasive, irreversible and expensive interventions or for presymptomatic diagnosis.

Recent simulation studies have also demonstrated some interesting features of genetic profiles that explain why the predictive value of a larger number of multiple weak susceptibility variants may not easily become much better (19,21–23). First, when multiple genes are considered simultaneously, one typically finds that all individuals in a population carry at least one or more risk genotypes, even those persons with a lower than average risk of disease (bar chart in Fig. 1C). Second, the more risk genotypes the higher the risk of disease, but substantial variation in disease risk may be seen between individuals with the same number of risk genotypes resulting from differences in effect sizes between risk genotypes (scatterplot in Fig. 1C). And third, in sharp contrast to genetic testing for monogenic diseases, most individuals have profiles that are associated with disease risks that are only slightly higher or lower than the average risk in the population (combination of bar chart and scatter plot, Fig. 1C). Figure 1C shows that also individuals who have e.g. 10 risk genotypes may have

Table 1. Recent studies on the prediction of complex diseases using multiple genes

Disease	Genetic variants	Variant selection[a]	AUC	Reference
Age-related macular degeneration	*CFH* Y402H, *CFH* rs1410996, *LOC387715* A69S, *C2-CFB*	5 (out of 1536 tag SNPs in established genes)	0.80[b]	(8)
Coronary heart disease	*UCP2* G(-866)A, *APOE* e2/3/4, *LPL* D9N, *APOA4* T347S	4 (out of 12)	0.62	(9)
Coronary heart disease	*AGT* T4072C, *ACE* I/D, *AGTR1* A1166C, *CYP11B2* C(-344)T, *ADD1* G614T, *GNB3* C825T	6 established variants	0.55[c]	(10)
Hypertriglyceridemia	*APOA5* S19W, *APOA5* T(-1131)C, *APOE e/3/4, GCKR* rs780094, *TRIB1* rs17321515, *TBL2/MLXIPL* rs17145738, *GALNT2* rs4846914	7 established variants	0.80	(11)
MI after surgery	*IL6* G572C, *ICAM1* K469E, *SELE* G98T	3 (out of 48)	0.70	(12)
Systemic lupus erythematosus	*PXK* rs6445975, *HLA* region rs3131379 and rs9275572, *IRF5/TNPO3* rs12537284, *KIAA1542* rs4963128, *ITGAM* rs9888739	From GWAS	0.67	(13)
Type 2 diabetes	*KCNJ11* G23L, *PPARG* P12A, *TCF7L2* rs7903146	3 established variants	0.55	(14)
Type 2 diabetes	*GCK* G(−30G)A, *IL6* G(−174)C, *TCF7L2* rs7903146	3 (out of 19)	0.56	(15)
Type 2 diabetes	*SNPs in TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX, THADA, JAZF1*	18 established variants	0.60	(16)
Type 2 diabetes	*SNPs in TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX, THADA, JAZF1*	18 established variants	0.60	(17)

MI, myocardial infarction; AUC, area under the receiver operating characteristic curve; SNP, single nucleotide polymorphism. AUC indicates the discriminative accuracy, 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 discrimination).

[a]Numbers between brackets indicate the total number of variants at the start of the analysis from which the most predictive variants were selected. Established means that the variants were selected from the literature, based on association with disease risk in other studies.

[b]AUC was calculated in letter to the editor based on the original data (18).

[c]AUC obtained from re-analysis of the original data, not published in the original article.

disease risks that are lower than the average risk in the population. This is explained by the fact that individuals who have 10 risk genotypes out of 40 variants tested also have 30 'protective' genotypes that decrease their risks. Whether the risk associated with a genetic profile is higher or lower than average depends on the extent to which the risk increase by the risk genotypes outbalances the risk decrease by the protective genotypes.

IMPROVING DISEASE PREDICTION

While testing multiple susceptibility variants alone may not yield perfect prediction of complex diseases, the question remains whether it will improve the prediction of disease beyond classical risk factors. Although the construction of profiles consisting of genetic and environmental risk factors appears an obvious solution, studies so far showed that genetic factors do not substantially improve the prediction of type 2 diabetes, coronary heart disease and prostate cancer, but again the number of genes investigated was small (Table 2) (9,12,15,16,24–27). However, from a theoretical perspective, it can be argued that also a large number of genes will unlikely have substantial added predictive value over traditional risk factors if these variants predispose the risk factors. For instance, genes associated with cardiovascular disease may also be involved in intermediate outcomes as dyslipidemia or hypertension or even smoking (26,28). According to the basic principles of epidemiological research, genetic variants involved in intermediate factors will not remain significant when they are entered in a regression model together with these intermediate factors (Fig. 2). Genetic variants may improve disease prediction beyond traditional risk factors when they are involved in unknown pathways or in pathways with unmeasurable intermediate factors. New yet unknown pathways may be more likely for some diseases than for others. A critical but not unlikely note is that gene discoveries may also identify novel etiological pathways and novel intermediate biomarkers, which consequently may be stronger predictors of disease than the genetic variant that led to its identification.

COMPLETE CAUSAL MECHANISMS

One of the paradigms in complex genetics is that the genetic prediction of common diseases can be substantially improved if we are able to identify genetic variants with strong effects, either on their own or in interaction with other variants or with environmental factors, i.e. gene–gene or gene–environment interaction. Yet, perfect prediction of disease, e.g. comparable to that of the genetic test for Huntington Disease, may only be achieved if we are able to understand the essential genetic and environmental factors in the causal mechanisms of the disease, to a similar extent as we understand the causal mechanism that leads to Huntington Disease.

Unraveling the underlying causal pathways implies that we are to understand why someone developed a certain disease. Rothman and Greenland (29) define a complete causal mechanism or a sufficient cause of disease as a set of minimal conditions and events that inevitably lead to disease, with 'minimal' implying that all component causes need to be

Table 2. Recent studies on the improvement of clinical prediction by testing multiple susceptibility genes

Disease	Clinical risk factors	Variant selection[a]	Genetic variants	AUC before	AUC after	Reference
Cardiovascular disease	Age, sex, family history of myocardial infarction, low density lipoprotein, high density lipoprotein cholesterol, triglycerides, systolic blood pressure, diastolic blood pressure, diabetes mellitus, body mass index, smoking, C-reactive protein, lipid-lowering therapy, antihypertensive treatment	9 (out of 11) established SNPs in 9 genes	*APOB* rs693, *APOE* cluster rs4420638, *HMGCR* rs12654264, *LDLR* rs1529729, *PCSK9* rs11591147, *ABCA1* rs3890182, *CETP* rs1800775, *LIPC* rs1800588, and *LPL* rs328	0.80	0.80	(26)
Coronary heart disease	Age, triglycerides, cholesterol, systolic blood pressure, smoking	4 (out of 12)	*UCP2* G(−866)A, *APOE* e2/3/4, *LPL* D9N, *APOA4* T347S	0.66	0.70	(9)
Coronary heart disease: in whites	Age, systolic blood pressure, total cholesterol, high density lipoprotein cholesterol, diabetes, use of antihypertensive medication, smoking	11 (out of 116)	*VAMP8, PALLD, KIF6, MKI67, MYH15, Loc646377, HPS1, SNX19, ADAMTS1 (2x), ADRB3*	0.76	0.77	(25)
Coronary heart disease: in blacks	Age, systolic blood pressure, total cholesterol, high density lipoprotein cholesterol, diabetes, use of antihypertensive medication, smoking	11 (out of 116)	*DMXL2, ZNF132, KIF6, F2, OR2A25, KRT5, CTNNA3, HAP1, GIPR, FSTL4, THBS2*	0.76	0.77	(25)
MI after surgery	AXT time, number of coronary grafts, previous cardiac surgery	3 (out of 48)	*IL6* G572C, *ICAM1* K469E, *SELE* G98T	0.70	0.76	(12)
Prostate cancer	Age, geographic region, family history	5 (out of 16) in 5 established regions)	rs4430796 (in 17q12), rs1859962 (in 17q24.3), rs16901979, rs6983267 and rs1447295 (all in 8q24)	0.61	0.63	(27)
Type 2 diabetes	Body mass index, plasma glucose level	3 (out of 6)	*PPARG* P12A, *CAPN10* SNP43 and SNP44	0.68[b]	0.68[b]	(24)
Type 2 diabetes	Age, sex, body mass index	3 (out of 19)	*GCK* G(−30G)A, *IL6* G(−174)C, *TCF7L2* rs7903146	0.82	0.82	(15)
Type 2 diabetes	Age, sex, body mass index	18 established variants	SNPs in *TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/ NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX-IDE, THADA, JAZF1*	0.78	0.80	(16)
Type 2 diabetes	Age, sex, body mass index	18 established variants	SNPs in *TCF7L2*, 2 in *CDKN2A/2B, KCNJ11, PPARG, ADAM30/ NOTCH2, IGF2BP2, FTO, CDKAL1, SLC30A8, TSPAN8//LGR5, CDC123, WFS1, TCF2, ADAMTS9, HHEX-IDE, THADA, JAZF1*	0.66	0.68	(17)

AUC, area under the receiver operating characteristic curve. AUC indicates the discriminative accuracy, 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 discrimination).

[a]Numbers between brackets indicate the total number of variants at the start of the analysis from which the most predictive variants were selected. Established means that the variants were selected from the literature, based on association with disease risk in other studies.

[b]AUC was calculated in letter to the editor based on the original data (41).

present for the disease to develop. When only one causal pathway is involved, the risk of disease is 100% when all component causes are present, and 0% when one or more causes are absent. Chance or randomness does not exist in the Rothman and Greenland models of complete causal mechanisms.

Figure 3 gives schematic diagrams for complete causal mechanisms of monogenic and complex diseases. Figure 3A presents a sufficient cause diagram for Huntington disease, in which there is only one causal factor. CAG extensions in the huntingtin gene are a complete and sufficient cause for the development of the disease despite the fact that there may be genes that modify age of onset. Figure 3B presents a complete causal model for PKU, which only occurs when homozygous carriers of mutations in the PAH gene are on a normal diet that includes phenylalanine. From a statistical perspective, this model serves as a typical example of gene–environment interaction.

For common diseases that result from multiple genetic and environmental causes, the complete causal mechanisms are by far more complex (Fig. 3C–F). Not only will they consist of a large number of different component causes, but a specific disease may also result from different causal mechanisms. For example, a complex disease may be caused by the presence of four different risk variants in different genes (G1 to G4 in Fig. 3C), but if one of the risk variants is absent (G4) then still the disease may inevitably occur when instead four other genetic risk variants (G5 to G8) and an environmental risk factor (E1) are present (Fig. 3D). Thus, for complex diseases, there are not one but many distinct combinations of risk factors that lead to disease development, with major single risk factors emerging in multiple combinations.

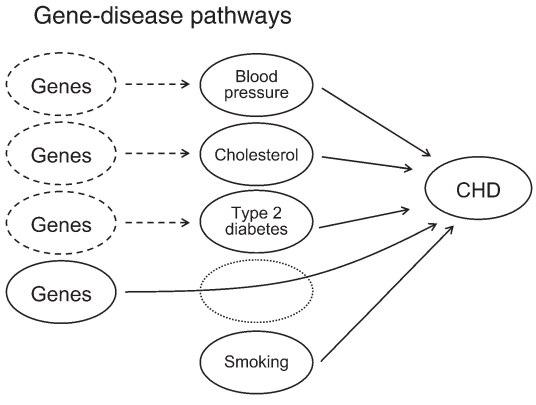

Figure 2. Predictive factors in a pathway model for complex diseases. Schematic (and incomplete) presentation of pathways that are involved in coronary heart disease (CHD). All interactions between the risk factors have been omitted. The dotted circle indicates unmeasured or unknown intermediate factors in other pathways.

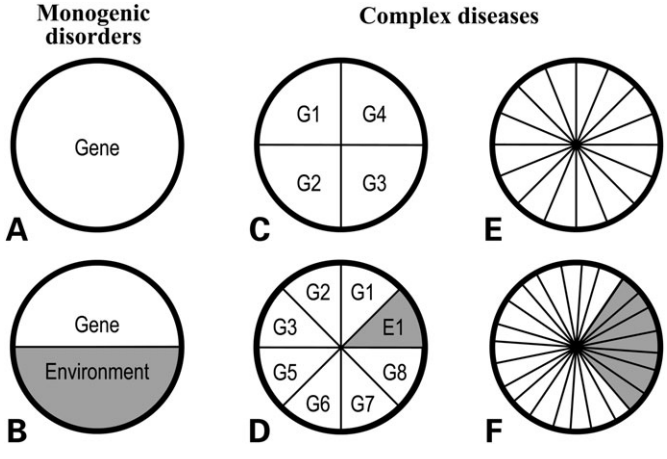

Figure 3. Complete cause models or sufficient causes of disease development. Complete causal models for (**A**) Huntington Disease; (**B**) Phenylketonuria; (C–F) Hypothetical examples for complex diseases. White areas refer to genetic factors and grey areas to environmental factors.

FROM CAUSAL MODELS TO DISEASE PREDICTION

Discovering complete causal mechanisms of common diseases implies the identification of specific combinations of causal factors among all possible combinations, namely identifying those combinations that inevitably lead to disease. And this may not be easy. As most multi-factorial diseases are caused by a complex interplay of many genetic and non-genetic factors, the number of potential combinations of these many factors is extremely large and easily outnumbers even the size of large cohorts or consortia. For example, the simultaneous testing of nine genetic variants, with three genotypes each, gives 3^9 or 19 683 potential genotype combinations and the testing of 12 variants gives 531 441 combinations. When a genetic profile of 12 variants is tested in a cohort of e.g. 30 000 individuals, all cases but also all controls will likely have unique profiles even when risk variants are common. This fact has two implications. First, it will be very difficult to prove that the profiles that are found only among cases actually are complete causal mechanisms, since it is extremely unlikely that the same combination of risk factors will be found in more than one person. Second, even if specific combinations could be identified as complete cause mechanisms, then still its usefulness for the prediction of common disease is limited. When combinations of risk factors are 'unique', only a few other persons in the world may have that exact same profile.

In the field of genetics, the uniqueness of profiles is not surprising as it is the basic rationale for current practice in forensic genetics and paternity testing. Forensic analysts of the Federal Bureau of Investigation (FBI) create DNA profiles (fingerprints) on the basis of a standard set of 13 specific short tandem repeat regions. The odds that two individuals have exactly the same profile are less than 1 in a billion. Potential suspects are identified when their 13-loci DNA profile matches that from the evidence left at the crime scene. If this concept of uniqueness is true for forensic and legal applications, it will also hold in medicine. One obvious exception is the inheritance of genetic profiles within families, but also in this case the probability of sharing the same combination of multiple risk genotypes is likely too small to be useful for disease prediction. For example, if a genetic profile of 12 homozygous risk genotypes constitutes a complete causal mechanism and one parent has this profile, then the probability of inheriting this same profile is 0.02% when the other parent is heterozygous carrier for all variants, and zero percent when the other parent is non-carrier for any variant.

The limited value of prediction of future occurrences based on a specific combination of multiple causal factors is not exclusive to medicine, but generally encountered in the prediction of complex events, such as the prevention of catastrophes and disasters. On March 6, 1987, the British roll-on roll-off car and passenger ferry *Herald of Free Enterprise* left the harbor of Zeebrugge in perfect technical state and in good weather conditions, and capsized less than 100 m out on open sea, killing 193 passengers and crewmembers (30). The ferry had left the harbor with its bow doors open. This fact undoubtedly contributed to the capsizing, but does not fully explain the catastrophe as several earlier successful crossings with open bow doors have been reported. From further investigation, researchers concluded that the ferry capsized because of a unique combination of 13 unfortunate and unlikely causal components, such as the assistant bosun had not closed the bow doors because he fell asleep, the ballast tanks that were filled with water for the car loading had not been emptied because of time pressure, and there were no subdividing bulkheads to secure the cars (31). This combination of causes explains *a posteriori* why the ferry capsized on that day and predicts with almost 100% certainty that any ferry will capsize again in these circumstances. Yet, the occurrence of this specific combination of causal factors is so rare that the combination will unlikely ever occur again in the future. Thus, even if the cause of this disaster is completely understood, the value for the prediction of future capsizing is virtually zero, because the specific combination is expected to be rare and too many other causal factors can contribute

to ferries capsizing. Similarly, given the diversity in causal factors and the random segregation of genetic risk variants, the reoccurrence of specific combinations of causal factors in complex diseases is also expected to be low. Accordingly, we argue that even perfect understanding of causal pathways may not lead to straightforward prediction of complex diseases as is possible for Huntington Disease. Predictive testing of common diseases, whether based on genetic variants only or genetic variants in combination with environmental risk factors, will remain based on statistical models, comparable to what has been applied in the empirical studies that are listed in Tables 1 and 2.

CONCLUSIVE REMARKS

Identification of sufficient cause mechanisms, and hence perfect risk prediction, is relatively straightforward for mono- and oligogenic disorders but immensely complicated for complex diseases. The diversity of the genetic origin of common diseases is too complex to unravel complete causal pathways and to reliably identify risk profiles. Even if sufficient cause genetic profiles could be identified, these profiles will be based on different combinations of multiple variants. As illustrated in this paper by analogy with the Herald of Free Enterprise and DNA identification in forensic genetics, each combination will be extremely rare, even if the variants by themselves are common, and for that reason of limited use for the prediction of common diseases.

Furthermore, we also argue that genetic testing may not improve the prediction of disease beyond classical risk factors or new biomarkers, if most of the genes that are involved in the disease play a role through these risk factors. This problem may concern some disorders more than others. Genetic variants may contribute to higher predictive value when the disease etiology and intermediate risk factors are poorly understood and when intermediate biomarkers cannot be easily assessed.

Does this mean that we advocate a halt to further investments in genomics research of complex diseases? Certainly not. Genomics research will substantially increase our understanding of disease pathogenesis, particularly through the identification of novel disease pathways and new biomarkers. This knowledge will likely lead to novel, more effective and more efficient preventive and therapeutic strategies. However, we argue that such interventions will more likely be made available to individuals who are eligible on the basis of classical risk factors or novel biomarkers rather than on genetic risk profiles. We may encounter situations in which genome-based prediction becomes more or less as good as prediction based on traditional risk factors. Through further reductions in genotyping costs, genetic profiling may become more economically than disease prediction based on traditional risk factors. Potential applications include decisions about preventive or therapeutic interventions, the (starting) doses of pharmacotherapy and the starting age of screening programs as recently proposed by Pharoah *et al*. (32) for breast cancer prevention. These applications have in common that they concern interventions that can be offered to populations at risk, accepting a share of false-positive and false-negative predictions and recognizing that the predictive value of genetic profiling will not be high enough for decisions about invasive, irreversible and expensive interventions.

Before implementation in health care, all applications of genetic profiling need appropriate evaluation to assess whether the predictive value is sufficient e.g. to improve population health or to improve the efficiency or quality of health care. It is clear that prediction studies so far have been rather simplistic in terms that most were based on a small number of variants which by themselves explain only a fraction of the genetic variability. We should not expect accidentally high predictive value from a small set of weak susceptibility genes, as the predictive value is merely a function of the risk of disease, the number of genetic variants, the frequency of their risk genotypes and the strengths of their effects (19). Studies investigating genome-based prediction of common diseases become of real interest when the number of associated variants is substantially larger, including at least tens or even hundreds of weak susceptibility variants among which preferably a few variants with stronger effects (19). Moreover, at this point of knowledge, we still have limited understanding of the exact causal variants, we cannot exclude that the weak effects of common variants on the risk of common disorders are in part explained by linkage disequilibrium of rare variants with major effects, we have limited insight in ethnic differences in gene expression and limited insight in gene–gene and gene–environment interactions. Genetic prediction may be further improved when proper account is taken of these issues.

Furthermore, there are two methodological issues that seriously hamper the validity of the results of many prediction studies. First, prediction studies must be conducted in populations that are representative for the settings in which the genetic testing will be applied. Case–control studies are not the appropriate study design for estimating risks and evaluating predictive value, particularly not when they have recruited hyperselected cases and controls as is often the case in genetic research for demonstrating gene-disease associations. Genetic profiling that is intended for prediction of future disease should be evaluated in large, long-term follow-up studies in which genetic variants are studied together with classical risk factors over time such as The Rotterdam Study, the Framingham Heart Study, the Atherosclerosis Risk in Communities (ARIC) study and the European Prospective Investigation into Cancer (EPIC) study among adults and the Generation R study among newborns (33–37). Second, none of the prediction studies had examined validation of the prediction in independent datasets, whereas this is an integral step in prediction research. The predictive value is generally overestimated when the prediction model is created and evaluated in the same study population, particularly if the same data were first used to select the strongest genetic predictors out of a large set of genotyped variants. Therefore, prediction studies, like genome-wide association studies, should be published with replication in at least one independent cohort.

Based on theoretical grounds, the predictive value of genetic profiling in complex diseases is limited by the heritability and the prevalence of the disease, and expected to be comparable to that of traditional risk factors at best. This level of predictive value may be enough to see some applications of genetic profiling in clinical or preventive health

care, but it will most likely be insufficient to personalize medical interventions at large and to revolutionize health care. Genomics research will lead to major advances in our understanding of the genetic basis of common diseases, but it will not make their etiology less complex.

FUNDING

This study was supported by the Centre for Medical Systems Biology (CMSB) in the framework of the Netherlands Genomics Initiative (NGI). A.C.J.W.J was sponsored by the VIDI grant of the Netherlands Organisation for Scientific Research (NWO).

Conflict of Interest statement. None declared.

REFERENCES

1. Manolio, T.A., Brooks, L.D. and Collins, F.S. (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.*, **118**, 1590–1605.
2. Guttmacher, A.E. and Collins, F.S. (2005) Realizing the promise of genomics in biomedical research. *JAMA*, **294**, 1399–1402.
3. Brand, A., Brand, H. and Schulte Schulte in den Bäumen, T. (2008) The impact of genetics and genomics on public health. *Eur. J. Hum. Genet.*, **16**, 5–13.
4. Khoury, M.J., Gwinn, M., Yoon, P.W., Dowling, N., Moore, C.A. and Bradley, L. (2007) 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.*, **9**, 665–674.
5. Haga, S.B., Khoury, M.J. and Burke, W. (2003) Genomic profiling to promote a healthy lifestyle: not ready for prime time. *Nat. Genet.*, **34**, 347–350.
6. Holtzman, N.A. and Marteau, T.M. (2000) Will genetics revolutionize medicine? *N. Engl. J. Med.*, **343**, 141–144.
7. Vineis, P., Schulte, P. and McMichael, A.J. (2001) Misconceptions about the use of genetic tests in populations. *Lancet*, **357**, 709–712.
8. Maller, J., George, S., Purcell, S., Fagerness, J., Altshuler, D., Daly, M.J. and Seddon, J.M. (2006) Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat. Genet.*, **38**, 1055–1059.
9. Humphries, S.E., Cooper, J.A., Talmud, P.J. and Miller, G.J. (2007) Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men. *Clin. Chem.*, **53**, 8–16.
10. van der Net, J.B., van Etten, J., Yazdanpanah, M., linga-Thie, G.M., Kastelein, J.J., Defesche, J.C., Koopmans, R.P., Steyerberg, E.W. and Sijbrands, E.J. (2008) Gene-load score of the rennin–angiotensin–aldosterone system is associated with coronary heart disease in familial hypercholesterolaemia. *Eur. Heart J.*, **29**, 1370–1376.
11. Wang, J., Ban, M.R., Zou, G.Y., Cao, H., Lin, T., Kennedy, B.A., Anand, S., Yusuf, S., Huff, M.W., Pollex, R.L. and Hegele, R.A. (2008) Polygenic determinants of severe hypertriglyceridemia. *Hum. Mol. Genet.*, July 1, Epub ahead of print.
12. Podgoreanu, M.V., White, W.D., Morris, R.W., Mathew, J.P., Stafford-Smith, M., Welsby, I.J., Grocott, H.P., Milano, C.A., Newman, M.F. and Schwinn, D.A. Perioperative Genetics and Safety Outcomes Study (PEGASUS) Investigative Team. (2006) Inflammatory gene polymorphisms and risk of postoperative myocardial infarction after cardiac surgery. *Circulation*, **114**, I-275.
13. Harley, J.B., Arcon-Riquelme, M.E., Criswell, L.A., Jacob, C.O., Kimberly, R.P., Moser, K.L., Tsao, B.P., Vyse, T.J., Langefeld, C.D., Nath, S.K. *et al.* (2008) Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat. Genet.*, **40**, 204–210.
14. Weedon, M.N., McCarthy, M.I., Hitman, G., Walker, M., Groves, C.J., Zeggini, E., Rayner, N.W., Shields, B., Owen, K.R., Hattersley, A.T. and Frayling, T.M. (2006) Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLOS Med.*, **3**, e374.
15. Vaxillaire, M., Veslot, J., Dina, C., Proenca, C., Cauchi, S., Charpentier, G., Tichet, J., Fumeron, F., Marre, M., Meyre, D. *et al.* (2008) Impact of common type 2 diabetes risk polymorphisms in the DESIR prospective study. *Diabetes*, **57**, 244–254.
16. Lango, H., Palmer, C.N., Morris, A.D., Zeggini, E., Hattersley, A.T., McCarthy, M.I., Frayling, T.M. and Weedon, M.N. (2008) Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes*, June 30, Epub ahead of print.
17. van Hoek, M., Dehgan, A., Witteman, J.C.M., Van Duijn, C.M., Uitterlinden, A.G., Oostra, B.A., Hofman, A., Sijbrands, E.J. and Janssens, A.C.J.W. (2008) Predicting type 2 diabetes based on polymorphisms from genome wide association studies: a population-based study. *Diabetes*, August, 11, Epub ahead of print.
18. Despriet, D.D.G., Klaver, C.C., van Duijn, C.M. and Janssens, A.C.J.W. (2007) Predictive value of multiple genetic testing for age-related macular degeneration. *Arch. Ophthalmol.*, **125**, 1270–1271.
19. Janssens, A.C.J.W., Aulchenko, Y.S., Elefante, S., Borsboom, G.J.J.M., Steyerberg, E.W. and Van Duijn, C.M. (2006) Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet. Med.*, **8**, 395–400.
20. Yang, Q., Khoury, M.J., Botto, L., Friedman, J.M. and Flanders, W.D. (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *Am. J. Hum. Genet.*, **72**, 636–649.
21. Janssens, A.C.J.W., Moonesinghe, R., Yang, Q., Steyerberg, E.W., Van Duijn, C.M. and Khoury, M.J. (2007) The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet. Med.*, **9**, 528–535.
22. Yang, Q., Khoury, M.J., Friedman, J.M., Little, J. and Flanders, W.D. (2005) How many genes underlie the occurrence of common complex diseases in the population? *Int. J. Epidemiol.*, **34**, 1129–1137.
23. Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.*, **17**, 1520–1528.
24. Lyssenko, V., Almgren, P., Anevski, D., Orho-Melander, M., Sjogren, M., Saloranta, C., Tuomi, T. and Groop, L.the Botnia Study Group. (2005) Genetic prediction of future type 2 Diabetes. *PLOS Med.*, **2**, e345.
25. Morrison, A.C., Bare, L.A., Chambless, L.E., Ellis, S.G., Malloy, M., Kane, J.P., Pankow, J.S., Devlin, J.J., Willerson, J.T. and Boerwinkle, E. (2007) Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *Am. J. Epidemiol.*, **166**, 28–35.
26. Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N.P., Roos, C., Hirschhorn, J.N., Berglund, G., Hedblad, B., Groop, L. *et al.* (2008) Polymorphisms associated with cholesterol and risk of cardiovascular events. *N. Engl. J. Med.*, **358**, 1240–1249.
27. Zheng, S.L., Sun, J., Wiklund, F., Smith, S., Stattin, P., Li, G., Adami, H.O., Hsu, F.C., Zhu, Y., Balter, K. *et al.* (2008) Cumulative association of five genetic variants with prostate cancer. *N. Engl. J. Med.*, **358**, 910–919.
28. Thorgeirsson, T.E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K.P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A. *et al.* (2008) A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, **452**, 638–642.
29. Rothman, K.J. and Greenland, S. (2005) Causation and causal inference in epidemiology. *Am. J. Public Health*, **95**, S144–S150.
30. Goulielmos, A.M. and Goulielmos, M.A. (2005) The accident of m/v *Herald of Free Enterprise*. *Disaster Prev. Manag.*, **14**, 479–492.
31. Braband, J., Evers, B. and de Stefano, E. (2003) Towards a hybrid approach for incident root cause analysis. *Proceedings of the 21st International System Safety Conference*, Ottawa, Ontario, Canada, pp. 203–211.
32. Pharoah, P.D., Antoniou, A.C., Easton, D.F. and Ponder, B.A. (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. *N. Engl. J. Med.*, **358**, 2796–2803.
33. Hofman, A., Breteler, M.M., Van Duijn, C.M., Krestin, G.P., Pols, H.A., Stricker, B.H., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R. and Witteman, J.C. (2007) The Rotterdam Study: objectives and design update. *Eur. J. Epidemiol.*, **22**, 819–829.
34. The ARIC investigators. (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am. J. Epidemiol.*, **129**, 687–702.

35. Dawber, T., Meadors, G. and Moore, F. (1951) Epidemiological approaches to heart disease: the Framingham Study. *Am. J. Public Health*, **41**, 279–281.
36. Riboli, E. and Kaaks, R. (1997) The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.*, **26** (Suppl. 1), S6–S14.
37. Jaddoe, V.W., Bakker, R., Van Duijn, C.M., van der Heijden, A.J., Lindemans, J., Mackenbach, J.P., Moll, H.A., Steegers, E.A., Tiemeier, H., Uitterlinden, A.G. *et al.* (2007) The Generation R Study Biobank: a resource for epidemiological studies in children and their parents. *Eur. J. Epidemiol.*, **22**, 917–923.
38. De Silva, N.M.G., Steele, A., Shields, B., Knight, B., Parnell, K., Weedon, M.N., Hattersley, A.T. and Frayling, T.M. (2007) The transcription factor 7-like 2 (TCF7L2) gene is associated with Type 2 diabetes in UK community-based cases, but the risk allele frequency is reduced compared with UK cases selected for genetic studies. *Diabet. Med.*, **24**, 1067–1072.
39. Narayan, K.M., Boyle, J.P., Thompson, T.J., Sorensen, S.W. and Williamson, D.F. (2003) Lifetime risk for diabetes mellitus in the United States. *JAMA*, **290**, 1884–1890.
40. Janssens, A.C.J.W. and Khoury, M.J. (2006) Predictive value of testing for multiple genetic variants in multifactorial diseases: implications for the discourse on ethical, legal and social issues. *Ital. J. Public Health*, **3**, 35–41.
41. Janssens, A.C.J.W., Gwinn, M., Subramonia-Iyer, S. and Khoury, M.J. (2006) Does genetic testing really improve the prediction of future type 2 diabetes? *PLOS Med.*, **3**, e114.