Risk Prediction of Complex Diseases from Family History and Known Susceptibility Loci, with Applications for Cancer Screening

Hon-Cheong So,[1] Johnny S.H. Kwan,[1] Stacey S. Cherny,[1,2,3] and Pak C. Sham[1,2,3,*]

Risk prediction based on genomic profiles has raised a lot of attention recently. However, family history is usually ignored in genetic risk prediction. In this study we proposed a statistical framework for risk prediction given an individual's genotype profile and family history. Genotype information about the relatives can also be incorporated. We allow risk prediction given the current age and follow-up period and consider competing risks of mortality. The framework allows easy extension to *any* family size and structure. In addition, the predicted risk at any percentile and the risk distribution graphs can be computed analytically. We applied the method to risk prediction for breast and prostate cancers by using known susceptibility loci from genome-wide association studies. For breast cancer, in the population the 10-year risk at age 50 ranged from 1.1% at the 5th percentile to 4.7% at the 95th percentile. If we consider the average 10-year risk at age 50 (2.39%) as the threshold for screening, the screening age ranged from 62 at the 20th percentile to 38 at the 95th percentile (and some never reach the threshold). For women with one affected first-degree relative, the 10-year risks ranged from 2.6% (at the 5th percentile) to 8.1% (at the 95th percentile). For prostate cancer, the corresponding 10-year risks at age 60 varied from 1.8% to 14.9% in the population and from 4.2% to 23.2% in those with an affected first-degree relative. We suggest that for some diseases genetic testing that incorporates family history can stratify people into diverse risk categories and might be useful in targeted prevention and screening.

Introduction

For complex diseases with substantial heritability, the prediction of disease risk has been largely based on family history information. However, because of the recent progress in genome-wide association studies (GWAS), an increasing number of susceptibility variants for complex diseases are being identified.[1,2] The genotyping of these variants is expected to contribute to risk prediction.

Genetic risk prediction has raised a lot of attention recently, and many companies (e.g., 23andMe, deCODEme, Navigenics) are already offering such services. However, family history is ignored in genetic risk prediction by these companies. Family history can often (though not always) be obtained easily and at no extra cost. However, family history is unable to stratify people into fine risk categories because in most cases people either have a negative family history or just have one or two close relatives with the disease. On the other hand, the number of genotype combinations from SNPs is large. For example, 5, 10, and 20 biallelic loci will give rise to 243 (3^5), 59,049, and 3,486,784,401 combinations respectively.

In addition, people with a positive family history have a greater incentive to request genetic testing. Because they have a higher baseline risk, the range of predicted risks for their given genotype profile is usually larger than that of other individuals without a family history of the disease (this will be illustrated in later examples). Therefore, a risk prediction algorithm that takes into account both family history and an individual's genotype is clearly clinically relevant.

Here, we present a statistical framework for integrating family history and known susceptibility variants into risk prediction; genotype information of relatives can also be incorporated into the model. In addition, to account for competing risks of mortality, we allow risk prediction given the current age and a period of follow up. The framework we propose is very flexible and allows easy extension to *any* family size and structure. One can readily extend the model to handle markers in linkage disequilibrium (LD), haplotypes, or multilocus genotypes (when interactions are present). In addition, the predicted risk at any percentile and the risk distribution graphs can be computed analytically without simulations.

Our risk prediction framework is based on the liability threshold model, which assumes a latent continuous liability underlying each disease.[3] The use of the liability threshold model in risk prediction has more than 30 years of history.[4–6] For example, Mandell and Elston[6] previously proposed the use of the Pearson-Aitken (PA) formula in estimating recurrence risks in relatives, and the same formula is employed in our model. Our proposed framework can thus be regarded an extension of these previous works.

We apply the framework to breast cancer (MIM 114480) and prostate cancer (MIM 176807) and discuss the possible impact of genetic testing on screening based on family history and known susceptibility variants. We computed the lifetime risk and 10-year risk of breast and prostate cancer at age 50 or 60 for people at different risk percentiles. From these data, we estimated the age at which screening should start. We found that the screening age,

[1]Department of Psychiatry, University of Hong Kong, Hong Kong SAR, China; [2]Genome Research Centre, University of Hong Kong, Hong Kong SAR, China; [3]State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong SAR, China
*Correspondence: pcsham@hkucc.hku.hk

DOI 10.1016/j.ajhg.2011.04.001. ©2011 by The American Society of Human Genetics. All rights reserved.

the absolute disease risks, and the number of individuals that need to be screened in order to prevent one cancer death (NNS) all vary considerably for different risk percentiles, implying that genetic tests with family history might be useful in risk stratification and targeted prevention of some cancers.

Material and Methods

Model Assumptions

Underlying the disease is a latent construct called liability (L), which is determined by multiple genetic and environmental factors and is normally distributed in the population with a mean of 0 and a variance of 1; each disease-associated locus explains a certain proportion of variance (Vg), and the aggregate-over all disease-associated loci is defined as the heritability of the disease. There is a threshold value (T) in liability, which when exceeded leads to the development of disease.

The total liability L can be partitioned into two components, a measurable liability (M), produced by known susceptibility variants, and another component (U), produced by genetic variants yet-to-be discovered as well as environmental factors. That is $L = M + U$. The measurable liability, M, has a mean of 0 and a variance of V in the population, so that it explains a proportion V of the variance in the total liability L (which has a variance of 1).The value V is equal to the sum of the variance in liability explained (or heritability explained) by *individual* known loci. The method of calculating Vg by a single locus is described in So et al.[7] Note that our model can also be generalized to take account of known measurable environmental risk factors (e.g., smoking status, lipid levels, etc.), but for simplicity of exposition, we assume here that environmental factors are not measured. The appendices include a brief discussion on the incorporation of environmental risk factors into the prediction model.

In this study we also assume that the liability refers to a *lifetime diagnosis of the disease*. Hence, the threshold value (T) is determined by the lifetime risk of disease in the population (denoted by K). In notations $T = \Phi^{-1}(1-K)$, where Φ^{-1} is the inverse normal distribution function. Also, the variance in liability explained by known susceptibility variants is evaluated based on *lifetime* risk and *lifetime* relative risk estimates as inputs. The risks calculated are therefore *lifetime* disease risks. However, the method can be generalized to take into account the current age of the individual and predict risk in a specified period of time.

Construction of Measured Liability Score

We describe how to calculate the variance in liability explained and the liability score for each genotype elsewhere.[7] The method allows any odds ratios for genotypes.

If we assume a multiplicative model of allelic effects (e.g., if Aa has an odds ratio (OR) of 1.3 over aa, then AA will have an OR of $1.3^2 = 1.69$), the liability score for the ith locus can also be approximated by $\sqrt{V_i}$ (square root of the Vg by that locus) multiplied by the standardized genotype count (G_i). The total liability score is then given by the sum of individual scores, $\sum_i \sqrt{V_i} G_i$.

Prediction of Disease Risk from Measured Liability Score

The proportion of variance in liability explained by known genetic variants, V, is directly related to our ability to predict the risk of disease in an individual from genotype information. Because $\mathrm{Cov}(L,M) = \mathrm{Var}(M) = V$, we have from standard regression theory, $\mathrm{E}(L|M=m) = m$ and $\mathrm{Var}(L|M=m) = 1 - V$. Hence, $\Pr(L > T|M=m)$, or the absolute risk of disease, can be obtained from the standard normal distribution function once we know the liability score (m) from the confirmed genetic variants.

Risk Prediction with Consideration of Family History

Modeling Affected Relatives

Here, we examine risk prediction by using known susceptibility variants for individuals with a positive family history of disease. For simplicity, we first consider the case of an individual with one affected first-degree relative. The PA formula can be applied to model the joint impact of family history and measured susceptibility variants on disease risks. The PA formula describes how the mean vector and covariance matrix of set of variables are distorted by selection on a subset of the variables. In addition to the PA formula, it is also possible to use multivariate integration for modeling the effects of selection. The PA formula, however, is simpler to implement (because numerical integration programs are not needed) and enables explicit formulas to be given for the results. We will focus on the use of PA formula, and details of the alternative approach with multivariate integration are detailed in the Appendix A.

Consider three liability distributions including (1) the overall liability of the individual's First-degree relative (L_{rel}), (2) the measurable liability of the individual (M_{ind}), and (3) the overall liability of the individual (L_{ind}). The mean vector of the above three liability distributions $[L_{rel}, M_{ind}, L_{ind}]^{\mathrm{T}}$ before selection is simply $[0,0,0]^{\mathrm{T}}$. The covariance matrix is

$$\Sigma = \begin{bmatrix} 1 & V/2 & h/2 \\ V/2 & V & V \\ h/2 & V & 1 \end{bmatrix},$$

where the total Vg by the known susceptibility variants is denoted by V and the heritability denoted by h. For simplicity of exposition, we shall assume in this paper that heritability refers to the narrow-sense (additive) heritability only. Nonadditive effects are assumed to be negligible. One can relax this assumption by including dominant effects as well, but the covariance of parent-child and siblings will then be different.

The selection for affected first-degree relatives changes the distribution of L_{rel} from a standard normal to a truncated normal distribution in which the truncation point equals T. From theories of truncated normal distribution, the new mean and variance of L_{rel} are

$$E(L_{rel,new}) = \frac{\phi(T)}{1-\Phi(T)} = a, \quad \mathrm{var}\,(L_{rel,new}) = 1 - a^2 + aT = b.$$

One can then apply PA formula as described above to calculate the mean and covariance matrix of M_{ind} and L_{ind} after selection on L_{rel}. Then we perform the second selection based on the known measurable liability of the individual, denoted by m_I. The PA formula is applied again to give the new mean and variance of the individual's overall liability L_{ind}. The results are given below.

Mean of L_{ind} after selection ($\mu_{ind,fam}$) (first-degree relative affected)

$$= \frac{ah}{2} + \left[V - \frac{(1-b)Vh}{4}\right]\left[\frac{1}{V-(1-b)\left(V^2/4\right)}\right]\left(m_I - \frac{aV}{2}\right)$$

(Equation 1)

Variance of L_{ind} after selection ($\sigma^2_{ind,fam}$) (first-degree relative affected)

$$= 1 - (1-b)\left(\frac{h^2}{4}\right) - \left[V - \frac{(1-b)Vh}{4}\right]^2 \left[\frac{1}{V - (1-b)(V^2/4)}\right]$$

(Equation 2)

The absolute risk of disease for the individual is the probability that the overall liability (a normally distributed variable with the above mean and variance) exceeds the threshold, T.

$$\Pr(Disease) = 1 - \Phi\left(\frac{T - \mu_{ind,\,fam}}{\sqrt{\sigma^2_{ind,\,fam}}}\right).$$

Using the same approach, we can derive the risk estimate of an individual who has an affected *second-degree relative*. It can be shown that L_{ind}, given family history in which a second-degree relative is affected and the individual's measured liability, has the following mean and variance:

The mean of L_{ind} after selection (second-degree relative affected) is

$$\frac{ah}{4} + \left[V - \frac{(1-b)Vh}{16}\right]\left[\frac{1}{V - (1-b)(V^2/16)}\right]\left(m_I - \frac{aV}{4}\right).$$

The variance of L_{ind} after selection (second-degree relative affected) is

$$= 1 - (1-b)\left(\frac{h^2}{16}\right) - \left[V - \frac{(1-b)Vh}{16}\right]^2 \left[\frac{1}{V - (1-b)(V^2/16)}\right].$$

Extension to Arbitrary Family Structure

The above methods can be readily extended to arbitrary family structures and any number of affected and unaffected relatives. The covariance structure of the liabilities needs to be adjusted according to the family structure. Further examples of risk prediction with the PA formula will be given in later sections.

Taking into Account Current Age and Predicting Risk in a Specified Period of Time

This is an extension of our previous work on age-conditional risk prediction without considering family history.[8] Note that we assume proportional hazards, that is the hazard ratio of disease due to positive family history, is constant regardless of the subject's age.

The absolute disease risk for an individual with a current age of a in the next s years is

$$p(s,a,R) = R\int_a^{a+s} \lambda(b)\,\exp\left[-\int_a^b (R\lambda(x) + \mu(x))dx\right]db$$

(Equation 3)

The PA formula enables us to estimate the lifetime risk of an individual given his or her family history and known genetic factors. This gives $p(s,\ a,\ R)$. Because the disease incidence function ($\lambda(x)$, where x is the age) and the net mortality function ($\mu(x)$) are known, the only unknown is the hazard ratio R. The above equation can be solved numerically for R; then one can substitute any value of starting age (a) and period of follow up (s) into the formula to obtain the relevant risk estimates.

Risk Distribution in Individuals with an Affected First-Degree Relative, Given a Set of Known Susceptibility Variants

Suppose we predict an individual's risk based on two pieces of information: (1) a set of known genes and (2) whether the person has an affected first-degree relative. We ignore more complex family histories for simplicity. What is the distribution (i.e., the probability density function [pdf]) of predicted risks in those with an affected first-degree relative?

We have described before the mean of overall liability after selection by positive family history (see Equations 1 and 2). The predicted absolute risk (r) is

$$\mathrm{r} = 1 - \Phi\left(\frac{T - \mu_{ind,\,fam}}{\sqrt{\sigma^2_{ind,\,fam}}}\right).$$

The variable m_I in Equation 1 is equivalent to z, which will be used below to represent the measurable liability. For notational simplicity, we define

$$w = ah/2 \;\; q = \left[V - \frac{(1-b)Vh}{4}\right]\left[\frac{1}{V - (1-b)(V^2/4)}\right], \;\; s = aV/2,$$

$$\mathrm{r} = 1 - \Phi\left(\frac{T - [w + q(z-s)]}{\sqrt{\sigma^2_{ind,\,fam}}}\right),$$

and

$$\frac{d\mathrm{r}}{dz} = -\phi\left(\frac{T - [w + q(z-s)]}{\sqrt{\sigma^2_{ind,\,fam}}}\right)\left(-\frac{q}{\sqrt{\sigma^2_{ind,\,fam}}}\right);$$

z can be expressed as

$$z = \Phi^{-1}(p)\sqrt{v_{rel,\,aff}} + \mu_{rel,\,aff},$$

where $v_{rel,aff}$ and $\mu_{rel,aff}$ are the variance and mean of measurable liability, respectively, for an individual (M_{ind}) who has an affected first-degree relative. p is the percentile of the measurable liability given an affected first-degree relative. We have $v_{rel,aff} = V - (1-b)(V^2/4)$ and $\mu_{rel,aff} = aV/2$.

Note that

$$\frac{dz}{dp} = \frac{\sqrt{v_{rel,\,aff}}}{\phi(\Phi^{-1}(p))}$$

Thus giving

$$\frac{dp}{d\mathrm{r}} = 1/\left(\frac{d\mathrm{r}}{dz} \bullet \frac{dz}{dp}\right).$$

$dp/d\mathrm{r}$ is the derivative of the cumulative density function (cdf) of predicted risks and is equal to the pdf. The method can be extended to deal with any family structure and any number of affected relatives. Knowing the predicted risk distribution has practical implications. For example, the distribution allows us to assess the proportion of people exceeding certain risk thresholds for screening or interventions.

Risk Prediction with Consideration of Measurable Liability of Relatives

We have described how to predict disease risk for an individual with positive family history and have a set of susceptibility variants genotyped. The methodology can be extended naturally to incorporate the measurable liability of relatives. For example, some of the relatives might also have been genotyped on a set of risk genes, and this information can be used to further improve risk prediction for the individual.

We will illustrate with an example how the measurable liabilities of relatives can be incorporated into the risk prediction algorithm. Suppose a first-degree relative is affected with the disease. Both the relative and the individual are tested on the same set of known susceptibility variants. To predict the individual's risk, we have to consider four liability distributions: (1)L_{rel}, the overall liability of the relative; (2)M_{rel}, the measurable liability of the relative; (3)M_{ind}, the measurable liability of the individual; and (4)L_{ind}, the overall liability of the individual. The vector (L_{rel}, M_{rel}, M_{ind}, L_{ind}) has the following mean and covariance matrix:

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & V & V/2 & h/2 \\ V & V & V/2 & V/2 \\ V/2 & V/2 & V & V \\ h/2 & V/2 & V & 1 \end{bmatrix}.$$

The PA formula was applied iteratively for the following three levels of selection: (1) the relative being affected ($L_{rel} > T$), (2) the known measurable liability of the relative (M_{rel}), and (3) the known measurable liability of the individual (M_{ind}). The details of the calculation are given in the Appendix A.

After the selections, the mean and variance of the *individual's overall liability* are

$$\tilde{\mu}_{final} = \frac{ah}{2} - \frac{aV[1-(1-b)h]}{2[1-(1-b)V]} + \frac{1}{2}\left(\frac{1-(1-b)h}{1-(1-b)V} - 1\right)m_R + m_I$$

and

$$\tilde{V}_{final} = 1 - \frac{(1-b)h^2}{4} - \frac{[V-(1-b)Vh]^2}{4\left[V-(1-b)V^2\right]} - \frac{3V}{4},$$

where m_R and m_I are the measurable liabilities of the relative and the individual, respectively. The absolute risk of disease for the individual is the probability that the overall liability (a normally distributed variable with the above mean and variance) exceeds the threshold, *T*.

The individual's and the relative's measurable liabilities have different impacts on the final predicted risk, which can be deduced from the above results. It is obvious that if an individual has a high *measurable* liability, he or she will also have a high *overall* liability and higher disease risk. But what about the measurable liability of the affected relative? The question will be resolved by the following derivations.

Because the Vg by genetic variants cannot exceed the total heritability ($h \geq V$) and b is the variance of the truncated standard normal distribution, $b < 1$, then

$$\begin{gathered}(1-b)h \geq (1-b)V \\ 1-(1-b))h \leq 1-(1-b)V \\ \frac{1}{2}\left(\frac{1-(1-b)h}{1-(1-b)V} - 1\right) \leq 0\end{gathered}.$$

Because the above factor that accompanies m_R is always smaller than or equal to 0, a higher measurable liability of the affected relative leads to a lower predicted risk for the individual.

This phenomenon can also be explained intuitively. If it is known that the relative harbors a lot of risk variants, then on average the rest of the variants that are passed onto his or her children or other relatives should be more protective. (The measured risk genes already raise the disease risk to a sufficiently high level.) On the other hand, if it is known that measured genotypes are mostly protective, but the relative is still affected, then on average the ungenotyped variants are likely to contain more risk variants.

Here, we also consider the case when a first-degree relative is affected and his or her measurable liability is known. The individual's measurable liability is, however, *not* known. Conditioned on these factors, the mean and variance of the index individual's overall liability are

$$\tilde{\mu} = \frac{ah}{2} + \frac{V-(1-b)Vh}{2\left[V-(1-b)V^2\right]}(m_R - aV)$$

and

$$\tilde{V} = 1 - \frac{(1-b)h^2}{4} - \frac{[V-(1-b)Vh]^2}{4\left[V-(1-b)V^2\right]},$$

respectively. The derivation is very similar to above. In this case, the higher the relative's measurable liability, the bigger the risk for the index individual (see the Appendix A for the proof).

Dealing with Continuous Risk Factors

Our original motivation for evaluating the Vg was to quantify the contribution of individual genetic variants to heritability. However, the same concept can also be extended to continuous risk factors. This extension would provide us with a unifying statistical framework for risk prediction such that our proposed method could deal with all types of risk factors, categorical or continuous. The details are given in the Appendix A. Briefly, we can consider that the continuous risk factor is composed of a very large number of categories, as opposed to only three categories for a biallelic marker.

Application to Real Examples: Breast and Prostate Cancer Risk Estimates and Implications for Screening

We illustrate the above methodologies by applying them to breast and prostate cancers. The odds ratio estimates for the two diseases were extracted mainly from the GWAS catalog maintained by the National Human Genome Research Institute (NHGRI),[2] complemented by a manual PubMed search. We included only genetic variants passing the genome-wide significance level (7.2 × 10^{-8}).[9] The details of the included susceptibility variants are described in another manuscript.[7] Briefly, a total of 13 SNPs were included for breast cancer and 30 were included for prostate cancer. The details are given in Tables S1 and S2 (available online). Incidence and mortality data (for the years 2004–2006) were extracted from the Surveillance, Epidemiology and End Results (SEER) database, established by the National Cancer Institute in the United States. Risk estimates were based on a white population from the U.S., and only female breast cancers were considered.

The procedures for risk estimation were as follows. We first converted the reported odds ratios to lifetime relative risks, taking into account competing risks. On the basis of the lifetime risks and the lifetime RR, the total Vg was computed. This is also equal to the variance of the measurable liability. Then we applied the PA formula to handle the covariance between the relevant liability distributions. In fact, the formulas for the more typical cases (e.g., one first-degree relative is affected) have been derived. The lifetime risks for people at different percentiles were computed. Besides the lifetime risk, we also calculated the 10-year absolute risk for women whose current age was 50, for each percentile. To do so, we deduced the incidence rate ratio (i.e., the hazard ratio) from the lifetime risk estimates according to Equation 3. The age-specific absolute risk for any follow-up periods could then be obtained. We performed the same risk-estimation procedure for women with one affected first-degree relative whose genotype profile (i.e., measurable liability) was also known. The algorithm is summarized in Table 1.

Next, we assessed the implications of genetic risk prediction for screening and chemoprevention, inspired by the analysis in Pharoah et al.[10] Breast cancer screening by mammography has been shown to provide benefits.[11] The US Preventive Services

Table 1. Algorithm for Predicting Lifetime Risk and Age-Conditional Risk within a Certain Time Interval with Family History Information

Step	Procedure
1	Use the competing risk formula to calculate the lifetime RR for the genotypes in each variant (i.e., calculate the lifetime risk for each genotype by using POR, then divide the lifetime risk of each group by that of a baseline group to obtain the lifetime RR).[a]
2	Compute the Vg of all variants; Vg is calculated with overall lifetime risk and the allele frequency and lifetime RR of genotypes as inputs. This forms the measurable liability.
3	Use the PA formula to take into account covariance between total and measurable liability, and obtain the predicted lifetime risk.
4	Deduce the R for each year from the overall lifetime absolute risk by solving the competing risk equation (Equation 3).
5	Knowing R, we can estimate the risk by using any start age and specifying any time period.

The following abbreviations are used: RR, relative risk; POR, prevalence odds ratio; Vg, variance explained; R, the hazard ratio; PA, Pearson-Aitken. Please refer to So and Sham[8] for more details concerning the calculation of lifetime RR. Note that both the heritability and measurable liability refers to liability to a lifetime diagnosis of the disease.

[a] Assumes the genetic loci to have constant effects on liability and hazard ratios regardless of age.

Task Force (USPSTF) recommends biennial mammography screening for women between 50 and 74 years old.[11] The UK National Health Service also offers screening to women above age 50. We calculated the average 10-year risk for a 50-year-old women to develop breast cancer as an approximate risk threshold for mammography screening, as in Pharoah et al.[10] The 10-year risk is 2.39%, and we evaluated the age at which women at different risk percentiles would reach this risk level. Adding to the work of Pharoah et al.,[10] we have updated the number of loci to 13 and described in detail the methodology to compute the absolute risks, accounting for competing causes of mortality (the methods for deriving the absolute risks were not detailed by Pharoah et al.[10]). Importantly, we extended the previous work to incorporate family history and the relatives' genetic profile in risk prediction. The NNS to prevent one cancer death was estimated. Unlike Pharoah et al.[10], we provided a general analytic approach to perform the calculations without the need to consider all genotype combinations. The method was also applied to prostate cancer. It is also much easier to evaluate the risks at different percentiles with our approach, because genotype combinations are conceptualized as the measurable liability.

Here, we briefly describe the concept of the NNS to prevent one cancer death.[12] NNS is an extension of the idea of number needed to treat. It is equal to 1 divided by the *absolute* reduction in mortality rate from the disease, if one compares the screening with the no-screening group. Following our previous calculations, we considered the NNS of a group of 50-year-old women with 10-year follow up. Instead of estimating the age at which one should start screening, we changed the angle of view and considered a group of women who start screening at the same age (50) and calculate how many women need to be screened such that one cancer death can be avoided. The rate ratio for death from breast cancer in the screening group is taken as 0.86.[13] In other words, when compared to women who do not receive any screening, the mortality rate from breast cancer is 14% lower in the screening group. This is, however, only the proportional decrease in mortality. The *absolute* mortality reduction is equal to the actual 10-year mortality rate in the unscreened group multiplied by 14%. One divided by this absolute mortality reduction is taken as the NNS in the general population (we assume the 50th percentile represents the general population).

Note that the NNS is lower if the disease is more common, for example in high-risk groups. We assumed that the susceptibility variants and family history do not affect the mortality and the mortality rate ratio is constant. As a result, if the condition is ten times more common in a subgroup, the *absolute* mortality reduction is also ten times higher and the NNS will be 1/10 of the control group. Using this principle, we computed the NNS at different risk levels as derived from the genotype profile and family history.

It has been shown that the risk of invasive breast cancer could be reduced by the drugs tamoxifen and reloxifene.[13] The Food and Drug Administration has approved the use of these drugs for high-risk women, including those with a five-year cancer risk of 1.66% or more.[13] Hence, we also considered this threshold and estimated the age at which this threshold will be passed.

A similar set of analyses was also performed on prostate cancer. However, for prostate cancer, the benefit of screening is less clear. Prostate-specific antigen (PSA) testing and digital rectal examination are commonly employed for screening. The USPSTF judged that there is insufficient evidence to assess the balance of benefits and harms of screening for men below age 75. Screening is not recommended above age 75.[14,15] The American Cancer Society recommended that asymptomatic men with at least 10-year life expectancy make an informed decision with their health care provider as to whether they should receive screening for prostate cancer after discussing the risks and benefits of screening with their health care provider. Average-risk men should receive such information starting at age 50.[16]

Similar to our calculations for breast cancer risk, we estimated the average 10-year risk of prostate cancer at age 50 as the risk threshold for screening. We repeated the procedures for risk estimation used in the breast cancer example, except that the 10-year risk at age 60 was also computed because of the later onset of the disease. The NNS was computed based on the consideration of 60-year-old men with 10 years of follow up. The rate ratio for death from prostate cancer in those who are screened was taken to be 0.80.[17]

Construction of Confidence Intervals

To provide a sense of the uncertainty of risk estimates, we constructed the confidence intervals of the predicted risks by a simulation approach. The standard deviation of log odds ratio (ln OR) for each SNP is derived from its confidence interval, and ln OR is assumed to follow a normal distribution. A random ln OR is simulated for every SNP in each run, and the entire procedure for risk estimation is repeated. A total of 1,000 simulations are performed.

Results

Table 2 shows the absolute risks of breast cancer at different risk percentiles in the general population. The total Vg by the known genetic variants is 5.70%. The predicted lifetime risk ranged from 5.7% at the 5th percentile to 22.1% in the 95th percentile. The range of 10-year risks at age 50 is smaller. Women at the 5th percentile had a risk of 1.1%, whereas those at the 95th percentile had a

Table 2. Predicted Risk of Breast Cancer in the General Population from Genetic Profiles

Percentile	Lifetime Risk (95% CI)	Incidence Rate Ratio (95% CI)	10-Year-Risk at Age 50 (95% CI)	Age at which 10-Year Risk > 2.39%	Age at which 5-Year Risk > 1.66%	NNS
5	0.057 (0.049–0.061)	0.46 (0.39–0.49)	0.011 (0.009–0.012)	NA	NA	3913
10	0.068 (0.061–0.072)	0.55 (0.49–0.58)	0.013 (0.012-0.014)	NA	NA	3265
20	0.084 (0.077–0.087)	0.68 (0.63–0.71)	0.016 (0.015–0.017)	61.7	NA	2643
30	0.096 (0.091–0.099)	0.79 (0.75–0.81)	0.019 (0.018–0.019)	55.9	NA	2281
40	0.108 (0.104–0.110)	0.90 (0.86–0.91)	0.021 (0.021–0.022)	52.6	65.5	2018
50	0.120 (0.118–0.121)	1.00 (0.98–1.01)	0.024 (0.024–0.024)	49.9	60.3	1805
60	0.133 (0.133–0.133)	1.12 (1.12–1.12)	0.027 (0.027–0.027)	47.1	57.4	1619
70	0.148 (0.147–0.150)	1.26 (1.25–1.28)	0.030 (0.030–0.030)	44.5	54.8	1445
80	0.167 (0.164–0.172)	1.44 (1.41–1.49)	0.034 (0.034–0.035)	42.1	51.6	1269
90	0.195 (0.190–0.206)	1.71 (1.67–1.82)	0.041 (0.039–0.043)	39.6	47.1	1067
95	0.221 (0.214–0.236)	1.97 (1.90–2.14)	0.047 (0.045–0.050)	38.0	44.4	930

The percentile column refers to the percentile of the measurable liability derived from known susceptibility variants. The following abbreviations are used: CI, confidence interval; NNS, number of individuals that need to be screened in order to prevent one cancer death; NA, not applicable. Because of competing risks of mortality, the absolute risk will not reach the designated threshold in these groups of women. Also note that the incidence rate ratio is computed with reference to the general population.

risk of 4.7%. If we consider the average 10-year risk at age 50 (2.39%) as the threshold for screening, the age at which such a threshold is reached differs considerably at different percentiles. The age ranges from 62 for women at the 20th percentile to 38 at the 95th percentile. Women at or below the 15th percentile will never meet this threshold because of competing risks of mortality. Similarly, the age at which the 5-year risk exceeds 1.66% differs considerably for women at different risk percentiles.

The risk estimates for women with one affected first-degree relative are shown in Table 3. The risks are higher than in the general population. The actual risks range from 13.2% at the 5th percentile to 35.4% at the 95th percentile. The relative risk when comparing women at the 5th and the 95th percentile is smaller for those with family history (2.7 times versus 3.9 times), but the range is larger (22.2% versus 16.3%). The age at which the 10-year risk exceeds the threshold (2.39%) ranges from 47 for women at the 5th risk percentile to 33 for women at the 95th percentile.

The NNS vary considerably across different risk levels. The NNS ranges from 930 (the risk at the 95th percentile) to 3913 (the risk at the 5th percentile) in the population. For women with a family history of breast cancer, the risks are higher and the NNS is reduced. The range is from 532 to 1624.

Table 3. Predicted Risk of Breast Cancer in Women Having One Affected First-Degree Relative from Genetic Profiles

Percentile	Lifetime Risk (95% CI)	Incidence Rate Ratio (95% CI)	10-Year-Risk at Age 50 (95% CI)	NNS	Age at which 10-Year Risk > 2.39%	Age at which 5-Year Risk > 1.66%
5	0.132 (0.119–0.138)	1.11 (0.99–1.17)	0.026 (0.024-0.028)	1634	47.3	57.6
10	0.150 (0.138–0.156)	1.28 (1.17–1.33)	0.030 (0.028-0.032)	1420	44.2	54.4
20	0.175 (0.165–0.179)	1.51 (1.43–1.55)	0.036 (0.034-0.037)	1206	41.3	50.3
30	0.194 (0.187–0.197)	1.70 (1.63–1.73)	0.040 (0.039-0.041)	1075	39.7	47.3
40	0.211 (0.207–0.213)	1.87 (1.83–1.89)	0.044 (0.043-0.045)	978	38.5	45.2
50	0.228 (0.226–0.229)	2.05 (2.03–2.06)	0.048 (0.048-0.048)	896	37.6	43.8
60	0.246 (0.245–0.246)	2.24 (2.23–2.25)	0.053 (0.052-0.053)	823	36.8	42.7
70	0.265 (0.263–0.269)	2.45 (2.43–2.50)	0.057 (0.057-0.059)	753	36.0	41.6
80	0.289 (0.286–0.297)	2.72 (2.68–2.82)	0.064 (0.063-0.066)	680	35.1	40.6
90	0.324 (0.318–0.338)	3.13 (3.06–3.31)	0.073 (0.071-0.077)	594	34.0	39.2
95	0.354 (0.345–0.374)	3.51 (3.40–3.77)	0.081 (0.079-0.087)	532	33.1	38.3

The incidence rate ratio is computed with reference to the general population. The following abbreviations are used: CI, confidence interval; NNS, number of individuals that need to be screened in order to prevent one cancer death.

Table 4. Predicted Risk of Breast Cancer in Women Having One Affected First-Degree Relative, with Consideration of the Relative's Genotype Profile

Percentile (Relative)	Percentile (Index Individual)	Lifetime Risk (95% CI)	Incidence Rate Ratio (95% CI)	10-Year-Risk at Age 50 (95% CI)	Age at which 10-Year Risk > 2.39%	Age at which 5-Year Risk > 1.66%
5	5	0.141 (0.128–0.147)	1.19 (1.07–1.25)	0.028 (0.026–0.030)	45.7	56.0
10	5	0.137 (0.124–0.143)	1.16 (1.04–1.21)	0.028 (0.025–0.029)	46.4	56.7
30	5	0.129 (0.115–0.135)	1.08 (0.96–1.14)	0.026 (0.023–0.027)	48.0	58.3
50	5	0.123 (0.110–0.130)	1.03 (0.91–1.09)	0.025 (0.022–0.026)	49.2	59.5
70	5	0.118 (0.104–0.124)	0.99 (0.86–1.04)	0.024 (0.021–0.025)	50.4	61.0
90	5	0.111 (0.097–0.117)	0.92 (0.80–0.98)	0.022 (0.019–0.023)	52.0	63.8
95	5	0.107 (0.093–0.114)	0.89 (0.77–0.95)	0.021 (0.018–0.023)	52.8	66.2
5	10	0.162 (0.151–0.167)	1.39 (1.29–1.44)	0.033 (0.031–0.034)	42.6	52.3
10	10	0.158 (0.146–0.163)	1.35 (1.25–1.40)	0.032 (0.030–0.033)	43.2	53.1
30	10	0.149 (0.137–0.154)	1.27 (1.16–1.32)	0.030 (0.028–0.031)	44.4	54.6
50	10	0.143 (0.131–0.148)	1.21 (1.10–1.26)	0.029 (0.026–0.030)	45.3	55.7
70	10	0.137 (0.125–0.143)	1.16 (1.04–1.21)	0.028 (0.025–0.029)	46.3	56.7
90	10	0.129 (0.116–0.135)	1.08 (0.97–1.14)	0.026 (0.023–0.027)	48.0	58.2
95	10	0.125 (0.112–0.131)	1.05 (0.93–1.11)	0.025 (0.022–0.026)	48.8	59.1
5	90	0.366 (0.357–0.386)	3.66 (3.54–3.93)	0.085 (0.082–0.091)	32.8	38.0
10	90	0.359 (0.350–0.378)	3.57 (3.46–3.82)	0.083 (0.080–0.088)	33.0	38.2
30	90	0.345 (0.337–0.362)	3.39 (3.30–3.61)	0.079 (0.077–0.084)	33.4	38.6
50	90	0.335 (0.328–0.351)	3.27 (3.19–3.47)	0.076 (0.074–0.080)	33.6	38.9
70	90	0.326 (0.319–0.340)	3.16 (3.08–3.33)	0.073 (0.072–0.077)	33.9	39.2
90	90	0.312 (0.307–0.325)	2.99 (2.93–3.14)	0.070 (0.068–0.073)	34.3	39.7
95	90	0.306 (0.301–0.318)	2.92 (2.86–3.06)	0.068 (0.067–0.071)	34.5	39.9
5	95	0.401 (0.389–0.426)	4.13 (3.97–4.51)	0.095 (0.091–0.103)	32.0	37.1
10	95	0.394 (0.382–0.419)	4.03 (3.88–4.39)	0.093 (0.089–0.101)	32.1	37.2
30	95	0.379 (0.369–0.402)	3.84 (3.70–4.15)	0.089 (0.086–0.095)	32.5	37.6
50	95	0.369 (0.359–0.391)	3.71 (3.58–4.00)	0.086 (0.083–0.092)	32.7	37.9
70	95	0.359 (0.350–0.380)	3.58 (3.46–3.85)	0.083 (0.080–0.089)	33.0	38.1
90	95	0.345 (0.337–0.364)	3.40 (3.30–3.63)	0.079 (0.077–0.084)	33.4	38.6
95	95	0.339 (0.331–0.356)	3.31 (3.22–3.53)	0.077 (0.075–0.082)	33.5	38.8

CI is an abbreviation for confidence interval.

Another interesting point to note is that if we compare the lifetime risk of all women against those with an affected first-degree relative at every risk percentile, the ratio of risks decreases with higher percentiles (Table S3). Intuitively, when a woman is at a high-risk percentile, she has already possessed many of the risk genotypes that would account for her familial risk. Hence, the increase in risk because of positive family history will be smaller than average.

Finally, for individuals with an affected first-degree relative, we considered adding the affected relative's genotype profile to the risk prediction model (Table 4 and Table S4). We assumed both the individual and the relative were genotyped on the same set of susceptibility variants. Adding the relative's genomic profile helps to further stratify the individuals' risks into finer categories. For instance, a woman at the 95th risk percentile might have a predicted risk ranging from 33.9% to 40% after considering her relative's risk percentile. The highest and lowest predicted risks are from the most discordant pairs (one at 5th and the other at 95th percentile). Intuitively, if the relative is at a high percentile of the measured liability (reflecting known genes), the *un*measured liability (reflecting unknown genes) tends to be lower, and the individual will have a lower risk. The age at which the 10-year predicted risks exceed the threshold of 2.39% also differs, the difference

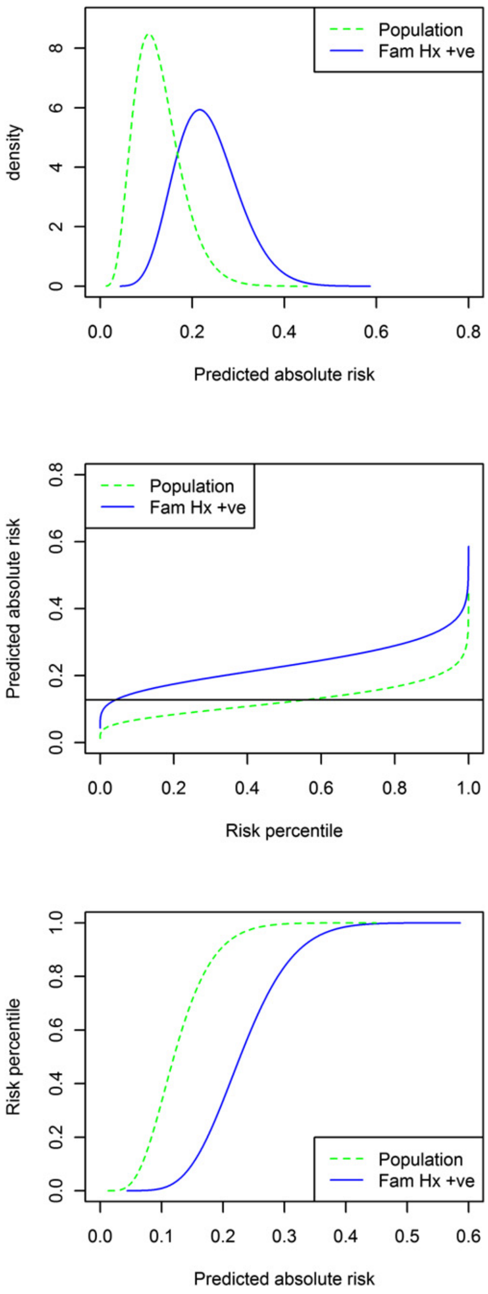

Figure 1. Plots of Predicted Risks of Breast Cancer in the General Population and in Individuals with a Family History of Disease and with One Affected First-Degree Relative
Top: The probability density functions of predicted risks. Middle: The predictiveness curve (predicted risk plotted against the risk percentile). Bottom: the cumulative density functions of predicted risks. The horizontal line in the middle graph represents the average lifetime risk in the whole population. People with an affected first-degree relative are denoted by "Fam Hx +ve."

being larger for lower-risk individuals. This is because their risks increase more slowly with age, whereas the high-risk women pass the threshold quite early in their lives. For women at the 5th percentile, the age at which their 10-year risks pass the threshold ranges from 45.7 to 52.8 when the relative's risk percentile ranges from the 5th to the 95th.

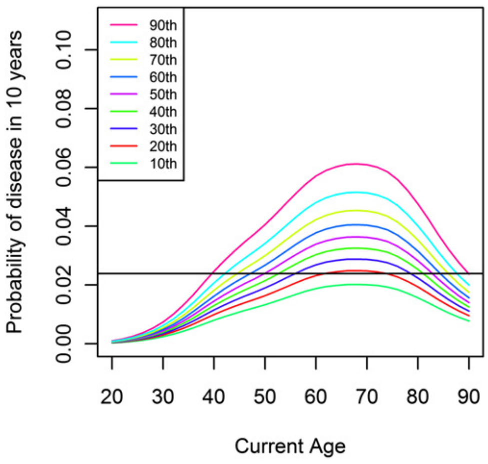

Figure 2. Ten-Year Risk of Breast Cancer at Different Risk Percentiles for the General Female Population
The horizontal line represents the average 10-year risk of breast cancer for a 50-year-old woman.

Figure 1 shows the distribution of predicted risks in the general population and in those with one affected first-degree relative. Three types of graphs are shown, including the risk distribution (or probability density function), the plot of predicted risk versus risk percentile (predictiveness curve), and the cumulative density of risks. Figures 2 and 3 show the 10-year risk of breast cancer at different ages, in the population and within those with an affected first-degree relative. The curve goes down at older ages because of competing risks of mortality that reduce the absolute risk of breast cancer. From Figure 2, it is clear that a woman might never reach the risk threshold for screening. The graphs also show the rate of increase in predicted risks with age; there are steeper curves for people at higher percentiles.

The observations and conclusions from analyses on prostate cancer are similar to those for breast cancer (see

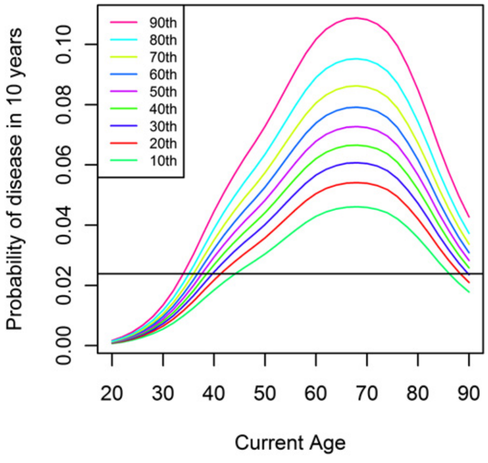

Figure 3. Ten-Year Risk of Breast Cancer at Different Risk Percentiles for Women with One Affected First-Degree Relative
The horizontal line represents the average 10-year risk of breast cancer for a 50-year-old woman.

Table 5. Predicted Risk of Prostate Cancer in General Population

Percentile	Lifetime Risk (95% CI)	Incidence Rate Ratio (95% CI)	10-Year Risk at Age 50 (95% CI)	10-Year Risk at Age 60 (95% CI)	NNS	Age at which Risk > 2.11%
5	0.044 (0.043–0.053)	0.29 (0.28–0.35)	0.006 (0.006–0.007)	0.018 (0.017–0.021)	5477	64.3
10	0.059 (0.057–0.068)	0.39 (0.38–0.45)	0.008 (0.008–0.010)	0.023 (0.023–0.027)	4103	58.5
20	0.081 (0.079–0.089)	0.54 (0.53–0.60)	0.012 (0.011–0.013)	0.033 (0.032–0.036)	2949	54.8
30	0.100 (0.099–0.108)	0.69 (0.67–0.74)	0.014 (0.014–0.016)	0.041 (0.040–0.044)	2353	52.8
40	0.120 (0.119–0.125)	0.83 (0.82–0.87)	0.017 (0.017–0.018)	0.049 (0.049–0.051)	1956	51.3
50	0.140 (0.139–0.143)	0.98 (0.98–1.01)	0.021 (0.021–0.021)	0.058 (0.058–0.059)	1656	50.1
60	0.162 (0.162–0.163)	1.16 (1.16–1.17)	0.024 (0.024–0.025)	0.068 (0.068–0.068)	1411	49.0
70	0.189 (0.186–0.189)	1.38 (1.36–1.38)	0.029 (0.029–0.029)	0.080 (0.079–0.080)	1196	48.0
80	0.223 (0.216–0.224)	1.67 (1.61–1.68)	0.035 (0.034–0.035)	0.097 (0.093–0.097)	994	46.9
90	0.275 (0.261–0.278)	2.17 (2.03–2.19)	0.045 (0.042–0.046)	0.123 (0.116–0.125)	779	45.6
95	0.323 (0.302–0.327)	2.66 (2.43–2.71)	0.055 (0.051–0.056)	0.149 (0.137–0.151)	645	44.6

The incidence rate ratio is computed with reference to the general population. The following abbreviations are sued: CI, confidence interval; NNS, number of individuals that need to be screened in order to prevent one cancer death.

Tables 5 and 6; Table S5; and Figures S1, S2, and S3). The total Vg for prostate cancer is higher than that for breast cancer, and the range of the predicted risks is larger. For men with an affected first-degree relative, the age at which 10-year risk exceeds the threshold (2.11%) ranges from 42 at the 95th percentile to 53 at the 5th percentile. For the general population, the corresponding age range is from 45 to 64. Notably, the 10-year risks at age 60 vary widely. The risks range from 4.2% (5th percentile) to 23.2% (95th percentile) in those with an affected first-degree relative and from 1.8% to 14.9% in the population. It is also worth noting that the NNS varies substantially from 645 (95th percentile) to 5477 (5th percentile) in the population and from 414 to 2282 in those with an affected first-degree relative. These results suggest that it might be more cost-effective to screen high-risk individuals as stratified by their genotype and family history.

We also observed that for people with one affected first-degree relative but at the lowest risk percentiles (<10th), their predicted lifetime risks are even lower than the median population lifetime risk. The influence of the known susceptibility loci begins to outweigh that of a positive family history.

Discussion

By considering the connections between the variance (heritability) explained and absolute risk based on the liability threshold model, we have proposed a statistical

Table 6. Predicted Risk of Prostate Cancer in Men Having One Affected First-Degree Relative

Percentile	Lifetime Risk (95% CI)	Incidence Rate Ratio (95% CI)	10-Year Risk at Age 50 (95% CI)	10-Year Risk at Age 60 (95% CI)	NNS	Age at which 10-Year Risk > 2.11%
5	0.103 (0.100–0.118)	0.71 (0.68–0.82)	0.015 (0.014–0.017)	0.042 (0.041–0.048)	2282	52.6
10	0.128 (0.125–0.142)	0.89 (0.87–1.00)	0.019 (0.018–0.021)	0.053 (0.052–0.059)	1818	50.8
20	0.163 (0.161–0.174)	1.17 (1.15–1.26)	0.025 (0.024–0.026)	0.069 (0.067–0.074)	1401	49.0
30	0.192 (0.190–0.200)	1.41 (1.39–1.48)	0.030 (0.029–0.031)	0.082 (0.081–0.086)	1172	47.9
40	0.219 (0.218–0.225)	1.64 (1.63–1.69)	0.034 (0.034–0.035)	0.095 (0.094–0.098)	1012	47.0
50	0.246 (0.246–0.249)	1.89 (1.88–1.91)	0.039 (0.039–0.040)	0.108 (0.108–0.110)	887	46.3
60	0.275 (0.274–0.275)	2.16 (2.16–2.16)	0.045 (0.045–0.045)	0.123 (0.123–0.123)	780	45.6
70	0.308 (0.303–0.309)	2.50 (2.44–2.50)	0.052 (0.051–0.052)	0.140 (0.138–0.141)	684	44.9
80	0.348 (0.338–0.350)	2.94 (2.82–2.96)	0.061 (0.058–0.061)	0.163 (0.157–0.164)	589	44.1
90	0.406 (0.389–0.410)	3.65 (3.43–3.70)	0.075 (0.070–0.076)	0.198 (0.187–0.201)	484	43.2
95	0.456 (0.432–0.461)	4.36 (4.00–4.43)	0.089 (0.082–0.090)	0.232 (0.215–0.235)	414	42.4

The incidence rate ratio is computed with reference to the general population. The following abbreviations are sued: CI, confidence interval; NNS, number of individuals that need to be screened in order to prevent one cancer death.

framework that allows the combination of family history and the genetic profiles of the individual and his or her relative in disease risk predictions.

A distinct advantage of our method is the flexibility and ease that allows for any family structure and size. Because the known risk factors are conceptualized as a continuous variable of measurable liability, the algorithm can be easily extended to deal with more complex pedigrees within the same framework of liability distributions. The computation requirement is not substantially increased.

Because the prediction model is based on the variance-explained framework, we can also readily extend the model to handle markers in LD, haplotypes, or multilocus genotypes (when interactions are present). The extensions are described in detail elsewhere.[7]

Another feature of our approach is that not only the affection status but also the genotype information (or potentially other risk factor profiles) of family members can also be incorporated in risk prediction. Formulas were provided for risk estimates given the measurable liability of the individual and his or her relative. We have also extended the method of calculating Vg to continuous predictor variables. Although the primary motivation of the study is to incorporate genetic factors in risk prediction, the method can also be extended to incorporate other categorical and continuous risk factors if the covariance structures between the risk factors are specified. In addition, we suggest methods to provide age-specific disease risk estimates over a specified period of time.

Comparison with Other Risk Prediction Approaches that Take Family History into Account

Ruderfer et al.[18] recently proposed an innovative approach to predict risk that incorporates family history and, possibly, the relative's genotype profiles Their method was also based on a liability threshold model and was applied to Crohn disease (MIM 266600) data. Our approach, however, is largely different, and we believe it carries numerous further advantages. To combine information across multiple loci, the method of Ruderfer et al.[18] requires computing the likelihood ratio for each locus. This step involves complex conditional probabilities and requires calculation of the joint genotype frequency of the individual and family members and consideration of all possible parental mating and transmission types. Extension of the method to more family members and complex family structure is probably more tedious and complicated. In contrast, our approach is built on the concept of a continuous measurable liability, which summarizes the contribution for all genetic (and/or nongenetic) risk factors. The computation is greatly simplified and extension to complex pedigree structure is very straightforward. Our algorithm also easily handles markers in LD, haplotypes, multiallelic markers, interactions, etc. by the appropriate calculation of Vg. Another advantage is that we can visualize the distribution of predicted risk for individuals with a family history of the disease by analytic calculations. Risk at any percentile can be readily computed. These measures are often of interest from a public health point of view. Importantly, we also studied age-conditional risk estimation and the implications of our risk prediction approach for cancer screening; these were not explored in previous works.

The most standard and commonly used approach for risk prediction is probably logistic regression. Family history can also be added as a covariate in the regression equation. This approach is relatively straightforward and can be easily implemented in most statistical software. It is free of the liability threshold model assumption. Obviously, we need to assume the data can be fit well by the standard logistic model (or other statistical models employed).

It is, however, difficult to deal with complex family structures with logistic regression. Usually we just consider whether an individual has a family history or not without considering the detailed pedigree structure. For example, a person might have one first-degree relative and one second-degree relative affected with the disease but such family history information is hard to be incorporated into a logistic model.

Moreover, building a logistic regression model requires information on genetic and environmental risk factors and family history to be available for all subjects. Therefore, a clinical sample must be available. Such clinical studies are often expensive and time consuming to perform because information on family history and genotypes (or other risk factors) must be ascertained from every individual. Because the effect sizes and frequencies of variants can differ across populations, these clinical studies might need to be repeated in different populations as well. Also, such models are less flexible; for example, it is difficult to deal with complex family structures and update the predicted risks when new susceptibility loci are discovered.

On the other hand, our proposed approach can model the effect of family history by knowing only the allele frequencies and relative risks of the susceptibility variants and the disease heritability. These summary statistics are usually available from published studies and public databases. Raw data are not required. As such, one can combine effect size estimates of risk factors from independent studies or meta-analyses. Programs for implementing the described methodologies will be available at the "Risk prediction from family history and genetic variants" website by H.-C.S. A web interface is also under development.

An advantage of a regression approach is that it takes into account of the covariance structure of all predictor variables (because the raw data is available). The risk prediction approach based on the liability model can also, in theory, model any categorical and continuous predictor variables, but their covariance structure will need to be specified, which is less straightforward.

Another method to take family history into account is via consideration of the residual familial risk.[19] This residual familial risk is the sibling relative risk after removing the contribution of the known susceptibility variants. The relative risk conferred by the genotype (or the aggregate relative risk conferred by a combination of genetic variants) is multiplied by the residual familial risk and the disease prevalence. This method can also combine risk factor information from independent studies. The advantage of risk prediction based on the liability model is that it can be easily extended to handle arbitrary family structure and size, whereas the extension of familial risk is not straightforward and to our knowledge has yet to be derived.

Limitations of the Prediction Model

No prediction models are perfect, and there are limitations for our proposed risk prediction approach. There are a number of assumptions that should be borne in mind before clinical applications.

Rare Variants

We have mainly focused on the use of common genetic variants in prediction. Rare mutations or structural variants are usually of higher penetrance, though they usually affect a small proportion of people. For breast cancer, rare mutations such as those in *BRCA1* (MIM 113705) and *BRCA2* (MIM 600185) have long been known to alter the disease risk. More comprehensive reviews of the genetics of familial breast cancer can be found elsewhere.[20–23] However, these high-penetrance mutations are found in only about 5% to 10% of all breast cancer cases.[20] The lifetime risk of breast cancer for *BRCA1* mutation carriers is about 65% and for *BRCA2* is about 45%.[24] For comparison, a woman with one affected first-degree relative and who is at the top 0.3% according to the genetic risk conferred by the 13 common variants will have the same lifetime risk. We also observed that the highest risk in Table 4 approaches 40%. We estimated that a woman at the 98th risk percentile who has one affected relative at the second percentile also has a lifetime risk of approximately 45%. We did not consider rare variants such as *BRCA* mutations in the current analysis. Our proposed method is best suited for those who do not harbor high-penetrance rare variants. Numerous programs are available to predict the probability of carrying *BRCA* mutations.[25–27]

Other Breast Cancer Prediction Models

Numerous prediction models have been developed for breast cancer, and we shall briefly review some of them here. A detailed review of different risk assessment models can be found elsewhere.[28] The Gail model,[29] which was based on data obtained from the Breast Cancer Detection Demonstration Project, was one of the earliest and continues to be one of the most widely used prediction models. The risk factors and relative risks were determined from logistic regression. The risk factors include hormonal factors (age at menarche and first birth), number of previous breast biopsies and the number of affected first-degree relatives. Another well-known model is the Claus model.[30] This model was derived from complex segregation analysis from a nested population-based case-control study and assumes a rare autosomal dominant (AD) locus that increases susceptibility to breast cancer. Only family history is used for risk prediction. The model allows both first- and second-degree relatives to be included and is able to adjust the risks based on their ages at onset. An extension of the Claus model[31] also allows for inclusion of ovarian cancer, bilateral breast cancer, and more than two affected relatives. The International Breast Cancer Intervention Study (IBIS) or Tyrer-Cuzick model[27] combines the features of the previous two models and accounts for both family history and other clinical risk factors. It assumes two AD loci (*BRCA1* and *2*) and one hypothetical low-penetrance gene. Mutation probabilities of *BRCA1/2* can be calculated as well.

The major differences between our model and others are that we incorporate genotype information from common variations and that our model is a general framework that can also be applied to other complex diseases. Unlike most other models, development of the current prediction framework does not require any raw data from population studies. Our proposed model is based on the theory of the liability threshold model, and whether the theory works well in practice will require further validation in external datasets.

Because there are yet no empirical studies on the combined effect of common variants and family history, we compare the risk ratios of positive family history as predicted from the liability threshold model against those obtained from population studies. We consider a large-scale collaborative study on breast cancer that combined individual data from 52 epidemiological studies including 58,209 women with the disease and 101,986 controls.[32] Compared to women with no affected relatives, the reported risk ratios for women with one, two, or three or more affected first-degree relatives are 1.80, 2.93, and 3.90, respectively.[32] Employing the liability threshold model, the corresponding risk ratios are 2.09, 2.94 and 3.65 respectively (the comparison group consists of women with one *un*affected female relative and we assumed exactly three affected relatives for the third scenario). The risk ratios are 1.84, 2.58, and 3.20, respectively, when comparison is made to the general population instead. These results suggest that the liability threshold model provides risk estimates that agree reasonably well with empirical data.

Inclusion of other established environmental or clinical risk factors in the prediction model will be another important step. For breast cancer, many risk factors such as hormonal factors and personal history of breast diseases have been identified. As a simple solution, the relative risks of nongenetic factors can be directly multiplied to the current risk estimates if we assume independence of these risk factors with known genetic variants and family history. A similar approach was also employed in the IBIS (Tyrer-Cuzick) model.[27] Note, however, that this assumption might not hold in practice.

The effect sizes of SNPs in this study are mainly based on the NHGRI catalog. The risk allele frequencies, ORs, and their corresponding confidence intervals were calculated from a joint analysis of the discovery and replication samples. If a replication sample was not available, then SNPs from the discovery sample were reported. It is known that effect size estimates are often biased upward for selected markers passing a stringent significance threshold, a phenomenon known as the winner's curse[33,34]. The ORs used in this study might also be overestimated because of this bias, and the reported risks at different percentiles might appear to more dispersed than they really are. Nevertheless, the bias is unlikely to be substantial because the sample size for replication is usually large compared to the original GWAS.

We have not considered age at onset of relatives into the prediction model. Earlier age at onset might suggest stronger a genetic influence, which could affect the risk estimate for relatives. There is evidence that earlier onset age of relatives increases the risk of breast cancer.[30,32] Many prediction models take this factor into account,[27,30] mainly by analyzing and fitting models to actual data. However, we do not have any raw data, and a framework to incorporate the age at onset of relatives into the liability threshold model is yet to be derived. One difficulty is that it is often hard to determine whether the same set of variants affect *both* the age at onset and disease risk or whether age at onset is a separate trait that is at least partially determined by other variants.

The calibration of a model measures how well the predicted risks agree with the actual risks and is an important criterion to be assessed. Because of the lack of external validation data, this criterion cannot be assessed here. It should also be noted that the risks reported here are based on incidence and mortality data from the US (from the SEER database) and might not be directly applicable to other populations.

A few other assumptions are also made in the risk model. We assumed proportional hazards in age-conditional risk estimates. In other words, the hazard ratios of the risk factors are assumed to be constant regardless of age. This might not be true in practice because genetic variants or family history might exert larger influences in younger age groups. For example, it has been observed that for breast cancer, the risk ratio conferred by positive family history generally decreases with the woman's age.[32] Of the SNPs for breast cancer identified to date, none of the ORs are found to vary by age. However, the hypothesis of age-dependent ORs is seldom explicitly tested in association studies. To overcome this assumption, Equation 3 can in fact be modified such that the hazard ratio (R) is function of age rather than a fixed value. We have performed a simple test to investigate the effect of this assumption on risk estimates. We extracted the age-specific effect sizes of having relative affected with breast cancer from the Collaborative Group on Hormonal Factors in Breast Cancer.[32] The lifetime risk estimate *without* the proportional hazard assumption was 19.47%. On the other hand, assuming a constant hazard ratio of 1.80 (the aggregate effect size reported), the lifetime risk was 20.36%. For simplicity, we assume the risk ratios are close to the hazard ratios. The results are reasonably close, suggesting that the assumption might not produce a substantial effect on lifetime risk estimates.

We assumed that the effects of the SNPs are additive on the liability scale, which is close to a multiplicative model of ORs. There is no strong evidence for gene-gene and gene-environmental interactions for the SNPs under study, but this is an area that warrants further exploration. If interactions are indeed present, they can potentially be accommodated in our model as well because the Vg can still be calculated (please refer to So et al.[7]).

We have assumed heritability refers to the variance in liability to disease at any time in life. This might not be true because of censoring in real studies, especially for later-onset traits. The study design can also differ, affecting the interpretation of heritability. We can correct this problem by using more sophisticated methods such as frailty models when computing heritability estimates.[35] The assumptions we used are inevitable because there is not enough data available to enable us to consider age-specific hazard ratios or recompute heritabilities.

Another caveat to note is that the age-specific incidence and mortality rates can only be cross-sectional. These figures can change with time. For instance, if a disease becomes more common in the future, the disease risks presented here will be underestimates. As a result, projection of risk over a long period of time might be less reliable.

Implications for Targeted Prevention and Screening

We showed that a collection of susceptibility variants helps to stratify the population into diverse risk categories, which can be useful in delivering screening programs. For those with a family history, the magnitudes of predicted risks are higher and the risk range is wider. It is possible to design a more individualized screening program that takes into account the actual predicted risks from family history and genetic risk factors. In particular, high-risk individuals might benefit from earlier screening. A number of societies have issued guidelines on breast cancer screening and prevention, for example the National Institute for Health and Clinical Excellence (NICE) in UK,[36] the American Cancer Society,[37] and the National Comprehensive Cancer Network.[38] These guidelines differ, but in general they recommend earlier surveillance for women at higher risks. For example, the NICE guideline recommends mammography screening for women age 40 or older who have 10-year risks greater than 3% or lifetime risks > 17% as predicted from their family history. Other women who do not have elevated risks are offered mammography beginning at age 50. From this study, it is clear that the inclusion of genetic information helps to refine a person's risk; for example, a woman with a positive family history but a favorable genetic profile might not have an elevated risk (Table 3), whereas another woman without a family history but who harbors multiple risk variants might need earlier screening (Table 2). Clinical decisions can be altered by a person's genetic profile.

Genomic profiles can also be used in other applications such as choosing individuals for inclusion in clinical trials[39] and determining whether chemoprophylatic agents should be taken.[40] Some interventions or screening methods are more expensive or can carry greater side effects (e.g., magnetic resonance imaging screening or chemoprevention) and hence might not apply to the whole population. It might be more efficient to target the high-risk individuals instead.

As shown in Table 4, a woman with an affected first-degree relative and a high-risk genotype profile might be advised to start screening in her 30s, as deduced from the 10-year disease risk. However, practically, the relative might not have been diagnosed by that time, especially if the relative is a sister. In that case, the woman can be advised to start screening as early as possible after the diagnosis of the relative. The data presented in the tables of this paper are theoretical only and some flexibility might be required when applied in real cases.

Although genomic profiling has potential to be applied to screening and targeted prevention, the limitations and assumptions must be carefully considered. As detailed before, our proposed framework has a number of limitations (some of which also apply to other prediction models). It should be emphasized that external validation is crucial to assess the performance of our model.

Even if a prediction model is well-validated, the harms and benefits of screening must be carefully weighed before applications. Some possible sources of harm from mammography screening include radiation exposure, anxiety and distress from the screening procedure (especially when the result is a false positive), additional biopsies from false positives, and overdiagnosis.[41] Overdiagnosis can occur when early-stage breast cancer or ductal carcinoma in situ is detected in a woman (typically an elderly woman) who is likely to die from other causes or when a detected early-stage lesion never advances to a clinically significant cancer.[11]

In this study, we followed Pharoah et al.,[10] who took the 10-year risk at age 50 as the threshold at which screening has a net benefit. However, the choice of 10 years instead of a longer or shorter term of risks is somewhat arbitrary. One justification might be that clinical trials for mammography screening often have a follow up of around 10 years. Also, considering a very short-term risk (e.g., the 1-year risk) is probably not justified because the predicted risks will still be very close for people with diverse risk factors. On the other hand, risk estimates over a long period of time can be less reliable. Ideally, risk thresholds should be determined based on a detailed analysis on the potential harms and benefits. For example, Gail and Pfeiffer[42] have proposed a decision theoretic model for screening that considers the losses (or gains) that result from the false positives, false negatives, true positives, and true negatives.

The considerations are even more complicated if the targeted screening program is to be applied at the population level. For instance, we need to consider the cost saved from early detection and treatment of the diseases, the amount of morbidity and mortality reduced, the cost of implementation of the program (e.g., genotyping and personnel), and so on. Education of the public and health care professionals is also necessary.

The susceptibility variants found to date do not enable us to predict whether an individual will develop the disease with very high certainty; hence, it is inappropriate for *diagnosis* of diseases or for determining if an invasive procedure (such as surgery) should be performed. We can tell the probability that one will be affected, but these probabilities are far from 1 (definitely affected) or 0 (definitely disease-free). It is noteworthy that the very commonly employed metric of area under the receiver operation characteristic curve (AUC) is more suited for evaluating the power of a test for diagnosis than the suitability for targeted prevention. Hence, a relatively low AUC does not preclude the clinical usefulness of a prediction test.[43,44] We have not evaluated the improvement in predictive power by inclusion of genetic variants into a prediction model, for example the net reclassification improvement, AUC change, etc. Readers can refer to reference[45] for a more detailed discussion on this matter.

We have not delved into a number of complexities regarding breast cancer screening. For example, mammography is more sensitive for women over 50 years of age,[46] and the actual benefits of screening should take this factor into account. We have not differentiated estrogen-receptor (ER) positive and negative breast tumors. The effect size of SNPs can differ, depending the ER status of the tumor.[23]

For prostate cancer, the benefits over harms for screening are not clear yet.[15] However, it is possible that targeting high-risk men might improve the efficiency of screening.

The main aim of this study is to propose a statistical framework to predict risks based on family history and genomic profiles. The validity of the model will need to be confirmed empirically in actual samples. If a large clinical sample is available, one can directly use standard techniques such as logistic regression to build a risk model with genetic or other risk factors and family history as predictors. Such models will not depend on the liability threshold model. However, this approach also suffers from a number of limitations as discussed previously. For example, large clinical samples are needed, and the study is often costly and time-consuming. It is often hard to collect a sufficient sample size for rarer diseases and the model is also less flexible. The framework we propose aims to provide a reasonable estimate of the disease risk, particularly when large clinical studies are not available, that considers both family history and genotype profiles.

In the long run, clinical trials and long-term follow up are required to confirm whether the targeted prevention strategy suggested here would ultimately offer more benefits than harms. In conclusion, we suggest that for some diseases, such as breast cancer, by combining genetic profiling and family history, we can stratify the population into diverse risk categories, which can be helpful in targeted disease prevention and the delivery of more individualized screening programs. We believe that it is important to examine the potential of genetic testing for diseases on a case-by-case basis. The usefulness of such testing will depend on the clinical context, for example whether the test is intended for diagnosis or targeted prevention.

Appendix A

PA Formula

The PA selection formula[47,48] is a generalization of regression and describes how a mean vector and covariance matrix of a set of variables are distorted by selection on a subset of the variables. Suppose we have a set of random variables, partitioned into **x** and **y.** The mean vector and covariance matrices of the variables are

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} V_x & C_{xy} \\ C_{yx} & V_y \end{bmatrix}.$$

Assume that selection is performed based on the vector **x.** If the selection transforms the mean of **x** vector from $\boldsymbol{\mu}_x$ to $\boldsymbol{\mu}_x^*$, the mean vector for **y** will be changed to

$$\mu_y^* = \mu_y + C_{yx} V_x^{-1} \left(\mu_x^* \text{-} \mu_x \right).$$

If the selection transforms the covariance matrix of **x** from V_x to V_x^*, then the covariance matrix Σ will be changed to

$$\Sigma^* = \begin{bmatrix} V_x^* & V_x^* V_x^{-1} C_{xy} \\ C_{yx} V_x^{-1} V_x^* & V_y \text{-} C_{yx} \left(V_x^{-1} \text{-} V_x^{-1} V_x^* V_x^{-1} \right) C_{xy} \end{bmatrix}.$$

Remarks on the PA Formula and an Alternative Approach to Dealing with Selection via the Use of Multivariate Integration

Strictly speaking, the PA formula provides only an approximation to the predicted risks, although the approximation is very good in most cases. For instance, we consider the simplest case of predicting an individual's risk given that he or she has one affected first-degree relative. We assume the liabilities for the individual and the relative are bivariate normal with a covariance equal to $h/2$, h being the heritability.

Applying PA formula, we can estimate the mean and variance of the individual's liability conditioned on the affected relative's liability, which is truncated normal. Because the individual's and the relative's liabilities are correlated, the individual's liability will no longer be normal given that the relative is affected. The deviation from normality is, however, usually very small and can generally be neglected. This phenomenon was first noticed by Falconer.[49]

The deviation is greater if the disease is very rare (and hence has a high liability threshold and the distribution conditioned on this extreme truncated normal is very far from normality) and the covariance between the liabilities of the relative and the individual is large. In the special case in which one is interested in predicting for monozygotic twins (given that one of them is affected) under a high heritability, the PA formula is inaccurate and alternative approaches can be used.

An exact method for taking into account the correlation in liabilities is the integration of multivariate normal distributions. Fast numerical algorithms for computing normal multivariate integrals have been developed[50] and are implemented in the R package mvtnorm.

For example, the individual's risk given an affected first-degree relative can be expressed as

Pr(individual affected/relative affected) = Pr(individual AND relative affected)/Pr(relative affected)

In terms of liabilities, we are interested in the probability that the individual's liability (x_{ind}) exceeds T given that the relative's liability (x_{rel}) exceeds T. This probability can be computed by bivariate integration,

$$\begin{aligned} &\Pr(x_{ind} > T \,|\, x_{rel} > T) \\ &= \frac{\Pr(x_{ind} > T \text{ and } x_{rel} > T)}{\Pr(x_{rel} > T)} \\ &= \frac{\int_T^\infty \int_T^\infty \phi_2(x_{ind}, x_{rel}) dx_{ind}\, dx_{rel}}{1 - \Phi(T)} \end{aligned}$$

where ϕ_2 denotes the bivariate normal density.

The above method can be generalized to deal with any number of affected and unaffected relatives and the genotype information of the individual and other relatives. Assume we know that i relatives are affected, j relatives are disease-free, and the genotype profiles of k relatives are given. For notational simplicity, denote the overall liability of the i affected relatives as x_1, x_2 …x_i, the overall liability of the j disease-free relatives as y_1,y_2,…y_j, and the measurable liability (the liability score from a set of known risk variants) of the k relatives as m_1, m_2, …m_k (the corresponding random variable is denoted by M):

$$\begin{aligned} &\Pr\Big(x_{ind} > T \,|\, x_1 > T \ldots x_i > T,\, y_1 < T \ldots\, y_j < T,\, M_{ind} = m_{ind},\, M_1 = m_1 \ldots\, M_k = m_k\Big) \\ &= \frac{\Pr\Big(x_{ind} > T,\; x_1 > T \ldots x_i > T,\, y_1 < T \ldots y_j < T,\, M_{ind} = m_{ind},\, M_1 = m_1 \ldots M_k = m_k\Big)}{\Pr\Big(x_1 > T \ldots x_i > T,\, y_1 < T \ldots\, y_j < T,\, M_{ind} = m_{ind},\, M_1 = m_1 \ldots M_k = m_k\Big)}. \\ &= \frac{\Pr\Big(x_{ind} > T,\, x_1 > T \ldots x_i > T,\, y_1 < T \ldots y_j < T \,|\, M_{ind} = m_{ind},\, M_1 = m_1 \ldots M_k = m_k\Big)}{\Pr\Big(x_1 > T \ldots x_i > T,\, y_1 < T \ldots y_j < T \,|\, M_{ind} = m_{ind},\, M_1 = m_1 \ldots M_k = m_k\Big)} \end{aligned}$$

The multivariate distribution of x and y conditioned on values of measurable liability can be found by a standard result in multivariate statistics (for example see result 4.6 in Johnson and Wichern[51]). The theorem is restated here for easy reference.

Let $X = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ be distributed as $N_p(\mu, \Sigma)$. The mean vector is denoted by $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and the covariance matrix is denoted by $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ with $|\Sigma_{22}| > 0$. The conditional distribution of $\mathbf{X}_1$, given that $\mathbf{X}_2 = \mathbf{x}_2$, is normally distributed with the following mean and covariance

$$\text{Mean} = \mu_1 + \underset{12}{\Sigma} \overset{-1}{\underset{22}{\Sigma}} (\mathbf{x}_2 \text{-} \mu_2)$$

and

$$\text{Covariance} = \Sigma_{11} \text{ - } \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Having derived the mean and covariance matrix of the conditional distribution of x and y, the above expression equals

$$\frac{\int_T^\infty \dots \int_T^\infty \int_{-\infty}^T \dots \int_{-\infty}^T \phi_{1+i+j|M}\left(x_{ind},\ x_1 \dots x_i,\ y_1 \dots y_j\right) dy_1 \dots dy_j\, dx_{ind}\, dx_1 \dots dx_i}{\int_T^\infty \dots \int_T^\infty \int_{-\infty}^T \dots \int_{-\infty}^T \phi_{i+j|M}\left(x_1 \dots x_i, y_1 \dots y_j\right) dy_1 \dots dy_j\, dx_1 \dots dx_i}.$$

$$\mu_y^* = \mu_y + C_{yx}V_x^{-1}(\mu_x^* \text{-} \mu_x)$$
$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} V/2 \\ h/2 \end{bmatrix} \times 1 \times (a-0) = \begin{bmatrix} aV/2 \\ ah/2 \end{bmatrix}$$

and

$$V_y^* = V_y - C_{yx}(V_x^{-1} \text{-} V_x^{-1}V_x^*V_x^{-1})C_{xy}$$
$$= \begin{bmatrix} V & V \\ V & 1 \end{bmatrix} - \begin{bmatrix} V/2 \\ h/2 \end{bmatrix} \times (1-b) \times [V/2 \quad h/2]$$
$$= \begin{bmatrix} V-(1-b)(V^2/4) & V-(1-b)(Vh/4) \\ V-(1-b)(Vh/4) & 1-(1-b)(h^2/4) \end{bmatrix}$$

Then the second selection is performed based on the known measurable liability of the individual (M_{ind}). The new mean is equal to the observed measured liability score for the individual (m_I), whereas the new variance equals zero because the individual now has an exact liability score. The PA formula can be applied again, giving the new mean and variance of the individual's overall liability L_{ind} after selection. The results are given in the main text.

There are, however, several advantages of using the PA formula. The formula involves just matrix inversion and hence calculations can be carried out in most statistical software and spreadsheets (such as Excel). On the other hand, multivariate normal integration requires dedicated computer programs, although currently the speed of computation is also very fast. The PA approach also allows the final results to be presented in explicit closed formulas (as listed in the main text).

Risk Prediction for an Individual with One First-Degree Relative Affected

Suppose the individual has been genotyped on a set of susceptibility variants. Following the above notations, we have the following equations in the current application:

$$V_x = 1 \quad V_y = \begin{bmatrix} V & V \\ V & 1 \end{bmatrix},$$

$$C_{yx} = \begin{bmatrix} V/2 \\ h/2 \end{bmatrix} \quad C_{xy} = [V/2 \quad h/2],$$

and

$$\mu_x = 0 \quad \mu_y = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

One can then apply the PA formula to calculate the mean and covariance matrix of M_{ind} and L_{ind} after selection on L_{rel}.

Risk Prediction Given the Genotypes of Both the Individual and His or Her Affected Relative

The first selection is based on the relative's affection status. By PA formula, after the selection on L_{rel}, M_{rel}, M_{ind}, and L_{ind} has the following mean and covariance:

$$\tilde{\mu}_{1st\ sel} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} V \\ V/2 \\ h/2 \end{bmatrix} (1)(a-0) = \begin{bmatrix} aV \\ aV/2 \\ ah/2 \end{bmatrix}$$

and

$$\tilde{V}_{1st\ sel} = \begin{bmatrix} V & V/2 & V/2 \\ V/2 & V & V \\ V/2 & V & 1 \end{bmatrix} - \begin{bmatrix} V \\ V/2 \\ h/2 \end{bmatrix} (1-b)[V \quad V/2 \quad h/2]$$
$$= \begin{bmatrix} V-(1-b)V^2 & \dfrac{V-(1-b)V^2}{2} & \dfrac{V-(1-b)Vh}{2} \\ \dfrac{V-(1-b)V^2}{2} & V-\dfrac{(1-b)V^2}{4} & V-\dfrac{(1-b)Vh}{4} \\ \dfrac{V-(1-b)Vh}{2} & V-\dfrac{(1-b)Vh}{4} & 1-\dfrac{(1-b)h^2}{4} \end{bmatrix}.$$

The second selection is based on a particular value of the measurable liability of the relative (M_{rel}). Suppose we know that the relative's measurable liability equals m_R; the variance now becomes 0 as the relative has an exact liability score. The mean and covariance of M_{ind} and L_{ind} are

$$\tilde{\mu}_{2nd\ sel} = \begin{bmatrix} aV/2 \\ ah/2 \end{bmatrix} + \begin{bmatrix} \frac{V-(1-b)V^2}{2} \\ \frac{V-(1-b)Vh}{2} \end{bmatrix} \left(\frac{1}{V-(1-b)V^2} \right) \times (m_R - aV) = \begin{bmatrix} m_R/2 \\ \frac{ah}{2} + \frac{V-(1-b)Vh}{2[V-(1-b)V^2]}(m_R - aV) \end{bmatrix}$$

and

$$\tilde{V}_{2nd\ sel} = \begin{bmatrix} V - \frac{(1-b)V^2}{4} & V - \frac{(1-b)Vh}{4} \\ V - \frac{(1-b)Vh}{4} & V - \frac{(1-b)h^2}{4} \end{bmatrix} - \begin{bmatrix} \frac{V-(1-b)V^2}{2} \\ \frac{V-(1-b)Vh}{2} \end{bmatrix} \left(\frac{1}{V-(1-b)V^2} \right) \times \begin{bmatrix} \frac{V-(1-b)V^2}{2} & \frac{V-(1-b)Vh}{2} \end{bmatrix} = \begin{bmatrix} \frac{3V}{4} & \frac{3V}{4} \\ \frac{3V}{4} & 1 - \frac{(1-b)h^2}{4} - \frac{[V-(1-b)Vh]^2}{4[V-(1-b)V^2]} \end{bmatrix}.$$

The third and final selection is based on the actual value of the individual's measurable liability (M_{ind}). Let this actual value be m_I; the mean and variance of the individual's overall liability are

$$\tilde{\mu}_{final} = \frac{ah}{2} + \frac{V-(1-b)Vh}{2[V-(1-b)V^2]}(m_R - aV) + \frac{3V}{4}\left(\frac{3V}{4}\right)^{-1}\left(m_I - \frac{m_R}{2}\right) = \frac{ah}{2} - \frac{aV[1-(1-b)h]}{2[1-(1-b)V]} + \frac{1}{2}\left(\frac{1-(1-b)h}{1-(1-b)V} - 1\right)m_R + m_I$$

and

$$\tilde{V}_{final} = 1 - \frac{(1-b)h^2}{4} - \frac{[V-(1-b)Vh]^2}{4[V-(1-b)V^2]} - \frac{3V}{4}.$$

Risk Prediction Given the Measurable Liability of an Affected First-Degree Relative

We have considered the case when a first-degree relative is affected and his or her measurable liability is known. Conditioned on these factors, the mean and variance of the index individual's overall liability are

$$\tilde{\mu} = \frac{ah}{2} + \frac{V-(1-b)Vh}{2[V-(1-b)V^2]}(m_R - aV)$$

and

$$\tilde{V} = 1 - \frac{(1-b)h^2}{4} - \frac{[V-(1-b)Vh]^2}{4[V-(1-b)V^2]},$$

respectively, as given in the main text. Here, we show why higher measurable liability of the relative (m_R) would lead to a bigger risk for the index individual. Obviously, the variance is independent of m_R. We only need to consider the sign of the coefficient associated with m_R.

Note that $V - (1-b)Vh = V[1-(1-b)h]$. b is the variance of a truncated normal distribution and is between 0 and 1. We have $0 < (1-b) < 1$ and $0 < h < 1$ and $0 < V < 1$, hence $0 < V[1-(1-b)h] < 1$.

By a very similar argument, $V - (1-b)V^2 = V[1-(1-b)V]$, and we have

$$\begin{array}{l} 0 < (1-b) < 1 \quad \text{and} \quad 0 < V < 1 \\ 0 < V[1-(1-b)V] < 1 \end{array}.$$

To conclude, the coefficient associated with m_R is positive,

$$\frac{V-(1-b)Vh}{2[V-(1-b)V^2]} > 0,$$

and higher m_R would result in bigger risk for the individual.

Variance Explained for Continuous Risk Factors

We have considered the Vg for a biallelic genetic variant before. In many cases, the risk factor is continuous. For example, high blood pressure and lipid levels are risk factors for type 2 diabetes. In this case, the Vg enables us to have an idea of the contribution of the particular risk factor to the overall variance in liability to the disease. It also allows us to extend the risk prediction framework to accommodate continuous predictor variables.

Calculation Based on Original Vg Estimation Approach with Finer Risk Factor Categories

Previously we consider the calculation of Vg when the risk factor (i.e., genotype) has three categories. To extend the calculation to the scenario involving a continuous risk factor, we can use a straightforward approach of treating each level of the risk factor as a distinct category and calculating the odds ratio in each category. We can set any category with odds ratio 1. The odds ratio for another category having a difference of h units with the baseline group is given by exp(hβ) or exp(β)^{h}. In this case, the maximum number of categories is the total number of subjects. The frequency of each risk category is 1/number of subjects if there are no two people with the same level of the risk factor. The disadvantage is that this approach requires raw data.

A further generalization is to assume a certain distribution of the predictor variable. For many risk factors, we can assume a normal or log-normal distribution. We can divide the predictor variable into arbitrarily fine categories, and Vg can be calculated as usual. When the number of categories is increased, the Vg should converge to the true Vg. For simplicity, we assume the predictor variable follows a normal distribution with a given mean and variance. We can divide the distribution into k bins, and the mean in each bin is derived from a truncated normal distribution,

$$E(X \mid z_a < X < z_b) = \mu + \sigma \frac{\phi\left(\frac{z_a - \mu}{\sigma}\right) - \phi\left(\frac{z_b - \mu}{\sigma}\right)}{\Phi\left(\frac{z_b - \mu}{\sigma}\right) - \Phi\left(\frac{z_a - \mu}{\sigma}\right)}.$$

where μ and σ are the mean and standard deviation of the normal distribution, while z_a and z_b are the start and end points of the bin.

The odds ratio for each category is approximated by the odds ratio corresponding to the mean. (If the number of bins is large, the mean for each bin can also approximated by the midpoint). The frequency of falling into each category can be directly derived from the relevant cdf. For example, if the distribution is normal, this frequency is $\Phi(z_b) - \Phi(z_a)$. Then the same algorithm for calculating Vg with only three genotype categories can be followed. The limitation of this approach is that we need to assume a parametric form of the predictor variable.

Supplemental Data

Supplemental Data include three figures and five tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

The work was supported by the Hong Kong Research Grants Council General Research Fund (HKU 766906M and HKU 774707M), the European Community's Seventh Framework Programme under grant agreement No. HEALTH-F2-2010-241909 (Project EU-GEI), and the University of Hong Kong Strategic Research Theme of Genomics. H.-C.S. was supported by a Croucher Foundation Scholarship. The work was supported by the Hong Kong Research Grants Council General Research Fund.

Received: November 30, 2010
Revised: March 27, 2011
Accepted: April 4, 2011
Published online: April 28, 2011

Web Resources

The URLs for data presented herein are as follows:

Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/gwastudies/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
PubMed, http://www.pubmed.com
Risk prediction from family history and genetic variants, http://sites.google.com/site/honcheongso/fam_pred
Surveillance, Epidemiology and End Results (SEER) database, http://seer.cancer.gov/

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. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.
3. Falconer, D. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. *29*, 51–76.
4. Smith, C. (1971). Recurrence risks for multifactorial inheritance. Am. J. Hum. Genet. *23*, 578–588.
5. Curnow, R.N. (1972). The multifactorial model for the inheritance of liability to disease and its implications for relatives at risk. Biometrics *28*, 931–946.
6. Mendell, N.R., and Elston, R.C. (1974). Multifactorial qualitative traits: Genetic analysis and prediction of recurrence risks. Biometrics *30*, 41–57.
7. So, H.C., Gui, A.H., Cherny, S.S., and Sham, P.C. (2011). Evaluating the heritability explained by known susceptibility variants: A survey of ten complex diseases. Genet. Epidemiol. Published online March 3, 2011.
8. So, H.C., and Sham, P.C. (2010). Effect size measures in genetic association studies and age-conditional risk prediction. Hum. Hered. *70*, 205–218.
9. Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. Genet. Epidemiol. *32*, 227–234.
10. 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.
11. US Preventive Services Task Force. (2009). Screening for breast cancer: U.S. Preventive Services Task Force recommendation statement. Ann. Intern. Med. *151*, 716–726, W-236.
12. Rembold, C.M. (1998). Number needed to screen: Development of a statistic for disease screening. BMJ *317*, 307–312.
13. Nelson, H.D., Fu, R., Griffin, J.C., Nygren, P., Smith, M.E., and Humphrey, L. (2009). Systematic review: Comparative effectiveness of medications to reduce risk for primary breast cancer. Ann. Intern. Med. *151*, 703–715, W-226–W-235.
14. Lin, K., Lipsitz, R., Miller, T., and Janakiraman, S.; U.S. Preventive Services Task Force. (2008). Benefits and harms of prostate-specific antigen screening for prostate cancer: An evidence update for the U.S. Preventive Services Task Force. Ann. Intern. Med. *149*, 192–199.
15. U.S. Preventive Services Task Force. (2008). Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement. Ann. Intern. Med. *149*, 185–191.
16. Wolf, A.M., Wender, R.C., Etzioni, R.B., Thompson, I.M., D'Amico, A.V., Volk, R.J., Brooks, D.D., Dash, C., Guessous, I., Andrews, K., et al; American Cancer Society Prostate Cancer Advisory Committee. (2010). American Cancer Society guideline for the early detection of prostate cancer: Update 2010. CA Cancer J. Clin. *60*, 70–98.
17. Schröder, F.H., Hugosson, J., Roobol, M.J., Tammela, T.L., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., et al; ERSPC Investigators. (2009). Screening and prostate-cancer mortality in a randomized European study. N. Engl. J. Med. *360*, 1320–1328.

18. Ruderfer, D.M., Korn, J., and Purcell, S.M. (2010). Family-based genetic risk prediction of multifactorial disease. Genome Med *2*, 2.

19. Lewis, C.M., Whitwell, S.C., Forbes, A., Sanderson, J., Mathew, C.G., and Marteau, T.M. (2007). Estimating risks of common complex diseases across genetic and environmental factors: The example of Crohn disease. J. Med. Genet. *44*, 689–694.

20. De Grève, J., Sermijn, E., De Brakeleer, S., Ren, Z., and Teugels, E. (2008). Hereditary breast cancer: From bench to bedside. Curr. Opin. Oncol. *20*, 605–613.

21. Stratton, M.R., and Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. Nat. Genet. *40*, 17–22.

22. Ripperger, T., Gadzicki, D., Meindl, A., and Schlegelberger, B. (2009). Breast cancer susceptibility: Current knowledge and implications for genetic counselling. Eur. J. Hum. Genet. *17*, 722–731.

23. Mavaddat, N., Antoniou, A.C., Easton, D.F., and Garcia-Closas, M. (2010). Genetic susceptibility to breast cancer. Mol. Oncol. *4*, 174–191.

24. Antoniou, A., Pharoah, P.D., Narod, S., Risch, H.A., Eyfjord, J.E., Hopper, J.L., Loman, N., Olsson, H., Johannsson, O., Borg, A., et al. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: A combined analysis of 22 studies. Am. J. Hum. Genet. *72*, 1117–1130.

25. Parmigiani, G., Berry, D., and Aguilar, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. Am. J. Hum. Genet. *62*, 145–158.

26. Antoniou, A.C., Pharoah, P.P., Smith, P., and Easton, D.F. (2004). The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Br. J. Cancer *91*, 1580–1590.

27. Tyrer, J., Duffy, S.W., and Cuzick, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. Stat. Med. *23*, 1111–1130.

28. Amir, E., Freedman, O.C., Seruga, B., and Evans, D.G. (2010). Assessing women at high risk of breast cancer: A review of risk assessment models. J. Natl. Cancer Inst. *102*, 680–691.

29. Gail, M.H., Brinton, L.A., Byar, D.P., Corle, D.K., Green, S.B., Schairer, C., and Mulvihill, J.J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J. Natl. Cancer Inst. *81*, 1879–1886.

30. Claus, E.B., Risch, N., and Thompson, W.D. (1994). Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. Cancer *73*, 643–651.

31. van Asperen, C.J., Jonker, M.A., Jacobi, C.E., van Diemen-Homan, J.E., Bakker, E., Breuning, M.H., van Houwelingen, J.C., and de Bock, G.H. (2004). Risk estimation for healthy women from breast cancer families: New insights and new strategies. Cancer Epidemiol. Biomarkers Prev. *13*, 87–93.

32. Collaborative Group on Hormonal Factors in Breast Cancer. (2001). Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet *358*, 1389–1399.

33. Zollner, S., and Pritchard, J.K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. Am. J. Hum. Genet. *80*, 605–615.

34. Garner, C. (2007). Upward bias in odds ratio estimates from genome-wide association studies. Genet. Epidemiol. *31*, 288–295.

35. Yashin, A.I., and Iachine, I.A. (1995). Genetic analysis of durations: Correlated frailty model applied to survival of Danish twins. Genet. Epidemiol. *12*, 529–538.

36. National Institute for Health and Clinical Excellence (NICE). Clinical guideline 41. Familial breast cancer: The classification and care of women at risk of familial breast cancer in primary, secondary and tertiary care (2010). (http://guidance.nice.org.uk/CG41/NICEGuidance/pdf/English).

37. American Cancer Society. Detailed Guide: Breast Cancer (2010). (http://www.cancer.org/Cancer/BreastCancer/DetailedGuide/index).

38. National Comprehensive Cancer Network (NCCN). Clinical Practice Guidelines in Oncology. Breast Cancer Screening and Diagnosis (2011). (http://www.nccn.org/professionals/physician_gls/pdf/breast-screening.pdf).

39. Fisher, B., Costantino, J.P., Wickerham, D.L., Redmond, C.K., Kavanah, M., Cronin, W.M., Vogel, V., Robidoux, A., Dimitrov, N., Atkins, J., et al. (1998). Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J. Natl. Cancer Inst. *90*, 1371–1388.

40. Gail, M.H., Costantino, J.P., Bryant, J., Croyle, R., Freedman, L., Helzlsouer, K., and Vogel, V. (1999). Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J. Natl. Cancer Inst. *91*, 1829–1846.

41. Nelson, H.D., Tyne, K., Naik, A., Bougatsos, C., Chan, B.K., and Humphrey, L.; U.S. Preventive Services Task Force. (2009). Screening for breast cancer: An update for the U.S. Preventive Services Task Force. Ann. Intern. Med. *151*, 727–737, W237-42.

42. Gail, M.H., and Pfeiffer, R.M. (2005). On criteria for evaluating models of absolute risk. Biostatistics *6*, 227–239.

43. Cook, N.R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation *115*, 928–935.

44. Pepe, M.S., and Janes, H.E. (2008). Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. J. Natl. Cancer Inst. *100*, 978–979.

45. So, H.C., and Sham, P.C. (2010). A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. PLoS Genet. *6*, e1001230.

46. Kerlikowske, K., Grady, D., Barclay, J., Sickles, E.A., and Ernster, V. (1996). Effect of age, breast density, and family history on the sensitivity of first screening mammography. JAMA *276*, 33–38.

47. Pearson, K. (1903). Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. Philos. Transact. A. Math. Phys. Eng. Sci. *200*, 1–66.

48. Aitken, A.C. (1934). Note on selection from a multivariate normal population. Proceedings of the Edinburgh Mathematical Society B. *4*, 106–110.

49. Falconer, D.S. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. Ann. Hum. Genet. *31*, 1–20.

50. Genz, A. (1992). Numerical computation of multivariate normal probabilities. J. Comput. Graph. Stat. *1*, 141–149.

51. Johnson, R.A., and Wichern, D.W. (2007). Applied multivariate statistical analysis (Upper Saddle River, N.J.: Pearson Prentice Hall).