Effect size measures in genetic association studies and age-conditional risk prediction

Hon-Cheong So[1] and Pak C. Sham[1,2,3]

[1]Department of Psychiatry, [2]Genome Research Centre and [3]the State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong SAR, China

Corresponding author: Pak C. Sham
Address: Department of Psychiatry, 10/F Laboratory Block, LKS Faculty of Medicine, University of Hong Kong, Pokfulam, Hong Kong SAR, China
Tel: (852)2819-9557
Fax: (852)2818-5653
Email: pcsham@hkucc.hku.hk

Key words Genetic risk prediction, odds ratio, relative risk, hazard ratio, association study, competing risks, genetic counseling

Abstract

The interest in risk prediction using genomic profiles has surged recently. A proper interpretation of effect size measures in association studies is crucial to accurate risk prediction. In this study we clarify the relationship between the odds ratio(OR), relative risk and incidence rate ratios(IRR) in the context of genetic association studies.

We demonstrated that under the common practice of sampling prevalent cases and controls, the resulting OR approximate the IRR. Based on this result, we presented a framework to compute disease risk given the current age and follow-up period (including lifetime risk), with consideration of competing risks of mortality. We considered two extensions. One is correcting the incidence rate to reflect the person-years alive and disease-free, the other is converting prevalence to incidence estimates.

The methodology was applied to an example of breast cancer prediction. We observed that simply multiplying the OR by the average lifetime risk estimates yielded a final estimate >100% (101%), while using our method that accounts for competing risks produce an estimate of 63% only. We also applied the method to risk prediction of Alzheimer disease in Hong Kong.

We recommended that companies offering direct-to-consumer genetic testing employ more rigorous prediction algorithms considering competing risks.

Introduction

In recent years, there has been a rapid increase in the number of genetic association studies, and in particular genome-wide associations studies have identified many susceptibility variants for complex diseases [1,2]. The interest in genetic risk prediction has surged in view of advances in our knowledge of the genetic basis in many diseases.

Accurate measure and correct interpretation of effect sizes are crucial to proper risk prediction. There are numerous ways of quantifying effect sizes in association studies, such as odds ratios (OR), relative risks, hazard ratios (HR) (also referred to as incidence rate ratio in this paper) etc., but they are often confused. The differences and relationships between various types of effect size measures in genetic association studies is a fundamental question, yet seems to be largely unnoticed. In this study we will clarify and compare several effect size measures and how they could be used to derive the disease risk of an individual. In particular, we showed how to estimate an individual's risk within an arbitrary period of time (say 5 or 10 years, or lifetime), conditioned on his/her current age. For late-onset diseases, a person is particularly subject to competing risks of mortality and these are also taken into consideration.

This paper is organized as follows. First we review various concepts of study designs and effect size estimates. Next we propose that in prevalent case-control studies, a design very often employed in genetic association studies, yield odds ratios that approximate the incidence rate ratio, allowing for age-dependent incidences and disease durations. Based on this approximation, we presented a methodology to compute the age-conditional disease risk (including lifetime risk), with consideration of competing causes of mortality. From the formula we explained the relationship between the lifetime relative risk and incidence rate ratio. We also considered two possible complications. One is how to make correction to the incidence rate such that it is based on person-years alive *and disease-free*. The other is to convert prevalence to incidence estimates. Finally we applied the methodologies to risk prediction in breast cancer and Alzheimer's disease.

Many of the concepts presented here are built on previous works in epidemiology. In this paper we have organized the different pieces together into a coherent framework in the context of genetic risk prediction and made extensions or modifications to previous methodologies where necessary.

Methods

Measures of disease occurrence

First we review several terms often used for describing the occurrence of diseases. The incidence rate (or incidence density) is a measure of the number of new cases that arise in a population in terms of person-time units. A related quantity, the cumulative incidence (also known as the incidence proportion), is the proportion of subjects who develop the outcome of interest in the follow-up period. Thus it is also equal to the probability (or risk) of disease in the specified follow-up period. The third related concept is the incidence odds, defined as the ratio of the proportion of subjects who develop the outcome to the proportion of subjects who do not develop the outcome during follow-up (which in turn could be estimated by the ratio of the numbers of subjects in a sample).

Besides measures concerning incidence, the prevalence is another important notion. The prevalence is the proportion of people with a particular outcome at a specific time point. Analogous to the definition of incidence odds, the prevalence odds is the proportion (or number) of subjects having a particular outcome to the proportion (or number) without the outcome at a specific time point.

Effect size measures

The odds ratio is probably the most common way of summarizing the effect size of a locus in genetic association studies. Using notations in table 1, the odds ratio may be expressed as ad/bc or (a/c)/ (b/d).

The exact interpretation of the odds ratio depends on the sampling scheme. First we distinguish two main types of sampling, one sampling *prevalent* cases (people who are diseased at the time of recruitment) and the other sampling *incident* cases (new cases that arise in the follow-up of a cohort). When cases are prevalent, the odds ratio obtained is the prevalence odds ratio (POR).

When cases are incident, the interpretation of OR depends on how the controls are ascertained. One method is to choose controls who are disease-free at the *end* of the follow-up, also known as cumulative sampling. In this case, the OR obtained approximates the risk ratio (RR, sometimes also called relative risk) *under the rare disease assumption*. However, in this case one may also convert the OR obtained to RR, if the incidence of disease is known [3]. The second method is to sample controls at the *beginning* of follow-up from the entire study population who are at risk. Also known as case-cohort studies, this design measures the risk

ratio in the base population. Thirdly, controls can be sampled longitudinally with cases. When a case arises, a control is selected from the population at risk at that time point. The probability that a person is selected as a control should be proportional to his/her contribution to the person-years in the cohort. This approach provides an estimate of the incidence rate ratio in the population. For more details, readers may refer to other reviews and textbooks and the references therein [4-6].

It is commonly believed that the odds ratio serves as a good approximation to the relative risk only when the disease is rare. The rare disease assumption, however, is only needed under the cumulative sampling scheme (the 1st design described above), in which the controls are free of disease at the *end* of the follow-up.

Prevalent case-control genetic association studies

Genetic association studies commonly do *not* involve the follow-up of a cohort and *prevalent* cases and controls are sampled instead. In other words, the studies recruit one group of subjects who are diseased and another group who are disease-free at the time of recruitment. Thus the odds ratio is the prevalence odds ratio (POR).

We will focus on this type of design in the current study since sampling prevalent cases is very common for association studies and the POR obtained has a simple interpretation that facilitates evaluation of age-conditional or lifetime risks, as detailed in subsequent sections.

Approximation of incidence rate ratio by the prevalence odds ratio

The prevalence odds is equal to the product of incidence and average duration when a population is stationary, assuming incidences are independent of age. In a stationary population, the same number of newborns are added to the population per unit time, the age-specific mortality rates are constant, and net migration rates are zero for all ages [7]. Keiding [8] gave a rigorous mathematical proof of the above relationship, but this fact has been observed before in Miettinen [9] and Freeman & Hutchison [10]. If we consider two groups of people, one exposed to a risk factor and the other not exposed, we have

$$\begin{aligned} POR &= \left[K_e / (1-K_e)\right] / \left[K_{ue} / (1-K_{ue})\right] \\ &= \lambda_e D_e / \lambda_{ue} D_{ue} \end{aligned}$$

where K is the disease prevalence, λ is the incidence rate, D is the disease duration, and the subscripts e and ue represent the exposed and unexposed group respectively. For example in genetic association studies, the "exposure" refers to whether the subject has the risk genotype or allele. If we further assume that duration of disease is independent of exposure (i.e. $D_e = D_{ue}$) , it is easy to see that the POR is equal to the ratio of incidence rates [11].

However, this derivation is based on the restrictive assumptions that the population is stationary and did not consider age-specific incidences and durations.

Alho [12] gave a detailed account on the relationship of prevalence, incidence and duration of disease in a general stable population, allowing incidence and duration to be age-dependent. Based on his results, we propose that the prevalence odds ratio still approximates the incidence rate ratio under age-dependent incidence and duration, especially when both the disease incidence and the effect size of exposure are low.

We first give a note on stable population. Stable population is a more relaxed assumption than stationary population. When the growth rate in the annual number of births, age-specific mortality rates and age-specific net migration rates are constant, a population will finally converge to a stable state. The age composition of a stable population is constant over time. Note that the number of birth per unit time must be constant under *stationary* assumptions, whereas growth (no matter positive or negative) is allowed for a *stable* population. The stationary population is in fact a special case of a stable population [7].

The results given by Alho [12] are outlined below and connections will be made with the notions of prevalence odds and odds ratios. A few notations are introduced here. x refers to the age. The incidence hazard is denoted by $\lambda(x)$. The net mortality hazard (mortality in the disease-free population) is denoted by $\mu(x)$. Note that the hazard may be interpreted as the rate of outcome in a very short period of time. $f_\rho(x)$ is the age distribution of the disease-free population. $D_\rho(x)$ is the expected discounted duration of disease. The term was introduced by Alho[12]. Discounting refers to the fact that the expected period that the individual stays in the disease state in age $a>z$ is discounted depending on ρ, the intrinsic growth rate. If the population is stationary (i.e. $\rho=0$), $D_\rho(x)$ is equal to the normal expected disease duration. The formula of $D_\rho(x)$ is given in the supplementary material. Note that the subscript ρ to a function indicates that the quantities are dependent on the population intrinsic growth rate ρ. I denotes the ill population and H the healthy population. Table 2 shows a list of the main notations used in this paper.

The prevalence at the exact time t is $I(t)/[I(t)+H(t)]$. As proved in Alho [12], the prevalence odds is given by

$$\frac{I(t)}{H(t)} = \int_0^\infty f_\rho(x)\lambda(x)D_\rho(x)dx$$

Consider two groups of individuals, one group is exposed to a risk factor (in the example of genetic association studies, the exposed group is the group with a certain risk allele or genotype) while the other group is composed of unexposed individuals. It is well-known that the exposure odds ratio equals the disease odds ratio (see supplementary methods). So the exposure odds ratio obtained from case-control studies is equal to

$$\frac{I_e(t)/H_e(t)}{I_{ue}(t)/H_{ue}(t)} = \frac{\int_0^\infty f_{\rho,e}(x)\lambda_e(x)D_\rho(x)dx}{\int_0^\infty f_{\rho,ue}(x)\lambda_{ue}(x)D_\rho(x)dx} \quad (1)$$

The subscripts *e* and *ue* refers to the exposed and unexposed group respectively. The duration of disease is assumed to be independent of the exposure.

Assuming proportional hazards, i.e. the incidence rate ratio (we will also call it hazard ratio and abbreviate it as *R* in the following text) is constant across all ages, we have

$$\lambda_e(x)/\lambda_{ue}(x) = R$$

In practice $f_{\rho,e}(x)$ and $f_{\rho,ue}(x)$ (the age distribution of healthy individuals in the exposed and unexposed group) are usually similar when the hazard ratio conferred by the risk factor is not large and the disease is not very common. Intuitively, the majority of the population is healthy and the age distribution of the healthy population should not differ much unless the effect of exposure is large.

Mathematically, the age distributions of the healthy population at age *a* are

$$f_{\rho,ue}(a) = \left(e^{-\rho a}\right)\left(\frac{\exp\left(-\int_0^a (\mu(x)+\lambda_{ue}(x))\,dx\right)}{\int_0^\infty \exp\left(-\int_0^a (\rho+\mu(x)+\lambda_{ue}(x))\,dx\right)\,da}\right)$$

$$f_{\rho,e}(a) = \left(e^{-\rho a}\right)\left(\frac{\exp\left(-\int_0^a (\mu(x)+R\lambda_{ue}(x))\,dx\right)}{\int_0^\infty \exp\left(-\int_0^a (\rho+\mu(x)+R\lambda_{ue}(x))\,dx\right)\,da}\right)$$

in the unexposed and exposed groups. It is usually reasonable to assume that the population growth rate (ρ) and the mortality in the disease-free group [$\mu(x)$] are independent of the exposure. Hence the only quantity that differs in the two groups is $\lambda(x)$. When the disease is rare and the effect size is not large, we can have the following approximation for both exposed and unexposed groups

$$f_\rho(a) \approx \left(e^{-\rho a}\right)\left(\frac{\exp\left(-\int_0^a \mu(x)\,dx\right)}{\int_0^\infty \exp\left(-\int_0^a (\rho+\mu(x))\,dx\right)\,da}\right)$$

and the age distribution will be approximately the same in both groups. To further test whether the age distribution are similar, we also made use of epidemiological data on breast cancer from the Surveillance Epidemiology and End results (SEER) database (http://seer.cancer.gov) and plot the distributions under different hazard ratios and population growth rates.

To conclude, we may approximate the prevalence odds ratio by

$$\frac{I_e(t)/H_e(t)}{I_{ue}(t)/H_{ue}(t)} = \frac{\int_0^\infty f_{\rho,e}(x)[R\ \lambda_{ue}(x)]D_\rho(x)dx}{\int_0^\infty f_{\rho,ue}(x)\lambda_{ue}(x)D_\rho(x)dx} \approx R$$

Therefore the prevalence odds ratio approximates the incidence rate ratio (IRR), assuming the population is stable and the duration of illness is unaffected by the exposure. We show that the above is true even when we allow the incidence to be *age-dependent*. Also, the approximation works well particularly when the prevalence odds ratio and the disease incidence are low. Interestingly, we have invoked the rare disease assumption in this case, but now it is used for approximating POR to IRR (rather than OR to RR).

Calculating the absolute risk

Previously we have shown the incidence rate ratio can be approximated by the prevalence odds ratio, which is often directly available from association studies. This observation is useful when we perform risk prediction. We will proceed to the calculation of absolute disease risks in a specified period of time, taking into account the current age of the individual and competing risks of mortality. The core methods were developed by Gail et al. [13] and Dupont [14,15]. Dupont did not consider the application of multiple risk factors in prediction. Gail *et al.*[13] presented an approach to risk prediction using multiple risk factors derived from a logistic model, but the method requires estimation of the baseline hazard (the hazard from a subject with none of the risk factors). This method is cumbersome when the number of risk factors is large. For example, 20 biallelic loci results in 3^{20} = 3,486,784,401 combinations of genotypes, creating substantial difficulties for the baseline hazard estimation. We have made some modifications to facilitate risk prediction in the presence of multiple susceptibility loci or other risk factors.

Inputs for calculating absolute risks

The inputs required include the incidence hazard $\lambda(x)$, the net mortality hazard (mortality in the disease-free population) $\mu(x)$, and the incidence rate ratio or hazard ratio R. In the context of genetic risk prediction, R is dependent on the individual's genotypes. For notational simplicity, this dependency is not explicitly shown in the formulas.

The incidence function $\lambda(x)$ can be obtained from databases or other epidemiological studies. The incidence is usually given in 5-year intervals, but we can construct a continuous function by spline or linear interpolation between the mid-points in each age band. Alternatively, one may assume the incidence is constant in each age band.

Note that strictly speaking, $\lambda(x)$ should be the incidence in the population who are alive *and disease-free*, although sometimes the incidence may not be defined exactly in this way. For example, in the Surveillance Epidemiology and End results (SEER) program of the U.S. National Cancer Institute, the person-years are estimated by the mid-year population of the catchment area and include people previously diagnosed with the disease. In the next section we provided a simple analytic correction method. The inputs required are the total and net mortality rates as well as the incidence rates.

Incidence rate given alive and disease-free (correction of incidence rates when it includes whole population)

The basic approach is to multiply the denominator by the prevalence of being disease-free [16]. The corrected age-specific incidence rate can be expressed as

$$\lambda(x) = \frac{C(x)}{L(x)\left[1 - K(x)\right]}$$

where x is the age at the start of the interval, C is the number of first primary cases, L is the mid-year population for a given year and the specified age group and K is the prevalence of disease. Since the *un*corrected incidence is $C(x)/L(x)$, we have

$$\text{Corrected } \lambda(x) = \text{uncorrected } \lambda(x) / \Pr(\text{disease-free at age } x) \qquad (2)$$

The prevalence of being disease-free at an age x is given by the disease-free survival probability divided by the overall survival probability to age x:

$$\Pr(\text{disease-free at age } x) = \frac{\exp\left[-\int_0^x (\lambda(x) + \mu(x))dx\right]}{\exp\left[-\int_0^x \xi(x)dx\right]}$$

where $\xi(x)$ is the total mortality rate. Although the prevalence-correction approach was

described before and applied in life-tables [16,17], we offered an analytic formula for the correction. The subsequent steps for computing the absolute risk are exactly the same (detailed in subsequent sections).

Fay *et al.* [18,19] provided an alternative method to correct for this problem. Their approach directly estimates the absolute risk with consideration of competing risks, without finding the prevalence first. Their method is mathematically more complex than the one presented here.

In addition to the incidence rates, the net mortality $\mu(x)$ needs to be calculated. $\mu(x)$ is the mortality rate due to other causes among people who are alive *and disease-free*. $\mu(x)$ may not be directly available, but as in Dupont & Plummer[15] and Fay *et al.*[19], this is approximately equal to the overall mortality in population *minus* mortality from the disease in population. In essence we assume that the mortality rate from other causes *given alive* equals the mortality rate from other causes *given alive and disease-free*, i.e. the mortality rate from other causes is the same regardless of whether the person is affected or not.

Formulas for calculating the absolute risk

Now we present the formulas for computing absolute risks. The disease-free survival function F is given by

$$F(a,R,b) = \exp\left[-\int_a^b (R\lambda(x) + \mu(x))dx\right]$$

which is the probability that a person with current age a and incidence rate ratio of disease R will survive to age b free of the disease.

Figure 1 presents the relationship of the different quantities involved in the formulas. (modified from Keiding[8] p.378)

For a person whose current age is a with incidence rate ratio of disease R, the probability of disease in the next s years is given by

$$p(s,a,R) = R\int_a^{a+s} \lambda(b)F(a,R,b)db$$

If the lifetime risk is to be estimated, one may set a to be the lowest possible age-of-onset and $a+s$ to be a sufficiently large number (e.g. 110). The choice of the upper bound $a+s$ has very little effect on lifetime risk estimates, as long as the upper bound is large enough. We have tried different values of the upper bound to breast cancer data (described in detail below), and found the final risk estimate to be very close. When the upper bounds were 90, 100, 110, 120 and 130, the risk estimates were 12.21%, 12.60%, 12.63%, 12.64% and 12.64% respectively.

This is mainly due to the very high competing risks of mortality when a person gets to very old. Indeed, the estimate will converge to a certain limit since one can only live up to a certain age and the mortality rate beyond that age is essentially 100%.

Combining the above formulas, the absolute risk can be written as

$$p(s,a,R) = R\int_a^{a+s} \lambda(b) \exp\left[-\int_a^b (R\lambda(x) + \mu(x))dx \right] db \qquad (3)$$

The equation can be explained in an intuitive manner. Re-express the absolute risk as

$$p(s,a,R) = \int_a^{a+s} R\lambda(b) \times F(a,R,b)db$$

The probability of disease at exactly age *b* is the product of the following probabilities:

1. Pr(the person survive to age *b* free of the disease)
2. Pr(the person develops the disease at age *b*, given that his/her incidence rate ratio of disease is *R*)

When we say that the person develops the disease in the next *s* years, the person may get the disease at any age *b*, provided that $a <= b <= a+s$, so we have to add up or integrate all probabilities corresponding to the possible values of *b*.

Simulations on the change in absolute risk with increase in hazard ratio

To investigate how the absolute risk will change with increases in the hazard ratio, we performed a simple simulation again using incidence and mortality data from breast cancer (SEER database). We observed that even when the hazard ratio gets extremely large (say 10000), the resulting lifetime risk (risk from age 10 to 110) never exceed one (Figure 3 and 4). This is a favorable property as the actual disease risk obviously should not be more than one.

Additional remarks on the formulas

There are a few points that require attention. The *R* above is the incidence rate ratio *as compared to the general population*. If multiple risk factors are present, *R* is the aggregate hazard ratio estimate, usually obtained by multiplying the hazard ratio of each risk factor. Methods to derive *R* are described in the next section. As explained before, although it is possible to derive the baseline incidence rate (the incidence of the population at the lowest level of risk) and apply the incidence rate ratio directly from the original association studies, this approach is more troublesome when there are numerous risk factors.

Also, the above formulas assume proportional hazards, hence *R* is constant. The formulae can be generalized to accommodate age-dependent changes in the incidence rate ratio, as described in Gail *et al.* [13] . The absolute risk can be re-written as

$$p(s,a,R(b)) = \int_a^{a+s} R(b)\lambda(b)\exp\left\{-\int_a^b [R(x)\lambda(x)+\mu(x)]dx\right\}db \quad (4)$$

by considering R as a function of age. However, as age-specific odds ratios are seldom available or reported in genetic association studies, we will assume the proportional hazards assumptions hold.

Incidence rate ratios as compared to the general population

Association studies usually report the OR as compared to a reference genotype. However, to apply our formula for predicting risks, it is more convenient to express the OR relative to the general population rather than to a reference genotype. A simple method for such conversion has been described in the documentation by 23andMe (https://23andme.https.internapcdn.net/res/5166/pdf/23-01_Estimating_Genotype_Specific_Incidence.pdf). This method only requires the overall prevalence of disease and the odds ratios for two genotypes compared to the reference genotype. We also extended this approach to deal with unscreened controls and continuous risk factors.

Assume the controls are disease-free, so that the exposure odds ratio equals the disease odds ratio as proved previously. Denoting the three genotypes by aa, Aa and AA and the event indicator by D, we have

$$\Pr(D) = \Pr(D \mid aa)\Pr(aa) + \Pr(D \mid Aa)\Pr(Aa) + \Pr(D \mid AA)\Pr(AA) \quad (5a)$$

$$OR_2 = \frac{\Pr(D \mid Aa)/[1-\Pr(D \mid Aa)]}{\Pr(D \mid aa)/[1-\Pr(D \mid aa)]} \quad (5b)$$

$$OR_3 = \frac{\Pr(D \mid AA)/[1-\Pr(D \mid AA)]}{\Pr(D \mid aa)/[1-\Pr(D \mid aa)]} \quad (5c)$$

where Pr(D) is the prevalence of the event and OR_2 and OR_3 are the odds ratios for genotypes Aa and AA relative to aa. We can solve the three equations for the penetrances Pr(D|aa), Pr(D|Aa) and Pr(D|AA), given that Pr(D), OR_2 and OR_3 are known. This approach can also used when there are more than 3 genotypes. The solution to the above equations is given in the supplementary methods.

When population controls are used (controls are unscreened), we have

$$\Pr(D) = \Pr(D \mid aa)\Pr(aa) + \Pr(D \mid Aa)\Pr(Aa) + \Pr(D \mid AA)\Pr(AA)$$

$$OR_2 = \frac{\Pr(D \mid Aa)}{\Pr(D \mid aa)}$$

$$OR_3 = \frac{\Pr(D \mid AA)}{\Pr(D \mid aa)}$$

The solutions to the above equations are straightforward.

The genotype-specific OR relative to the general population (OR*) can then be obtained. For instance, for genotype aa,

$$OR^*_{aa} = \frac{\Pr(D \mid aa)/[1-\Pr(D \mid aa)]}{\Pr(D)/[1-\Pr(D)]} \quad (6)$$

Calculations for the other two genotypes are similar. It is straightforward to combine risk estimates from multiple loci using the above formulation. Assuming the loci are independent and their effects are multiplicative, the aggregate OR is equal to the product of the subject's OR at each locus. However, it should be noted that in practice the independence assumption may be violated since some loci may be in LD. In this case, one may consider the haplotypes formed by the loci or the genotype combinations (e.g. 9 combinations for 2 loci) and evaluate the effect size in each stratum. It is easy to obtain the corresponding OR of each stratum compared to the general population by the same equation-solving approach. If it is not possible to obtain data for haplotypes or genotype combinations, then one may pick up tag SNPs from each associated region and consider the effect of the "representative" SNPs only.

Alternatively, if the frequencies of the genotypes in cases and in non-cases/general population are known, the penetrance for each genotype can also be easily obtained by Bayes' law,

$$\begin{aligned}\Pr(D \mid aa) &= \frac{\Pr(aa \mid D)\Pr(D)}{\Pr(aa)} \\ &= \frac{\Pr(aa \mid D)\Pr(D)}{\Pr(D)\Pr(aa \mid D) + \Pr(\bar{D})\Pr(aa \mid \bar{D})}\end{aligned}$$

Continuous predictor variables

The OR compared to general population can still be computed when the predictor variable Z is continuous (like blood pressure). If the distributions of the predictor variable in cases and in non-cases or in the population are known, the penetrance for a particular level of the predictor variable is

$$\begin{aligned}\Pr(D \mid Z = z) &= \frac{f(z \mid D)\Pr(D)}{f(z)} \\ &= \frac{f(z \mid D)\Pr(D)}{\Pr(D) f(z \mid D) + \Pr(\bar{D}) f(z \mid \bar{D})}\end{aligned}$$

where f is the density function of the continuous risk factor Z.

If the regression coefficients β_0 and β_1 are given for a logistic regression model, the

fitted probability is $\Pr(Z = z)$ which in turn equals $\exp(\beta_0 + \beta_1 z) / \left[1 + \exp(\beta_0 + \beta_1 z)\right]$.

Note that for case-control studies the β_0 needs to be corrected as

$$\text{True } \beta_0 = \text{calculated } \beta_0 - \log r$$

where *r* is the ratio of sampling rate of cases to sampling rate of non-cases [20,21].

A summary of the risk prediction algorithm is presented in table 3.

Incidence rate ratio and (lifetime) relative risk

Although these two effect size estimates seem similar, they are not identical especially when the outcome is common. For convenience, we assume all three risks are calculated relative to the general population.

The lifetime relative risk is the lifetime risk conferred by a particular set of risk factors compared to lifetime risk of the general population. It is a very useful concept in risk prediction since an individual's lifetime risk of disease is simply the average lifetime risk in population multiplied by the lifetime relative risk. The lifetime RR can be expressed as

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

For R>=1 ,

$$\int_a^b (R\lambda(x) + \mu(x))dx \geq \int_a^b (\lambda(x) + \mu(x))dx$$

$$\exp\left[-\int_a^b (R\lambda(x) + \mu(x))dx\right] \leq \exp\left[-\int_a^b (\lambda(x) + \mu(x))dx\right]$$

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

hence $RR_{lifetime}$<= R for R>=1. Similarly, we can show that when R<1, $RR_{lifetime}$ >R. In other words, $RR_{lifetime}$ is always closer to 1 than the hazard ratio R. Note that this phenomenon also applies to relative risk in any specified period of time, since a and s can take any positive values. The phenomenon that relative risk is less than hazard ratio has been shown before [22], but competing risks have *not* been considered in the previous study.

When will the lifetime relative risk (or the relative risk in a specified period of time) be close to the hazard ratio? When the incidence of disease is low, we can make the following approximations:

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

by assuming that $R\lambda(x)+\mu(x) \approx \mu(x)$ and $\lambda(x)+\mu(x) \approx \mu(x)$. This approximation works better when *R* is small and the disease incidence is low, for example when predicting risk of disease in a short term.

Extension: conversion of prevalence to incidence

Since prevalence data are easier to collect, sometimes only prevalence instead of incidence data are available. Here we present an approach to convert prevalence to incidence, modified from Podgor and Leske [23]. The methodology will be applied to Alzheimer disease in Hong Kong, as detailed in a later section. In the current application, we considered a more "continuous" version of the original method [23] by spline interpolation of the original prevalence data given in 5-year intervals. In addition, we also considered the case when the net mortality rates are not available.

The following approach allows the disease incidence to be calculated from prevalence data for irreversible diseases with differential mortality. Consider a group of individuals currently aged *x* years followed for *N* years. The prevalence proportions at the start and end of the *N*-year interval are known and denoted by P_0 and P_1 respectively. The model is characterized by three *cumulative* hazards (or rates), *M*, Λ and *A*. Note that the upper case letters represent cumulative functions, while the corresponding lower case letters represent the hazard function. Therefore *M* is the cumulative mortality rate among disease-free individuals (as opposed to μ), Λ is the cumulative incidence rate (as opposed to λ) and *A* is the cumulative mortality rate among affected individuals (as opposed to α). The rates are expressed in units of (*N* years)$^{-1}$. We will set *N*=1 in the following discussions.

In practice, the prevalence and mortality data may be available in say, 5 year intervals. To construct continuous prevalence and mortality hazard functions, interpolation may be used.

To obtain the *cumulative* functions, simply integrate the corresponding hazard functions over the desired age range. For example, to obtain the incidence between ages a_1 and a_2, we will need to consider other cumulative hazards over this age range, which are given by

$$M(a_1,a_2) = \int_{a_1}^{a_2} \mu(t)dt \qquad\qquad A(a_1,a_2) = \int_{a_1}^{a_2} \alpha(t)dt$$

For simplicity here we assume that the mortality rate among affected individuals only depends on age.

Here we demonstrate how to deal with the case when the net mortality rates are *not* directly available but the overall population mortality rates and the hazard ratio for mortality due to the disease can be obtained. If net mortality data is available, one can directly obtain $\mu(t)$ and skip some of the steps below.

Denote the overall mortality hazard by h_{overall}, the mortality hazard among disease-free individuals by μ and the mortality hazard among affected individuals by α, we have

$$h_{overall}(t) = [1 - K(t)]\mu(t) + K(t)\alpha(t)$$

where t refers to the exact age which is continuous.

Suppose the hazard ratio for mortality in the affected group compared to the unaffected group is HR_{mort}, then $\alpha(t) = HR_{mort}\mu(t)$ and

$$h_{overall}(t) = [1 - K(t)]\mu(t) + K(t)HR_{mort}\mu(t)$$

$$\mu(t) = \frac{h_{overall}(t)}{1 - K(t) + HR_{mort}K(t)}$$

Podgor and Leske [23] showed that

$$\frac{(1-P_0)P_1 e^{-(M+\Lambda)}}{1-P_1} = \frac{P_0 e^{-A} + (1-P_0)[e^{-A} - e^{-(M+\Lambda)}]\Lambda}{M + \Lambda - A} \qquad (7)$$

For simplicity, the notations a_1 and a_2 showing the age range are dropped. P_0 and P_1 are the prevalences at ages a_1 and a_2 respectively. The equation can be solved numerically for Λ since all the other quantities are known. Since t=1, Λ equals $\lambda(t)$ for t within the age interval calculated. (For example if t=5, one can divide Λ by 5 to obtain $\lambda(t)$.)

Results

On the approximation of incidence rate ratio (hazard ratio) by prevalence odds ratio

We compared the distribution of the disease-free population $f_\rho(x)$ among exposed and unexposed group under different combinations of hazard ratios and population growth rate. We also simulated a more common disease by doubling the incidence in each age group. The incidence and mortality rates were all extracted from the 2004-2006 breast cancer data in SEER database. As shown in figure 2 and supplementary figures, the two curves were generally close, with greater discrepancy at larger hazard ratios and higher incidence. The

intrinsic population growth rate ρ, which ranges between -2.5% to 2.5%, did not have substantial effects on the separations of the two distributions. These figures were chosen according to a recent study [24] which reported that the intrinsic growth rate ranged from -2.44% to 2.16% in a list of selected countries and regions after taking into account of migration (table 2 in Preston & Wang [21]).

We also considered the actual approximation of IRR by POR by formula (2) (see supplementary methods for more detailed derivations). As suggested by Alho [12], the expected duration of disease can often be approximated by a simple linear function of age, i.e. $D_0(x) = \beta_0 + \beta_1 x$, where x is the age. The subscript 0 under D indicates the growth is zero, or the population is stationary, such that the discounted expected duration equals the expected duration. We took β_0=42.6 and β_1 = -0.485 according to a least square fit to the breast cancer data in Hakama & Hakulinen [25]; the same equation was applied in Alho's paper. The derivation of $D_\rho(x)$ (particularly where ρ is nonzero) requires the age- and disease duration-specific mortality function in affected individuals, but it is not easily available. For simplicity and illustration purposes, the same duration function $D_0(x)$ was used for ρ=0.025 and -0.025. The calculations were repeated for different IRR and for twice the original incidence.

Results were shown in table 4. In general the IRR were close to the POR, although the approximation was poorer when the effect size became larger and the incidence was doubled. Changes in the intrinsic growth rate did not affect the results to a large extent, but interestingly the approximation was better for negative growth rate and worse for positive growth rate in the case of breast cancer.

Applications to risk prediction in breast cancer

First the original incidence rates from the SEER database were corrected for the inclusion of people previously diagnosed with breast cancer. The corrected incidence rates and prevalences are shown in table 5. As expected, the effect of such adjustment is more prominent in the older age groups in which the prevalences are higher.

Next we illustrate how the methodologies described above can be used in practice. We considered 13 loci with compelling evidence of association to breast cancer (Table 6). Suppose we wish to predict risk for a US white woman who is homozygous for the risk allele at 10 loci (the first 10 loci of table 6) and heterozygous at the other 3 loci. The original OR are first converted to OR compared to the general population. The aggregate OR compared to the population is 8.01 under a multiplicative model. The lifetime risk estimate is 63.3%. If we

simply treat the aggregate OR as the lifetime relative risk, the resulting lifetime risk is 12.6% x 8.01 = 101%, which is obviously impossible.

The cumulative risks up to various ages are shown in figure 5. Figure 6 and 7 shows the residual lifetime risk and 10-year risk at different ages. The risks for the woman and for the population are both displayed.

Application to Alzheimer's disease risk in Hong Kong

We also illustrate how to convert prevalence to incidence estimates for risk estimation. Age-specific incidence disease for Alzheimer's disease (AD) is not available in Hong Kong, but previously a prevalence study has been carried out [26]. Since age-specific net mortality data was also not directly available, we considered the hazard ratio of mortality due to AD and combined with overall mortality data in the population. The age- and sex-adjusted hazard ratio was taken as 1.7 according to a previous study [27]. Mortality in the Hong Kong population was based on 2008 estimates from the Department of Health (http://www.dh.gov.hk/english/pub_rec/pub_rec_ar/pdf/0809/tabA05.pdf). Mortality rates beyond 85 were estimated by a 3rd-degree polynomial regression to mortality rates from 60 to 85.

The risk of AD from age 70 in 25 years was estimated to be 17%. We further considered the APOE loci in predicting risks. Because of lack of large-scale meta-analytic estimates of effect sizes of this loci in the Chinese, we took the effect sizes from the Japanese in a meta-analysis [28]. The risks from age 70 in 10, 20 and 25 years are displayed in table 7.

Discussion

In this study we provided a coherent and rigorous framework for age-conditional risk prediction using genomic profiles and clarified the interpretation of several effect size measures in association studies. The principles are also illustrated in two disease examples. To summarize, our novel findings and contributions include: (1) Showing mathematically that the OR approximate the incidence rate ratio in prevalent case-control studies, allowing the incidence and disease duration to be age-dependent instead of constant; (2) a practical approach to calculating the age-conditional disease risk from multiple genetic markers and other risk factors (categorical or quantitative), with consideration of competing risks; (3) demonstrating by simulations that the proposed method never produce absolute risks >1, even when the effect sizes are extremely large; (4) a new and simple analytic formula to compute the incidence rate based on people who are alive *and disease free*; (5) proving that the relative

risk is *always* closer to 1 than the incidence rate ratio, taking into account competing risks; (6) application to breast cancer and showing that the conventional risk prediction method by commercial companies may yield an estimate >100%; (7) application to Alzheimer's disease in Hong Kong and demonstrating the approach to risk prediction when only prevalence data is available.

It is commonly believed that the OR is good approximation to the RR (risk ratio) under the rare disease assumption. The statistical reasoning is that the RR = [$a/(a+b)$]/[$c/(c+d)$] (using notations in table 1) which is roughly equal to (a/b) / (c/d) when a and c are small. However, it should be emphasized that the above reasoning is valid only if incident cases are sampled and cumulative sampling is employed.

For a typical prevalent case-control association study, from our derivations the OR obtained indeed approximate the risk ratio (over a period time) when the disease is rare, but through a much more complex relationship than the "standard" theory mentioned above. The prevalent study records the *prevalence* OR, which approximates the incidence rate ratio, which in turns approximates the risk ratio. As detailed in the methods section, the mathematics and assumptions involved are more intricate than one may think.

Several companies have already offered direct-to-consumer genetic testing (e.g. deCODEme, 23andMe and Navigenics). However, the statistical methodologies they employed for risk prediction are not completely rigorous. None of these three companies considered competing risks in risk estimation, even for more common diseases. As an example, deCODEme takes the OR from association studies directly as the lifetime relative risk. We have shown in the breast cancer example that under this assumption, a person may get a risk estimate of >100%. Employing the competing risk formula will result in a risk less than 1. It is noteworthy that in our hypothetical example, the absolute risk estimates differ by ~38% for the two approaches. In addition, none of these companies allows users to enter their current age and estimate their risk in a certain period of time (e.g. 5 or 10 years), a quantity that may be of greater interest to the individual and the clinician. In addition, lifetime risk estimates may be less reliable than risk over a shorter term because of temporal trends of disease.

Caveats and limitations

In the first section of this chapter we have discussed the approximation of incidence rate ratio (IRR) by prevalence odds ratio (POR). We have performed a brief simulation study on the age distribution of healthy population in the exposed and unexposed. Theoretically, the exact difference between the IRR and POR may be computed based on formula (1). However, we will also need to know the duration function ($D\rho(z)$), which is usually not as easily

available as the incidence or mortality rates. More extensive methodological and simulation studies are required to more accurately quantify the impact of effect size, disease incidence and other assumptions such as proportional hazards on the accuracy of approximation.

There are some other possible errors that may affect the accuracy of the risk estimates. First, it should be noted that the age-specific incidence and mortality rates that one can obtain are only cross-sectional. It may be difficult to predict the trend of a disease in the long run. For instance, if a disease happens to become more common in the next 20 years, the current long-term risk estimate will be smaller than the actual value.

Also we have assumed that the hazard ratios of risk factors are constant and independent of age, as in formula (2). This assumption of proportional hazards however may not always hold. For example, genetic factors or family history may exert larger influences in younger age groups (e.g. in [29]). Older age groups on the other hand may be more affected by environmental factors. Few studies have investigated whether genetic variants exerted different effects in different age groups. For studies with large sample sizes, it is possible to consider if the effect size varied according to age or if age and genotype showed significant interactions.

The risk model requires input of incidence and net mortality rates. The accuracy of the final risk estimate will depend on the accuracy of these input data. It is therefore important to invest efforts for reliable epidemiological data for disease occurrences. A good example is the SEER database of the National Cancer Institute, which is the most comprehensive and authoritative source of information on cancer incidence and survival in the US. It collects cancer incidence and survival data from population-based cancer registries covering about 26 percent of the US population (~74 million people), ~50 million of which are Caucasians.

Recently, Yang et al. [30] have considered lifetime risk estimates using genetic profiles and the uncertainties of the estimates. Lifetime risks were computed using life-tables in which a large hypothetical cohort was constructed and the number of cases and deaths were estimated in 5-year age intervals. In this study we employed a more analytic approach, using explicit mathematical formulas based on competing risk models. Our analytic approach allows easy calculation of absolute risk with any current age and period of follow-up. The current method also allows smoother estimates of incidence and mortality functions by linear or spline interpolations, instead of assuming constant incidence in each age band. Moreover, the formulas provided better insights into the different effect size measures, such as the comparison of lifetime relative risk and incidence rate ratios.

The uncertainties of the risk estimates were not explored in the current study, but were the focus in the paper by Yang et al. They have shown that the changes in genotype risk ratio and allele frequencies have very limited impact on lifetime risk estimates, but changes in incidence rates have the largest influence on risk estimates. This result is unsurprising, although it raises the point that we should be cautious when applying a risk model to a different population or ethnic group. The paper termed variation in incidence as "uncertainty", which may not be the most appropriate since sometimes we do know that the differences in incidence in different populations. For example, it is often possible to obtain cancer or other disease statistics for a particular local population (like a city or a state). "Uncertainty" commonly refers to sampling variation in a statistical sense. This may be more pertinent to rare diseases that require larger sample sizes for reliable measures of disease occurrence.

It is also worthwhile to note that variation of incidence affects *all* sorts of prediction models based on competing risk methodology (like formula 2). The effects are not specific to the use of genomic profiles. For instance, the famous Gail model [13] of predicting breast cancer was based on the same competing risk methodology and is also affected by variations in incidence. The Gail model has been validated [31] and has been use to evaluate the risks and benefits for chemoprevention in breast cancer [32]. Therefore the criticism of "uncertainty" (particularly in incidence) should not be regarded as a specific reason against genomic profiling.

As a side note, in reference [30] it is mentioned that OR may be used as a proxy to the population risk ratio and for more common diseases, correction of OR to RR may be achieved by Zhang and Yu's [3] method. This correction however will only work if the association study samples *incident* cases and controls who are free of disease at the *end* of the cohort. This design is *not* commonly employed in genetic association studies though.

Despite the many caveats discussed, genomic profiling has the potential to allow individuals to be better informed of disease risks. Risk estimation however needs to be performed carefully with the proper statistical methodology and possible sources of error should be borne in mind.

Acknowledgements

The work was supported by the Hong Kong Research Grants Council General Research Fund grants HKU 766906M and HKU 774707M and the University of Hong Kong Strategic Research Theme of Genomics. Hon-Cheong So was supported by a Croucher Foundation Scholarship.

Web resources

Relevant R programs are available at

http://sites.google.com/site/honcheongso/software/gene-pred

References:

1 Hindorff L, Junkins H, Hall P, Mehta J, Manolio T: A catalog of published genome-wide association studies. Available at: Http://www.Genome.Gov/gwastudies/ accessed 9th july 2010.

2 Manolio TA, Brooks LD, Collins FS: A hapmap harvest of insights into the genetics of common disease. J Clin Invest 2008;118:1590-1605.

3 Zhang J, Yu KF: What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA 1998;280:1690-1691.

4 Pearce N: What does the odds ratio estimate in a case-control study? Int J Epidemiol 1993;22:1189-1192.

5 Knol MJ, Vandenbroucke JP, Scott P, Egger M: What do case-control studies estimate? Survey of methods and assumptions in published case-control research. Am J Epidemiol 2008;168:1073-1081.

6 Rothman KJ, Greenland S, Lash TL: Modern epidemiology, ed 3rd. Philadelphia, Lippincott Williams & Wilkins, 2008.

7 Preston SH, Heuveline P, Guillot M: Demography : Measuring and modeling population processes. Oxford, UK ; Malden, MA, Blackwell Publishers, 2001.

8 Keiding N: Age-specific incidence and prevalence - a statistical perspective. J Roy Stat Soc a Sta 1991;154:371-412.

9 Miettinen O: Estimability and estimation in case-referent studies. Am J Epidemiol 1976;103:226-235.

10 Freeman J, Hutchison GB: Prevalence, incidence and duration. Am J Epidemiol 1980;112:707-723.

11 Pearce N: Effect measures in prevalence studies. Environ Health Perspect 2004;112:1047-1050.

12 Alho JM: On prevalence, incidence, and duration in general stable-populations. Biometrics 1992;48:587-592.

13 Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879-1886.

14 Dupont WD: Converting relative risks to absolute risks: A graphical approach. Stat Med 1989;8:641-651.

15 Dupont WD, Plummer WD, Jr.: Understanding the relationship between relative

and absolute risk. Cancer 1996;77:2193-2199.
16 Merrill RM, Feuer EJ: Risk-adjusted cancer-incidence rates (united states). Cancer Causes Control 1996;7:544-552.
17 Feuer EJ, Wun LM, Boring CC, Flanders WD, Timmel MJ, Tong T: The lifetime risk of developing breast cancer. J Natl Cancer Inst 1993;85:892-897.
18 Fay MP: Estimating age conditional probability of developing disease from surveillance data. Popul Health Metr 2004;2:6.
19 Fay MP, Pfeiffer R, Cronin KA, Le C, Feuer EJ: Age-conditional probabilities of developing cancer. Stat Med 2003;22:1837-1848.
20 Neuhaus J: Bias due to ignoring the sample design in case-control studies. Australian & New Zealand Journal of Statistics 2002;44:285-293.
21 Prentice RL, Pyke R: Logistic disease incidence models and case-control studies. Biometrika 1979;66:403-411.
22 Symons MJ, Moore DT: Hazard rate ratio and prospective epidemiological studies. J Clin Epidemiol 2002;55:893-899.
23 Podgor MJ, Leske MC: Estimating incidence from age-specific prevalence for irreversible diseases with differential mortality. Stat Med 1986;5:573-578.
24 Preston S, Wang H: Intrinsic growth rates and net reproduction rates in the presence of migration. Population and Development Review 2007;33:657-666.
25 Hakama M, Hakulinen T: Estimating the expectation of life in cancer survival studies with incomplete follow-up information. J Chronic Dis 1977;30:585-597.
26 Chiu HF, Lam LC, Chi I, Leung T, Li SW, Law WT, Chung DW, Fung HH, Kan PS, Lum CM, Ng J, Lau J: Prevalence of dementia in chinese elderly in hong kong. Neurology 1998;50:1002-1009.
27 Ganguli M, Dodge HH, Shen C, Pandav RS, DeKosky ST: Alzheimer disease and mortality: A 15-year epidemiological study. Arch Neurol 2005;62:779-784.
28 Farrer L, Cupples L, Haines J, Hyman B, Kukull W, Mayeux R, Myers R, Pericak-Vance M, Risch N, Van Duijn C: Effects of age, sex, and ethnicity on the association between apolipoprotein e genotype and alzheimer disease. A meta-analysis. Apoe and alzheimer disease meta analysis consortium. JAMA 1997;278:1349-1356.
29 Li X, Sundquist J, Sundquist K: Age-specific familial risks of depression: A nation-wide epidemiological study from sweden. J Psychiatr Res 2008;42:808-814.
30 Yang Q, Flanders WD, Moonesinghe R, Ioannidis JP, Guessous I, Khoury MJ: Using lifetime risk estimates in personal genomic profiles: Estimation of uncertainty. Am J Hum Genet 2009;85:786-800.
31 Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS: Validation studies for models projecting the risk of invasive and total breast cancer incidence. J Natl Cancer Inst 1999;91:1541-1548.

32 Gail MH, Costantino JP, Bryant J, Croyle R, Freedman L, Helzlsouer K, Vogel V: Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. J Natl Cancer Inst 1999;91:1829-1846.

Tables

Table 1 A typical 2 by 2 contingency table for a genetic association study

	Case	Control
Risk variant present	a	b
Risk variant absent	c	d

Table 2 Main notations used in this paper

Symbol	Meaning
K	Disease prevalence
ρ	Population intrinsic growth rate (when used as a subscript to a function, means that the function depends on the intrinsic growth rate)
$f_\rho(x)$	Age distribution of the disease-free population
$D_\rho(x)$	Expected discounted duration of disease at age x
$\mu(x)$	Incidence rate at age x
$\lambda(x)$	Net mortality rate (i.e. mortality rate for disease-free individuals) at age x
$\alpha(x,d)$	Mortality rate for affected individuals at age x, which may be dependent on the illness duration d
$\mathrm{M}(a_1,a_2)$	Cumulative incidence (between ages a_1 and a_2)
$\Lambda(a_1,a_2)$	Cumulative net mortality (between ages a_1 and a_2)
$\mathrm{A}(a_1,a_2)$	Cumulative mortality rate for affected individuals(between ages a_1 and a_2)
R	Incidence rate ratio (also referred to as the hazard ratio)
$\xi(x)$	Total mortality rate at age x

Table 3 Algorithm for age-conditional risk estimates from susceptibility loci

(1)	Take the prevalence odds ratio (POR) as the incidence rate ratio
(2)	Obtain the POR as compared to the general population, instead of comparing to the baseline group (formulae 5 and 6) Continuous risk factors can also be used.
(3)	Multiply POR of each risk factor together to obtain "aggregate POR"
(Optional)	Obtain incidence from prevalence data (formula 7)
(Optional)	Correct the incidence such that it is based on the person-years alive and disease-free (formula 2)
(4)	Use the competing risk formula (i.e. formula 3) to calculate risk given current age and period of follow-up (e.g. 10 years)
(Optional)	Use age-dependent POR if necessary (formula 4)

Table 4 Comparison of incidence rate ratio and prevalence odds ratios

	Original incidence	Doubled incidence
IRR	POR	POR
ρ=0		
1.2	1.197	1.194
1.5	1.491	1.480
2	1.975	1.944
3	2.922	2.827
5	4.727	4.403
8	7.201	6.335
10	8.696	7.384
ρ=0.025		
1.2	1.195	1.189
1.5	1.484	1.467
2	1.956	1.914
3	2.871	2.751
5	4.582	4.220
8	6.880	6.018
10	8.256	7.010
ρ= -0.025		
1.2	1.202	1.203
1.5	1.506	1.508
2	2.015	2.016
3	3.038	3.025
5	5.084	4.950
8	8.076	7.463
10	9.975	8.861

The comparison was based on breast cancer data in US white females (SEER database). IRR, incidence rate ratio; POR, prevalence odds ratio; ρ, intrinsic growth rate for population.

Table 5 Corrected incidence for person-years alive and disease-free for breast cancer

Starting age	Ending age	Original Incidence Rate	Corrected incidence	Prevalence
0	1	0	0	0
1	5	0	0	0
5	10	0	0	0
10	15	3.20E-07	3.20E-07	8.00E-07
15	20	1.81E-06	1.81E-06	6.13E-06
20	25	1.29E-05	1.29E-05	4.00E-05
25	30	7.53E-05	7.53E-05	2.47E-04
30	35	2.52E-04	2.53E-04	9.89E-04
35	40	5.66E-04	5.68E-04	2.84E-03
40	45	1.16E-03	1.17E-03	6.60E-03
45	50	1.81E-03	1.83E-03	0.013
50	55	2.19E-03	2.24E-03	0.022
55	60	2.69E-03	2.78E-03	0.032
60	65	3.39E-03	3.54E-03	0.044
65	70	3.85E-03	4.09E-03	0.057
70	75	4.02E-03	4.33E-03	0.072
75	80	4.23E-03	4.62E-03	0.086
80	85	4.14E-03	4.60E-03	0.099
85	90	3.49E-03	3.92E-03	0.110
90	95	2.94E-03	3.32E-03	0.116
95	105	2.27E-03	2.58E-03	0.119

Table 6 Lifetime risk, lifetime relative risk and odds ratio for 13 breast cancer susceptibility loci

Region	Reported Genes	SNPs	RAF	OR of Aa	OR of AA	OR* of aa	OR* of Aa	OR* of AA	Life risk (aa)	Life risk (Aa)	Life risk (AA)	Lifetime RR (Aa)	Lifetime RR (AA)
10q26.13	FGFR2	rs2981582	0.38	1.260	1.588	0.828	1.044	**1.315**	0.106	0.131	0.162	1.240	1.531
11p15.5	LSP1	rs3817198	0.3	1.070	1.145	0.959	1.026	**1.098**	0.122	0.129	0.138	1.065	1.133
16q12.1	TNCR9	rs3803662	0.25	1.200	1.440	0.907	1.089	**1.306**	0.115	0.137	0.161	1.184	1.398
17q23	COX11	rs6504950	0.73	1.053	1.108	0.927	0.976	**1.027**	0.118	0.124	0.130	1.049	1.100
1p11.2	Intergenic	rs11249433	0.39	1.160	1.346	0.886	1.028	**1.192**	0.113	0.130	0.148	1.148	1.315
2q35	Intergenic	rs13387042	0.51	1.250	1.563	0.788	0.985	**1.231**	0.101	0.125	0.153	1.232	1.511
3p24	NEK10/SLC4A7	rs4973768	0.46	1.110	1.232	0.906	1.006	**1.116**	0.115	0.127	0.140	1.102	1.213
5p12	MRPS30/FGFR10	rs10941679	0.243	1.190	1.416	0.914	1.087	**1.294**	0.116	0.136	0.160	1.175	1.376
5q11.2	MAP3K1	rs889312	0.28	1.130	1.277	0.931	1.052	**1.189**	0.118	0.132	0.148	1.120	1.253
6q22.33	ECHDC1, RNF146	rs2180341	0.21	1.410	1.988	0.848	1.196	**1.686**	0.108	0.149	0.202	1.374	1.868
6q25.1	C6orf97	rs2046210	0.37	1.290	1.664	0.817	**1.054**	1.359	0.105	0.133	0.167	1.267	1.598
8q24.21	Intergenic	rs13281615	0.4	1.080	1.166	0.939	**1.014**	1.095	0.119	0.128	0.137	1.074	1.153
2q33-q34	CASP8	rs1045485	0.87	1.136	1.291	0.800	**0.909**	1.033	0.103	0.116	0.130	1.127	1.269

RAF, risk allele frequency; Life risk, absolute lifetime risk ; RR, relative risk.

The capital A refers to the risk allele and the genotype aa is the baseline group with odds ratio 1. OR* represents the OR as compared to the general population. The bolded numbers represent the genotypes possessed by a hypothetical case as discussed in the text. The aggregate OR from these genotypes equals 8.01.

Table 7 Estimated Alzheimer's disease risk by APOE genotype (in Hong Kong)

APOE genotype	Frequency	OR	OR to population	Risk at 70 in 10 years	Risk at 70 in 20 years	Risk at 70 in 25 years	Relative risk (for risk from 70 to 95)
3/3	0.491	1	0.200	0.007	0.030	0.039	1
2/2	0.003	1.1	0.220	0.008	0.033	0.042	1.10
2/3	0.039	0.9	0.180	0.007	0.027	0.035	0.90
2/4	0.009	2.4	0.480	0.018	0.071	0.089	2.30
3/4	0.369	5.6	1.120	0.040	0.155	0.188	4.88
4/4	0.089	33.1	6.621	0.215	0.554	0.588	15.25

Figure legends

Figure 1: The three states for an individual are depicted. H is the healthy (and alive) state, I is the illness/disease (and alive) state and M is the state of death (mortality). The greek symbols represent the hazard functions for transition to another state. x denotes the age and d denotes the duration of illness. $\lambda(x)$ is the incidence rate , $\mu(x)$ is the mortality rate for a disease-free individual, also called net mortality ; $\alpha(x,d)$ is the mortality rate for affected individuals which can be dependent on the illness duration.

Figure 2 Age distribution of disease-free individuals in the exposed group and in the general population. f(age), proportion of disease-free individuals at specific ages. The hazard ratio is 2 and the intrinsic growth rate is 0 (stationary population).

Figure 3 Increase in absolute lifetime risk for breast cancer with increases in IRR. Incidence rate ratio ranges from 1 to 50. The horizontal line represent absolute risk =100%.

Figure 4 Increase in absolute lifetime risk for breast cancer with increases in IRR. Incidence rate ratio ranges from 50 to 10000. The horizontal line represent absolute risk =100%.

Figure 5 Cumulative risks up different ages (x-axis) for a woman homozygous at 10 risk loci and in population

Figure 6 Residual lifetime risk for a woman homozygous at 10 risk loci and in population

Figure 7 Ten-year risks for a woman homozygous at 10 risk loci and in population

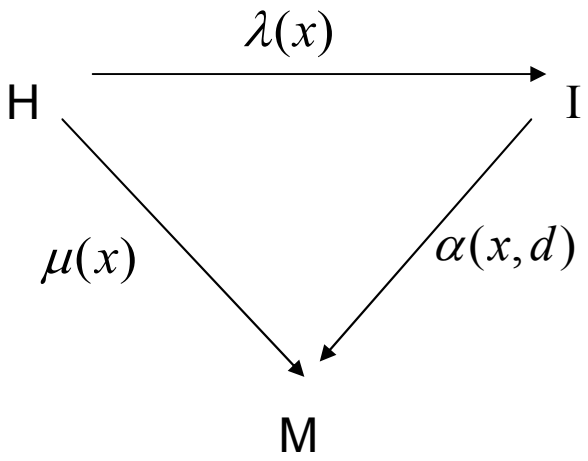
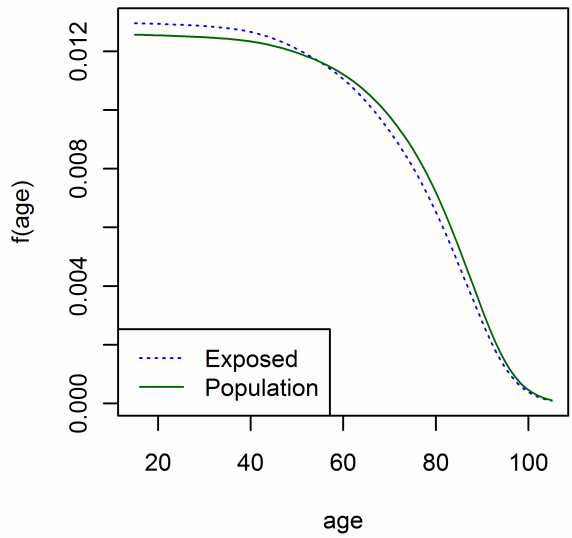


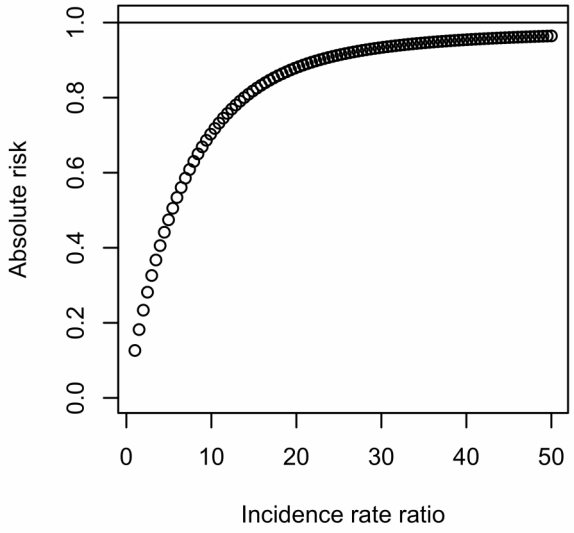
Figure 1: The three states for an individual are depicted. H is the healthy (and alive) state, I is the illness/disease (and alive) state and M is the state of death (mortality). The greek symbols represent the hazard functions for transition to another state. x denotes the age and d denotes the duration of illness. $\lambda(x)$ is the incidence rate , $\mu(x)$ is the mortality rate for a disease-free individual, also called net mortality ; $\alpha(x,d)$ is the mortality rate for affected individuals which can be dependent on the illness duration.

Figure 2　Age distribution of disease-free individuals in the exposed group and in the general population



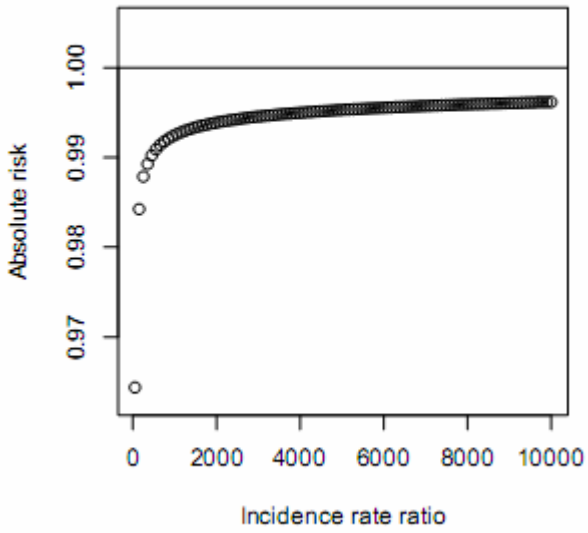
f(age), proportion of disease-free individuals at specific ages. The hazard ratio is 2 and the intrinsic growth rate is 0 (stationary population).

Figure 3 Increase in absolute lifetime risk for breast cancer with increases in IRR



Incidence rate ratio ranges from 1 to 50.

Figure 4



Incidence rate ratio ranges from 50 to 10000. The horizontal line represent absolute risk =100%.

Figure 5 Cumulative risks up different ages (x-axis) for a woman homozygous at 10 risk loci and in population

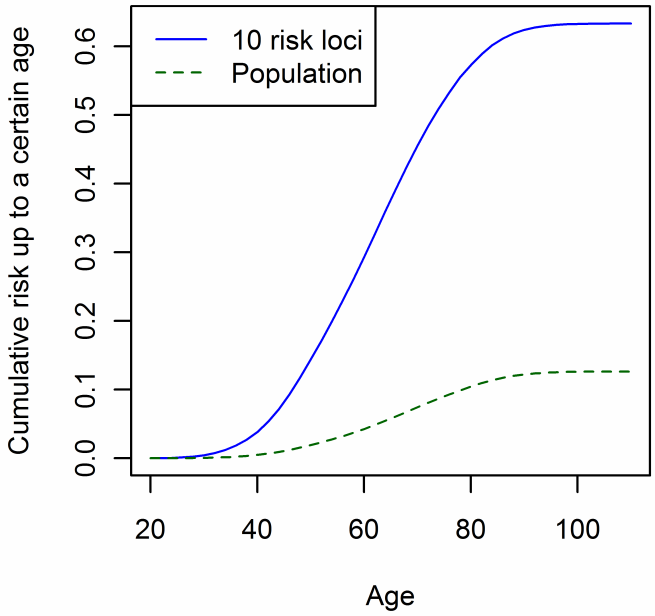


Figure 6 Residual lifetime risk for a woman homozygous at 10 risk loci and in population

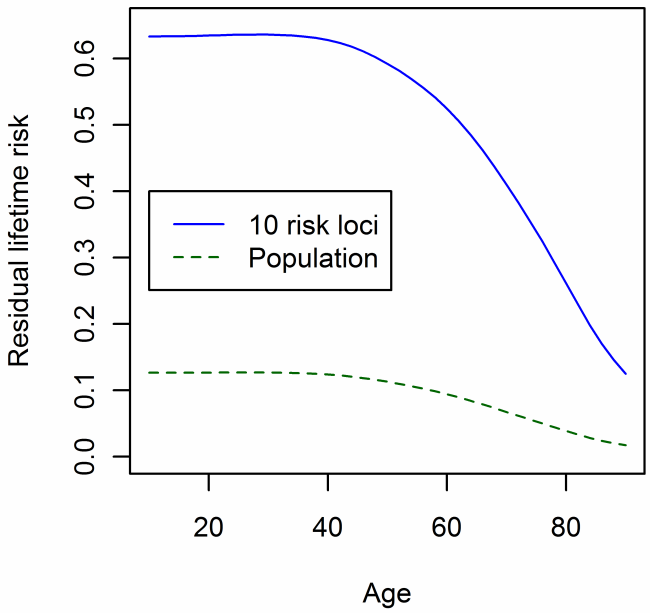
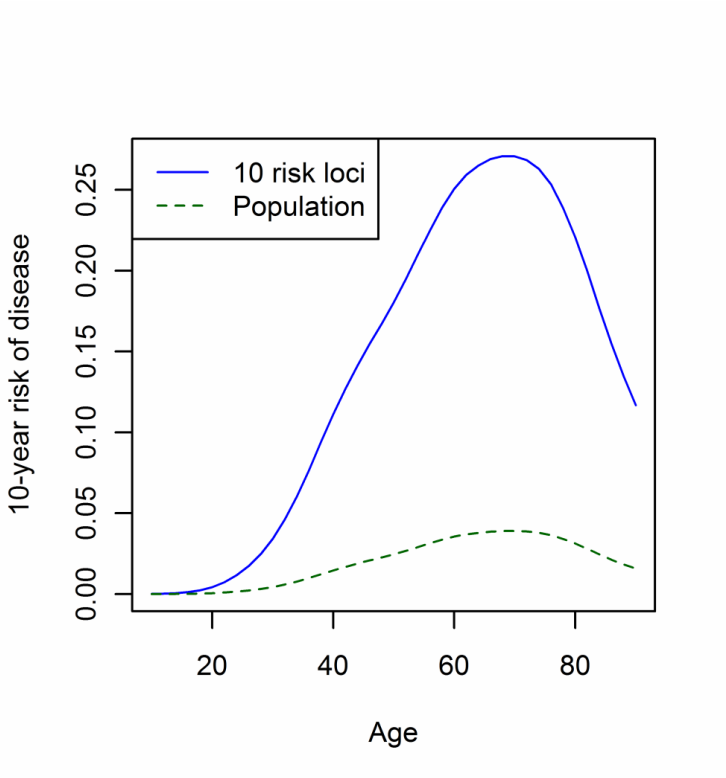


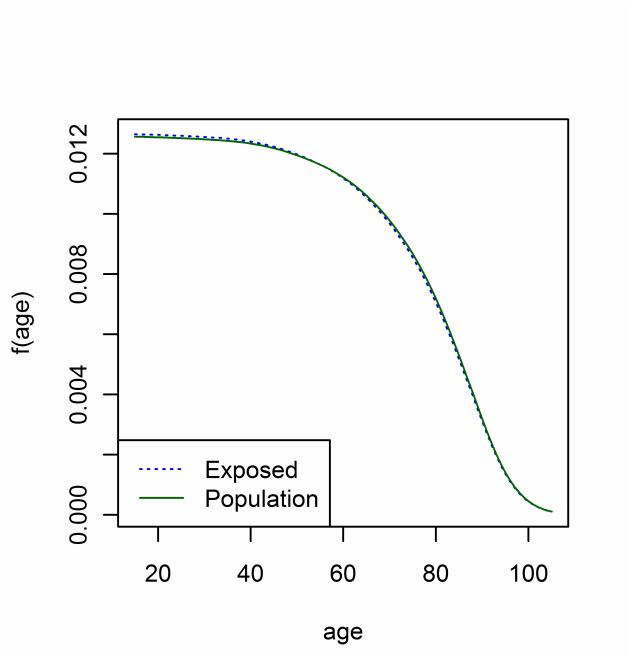
Figure 7

Ten-year risks for a woman homozygous at 10 risk loci and in population

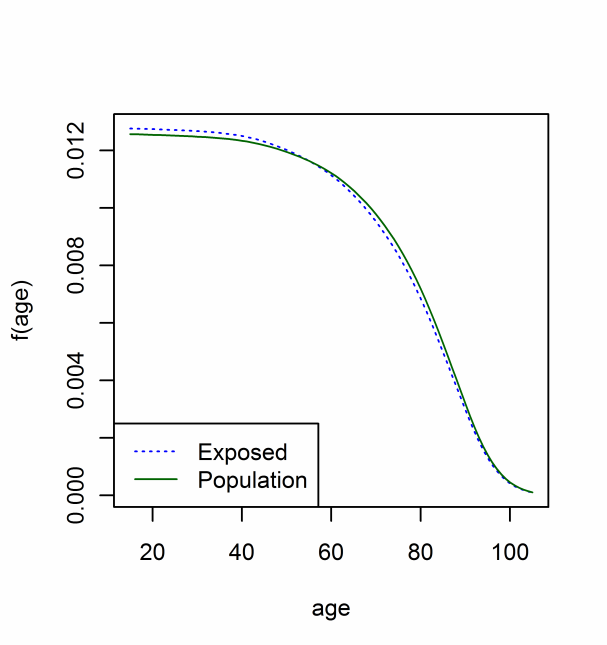


Supplementary figures

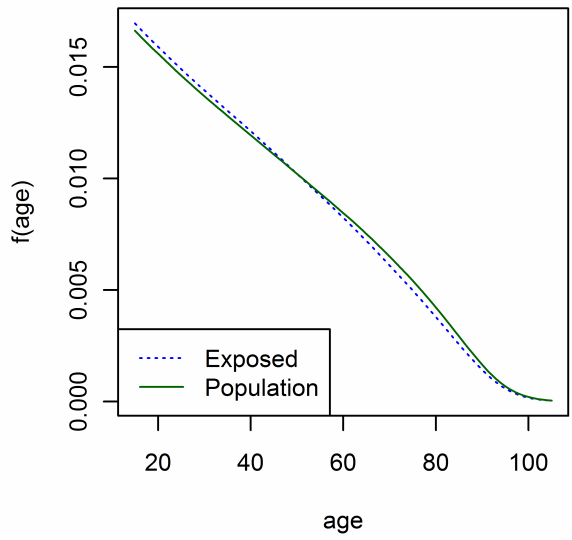
Supp Figure 1 Age distribution of disease-free individuals in the exposed group and in the general population for different intrinsic growth rates (rho) and hazard ratio(HR)



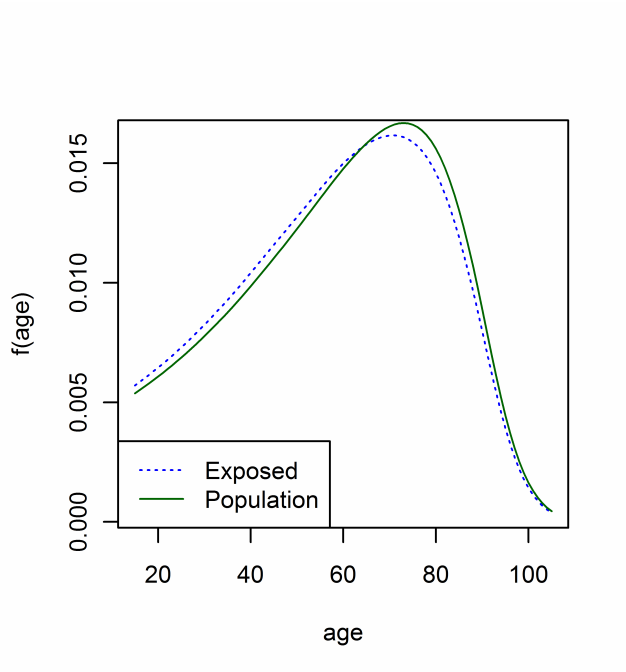
rho=0 , HR= 1.2



rho=0, HR=1.5



rho = 0.0125, HR=2



rho = -0.0125, HR=2