Theoretical Biology and Medical Modelling

BioMed Central

Research

Open Access

A model of gene-gene and gene-environment interactions and its implications for targeting environmental interventions by genotype

Helen M Wallace*

Address: GeneWatch UK, The Mill House, Tideswell, Buxton, Derbyshire, SK17 8LN, UK

Email: Helen M Wallace* - helen.wallace@genewatch.org

* Corresponding author

Published: 09 October 2006

Theoretical Biology and Medical Modelling 2006, **3**:35 doi:10.1186/1742-4682-3-35

Received: 13 April 2006
Accepted: 09 October 2006

This article is available from: http://www.tbiomed.com/content/3/1/35

© 2006 Wallace; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The potential public health benefits of targeting environmental interventions by genotype depend on the environmental and genetic contributions to the variance of common diseases, and the magnitude of any gene-environment interaction. In the absence of prior knowledge of all risk factors, twin, family and environmental data may help to define the potential limits of these benefits in a given population. However, a general methodology to analyze twin data is required because of the potential importance of gene-gene interactions (epistasis), gene-environment interactions, and conditions that break the 'equal environments' assumption for monozygotic and dizygotic twins.

Method: A new model for gene-gene and gene-environment interactions is developed that abandons the assumptions of the classical twin study, including Fisher's (1918) assumption that genes act as risk factors for common traits in a manner necessarily dominated by an additive polygenic term. Provided there are no confounders, the model can be used to implement a top-down approach to quantifying the potential utility of genetic prediction and prevention, using twin, family and environmental data. The results describe a solution space for each disease or trait, which may or may not include the classical twin study result. Each point in the solution space corresponds to a different model of genotypic risk and gene-environment interaction.

Conclusion: The results show that the potential for reducing the incidence of common diseases using environmental interventions targeted by genotype may be limited, except in special cases. The model also confirms that the importance of an individual's genotype in determining their risk of complex diseases tends to be exaggerated by the classical twin studies method, owing to the 'equal environments' assumption and the assumption of no gene-environment interaction. In addition, if phenotypes are genetically robust, because of epistasis, a largely environmental explanation for shared sibling risk is plausible, even if the classical heritability is high. The results therefore highlight the possibility – previously rejected on the basis of twin study results – that inherited genetic variants are important in determining risk only for the relatively rare familial forms of diseases such as breast cancer. If so, genetic models of familial aggregation may be incorrect and the hunt for additional susceptibility genes could be largely fruitless.

Background

Some geneticists have predicted a genetic revolution in healthcare: involving a future in which individuals take a battery of genetic tests, at birth or later in life, to determine their individual 'genetic susceptibility' to disease [1,2]. In theory, once the risk of particular combinations of genotype and environmental exposure is known, medical interventions (including lifestyle advice, screening or medication) could then be targeted at high-risk groups or individuals, with the aim of preventing disease [3].

However, there are also many critics of this strategy, who argue that it is likely to be of limited benefit to health [4-8]. One area of debate concerns the proportion of cases of a given common disease that might be avoided by targeting environmental or lifestyle interventions to those at high genotypic risk. Known genetic risk factors have to date shown limited utility in this respect [9]. However, some argue that combinations of multiple genetic risk factors may prove more useful in the future [10].

There are two possible approaches to considering this issue. The 'bottom-up' approach seeks to identify individual genetic and environmental risk factors and their interactions and quantify the risks. However, this approach is limited by the difficulties in establishing the statistical validity of genetic association studies and of quantifying gene-gene and gene-environment interactions: see, for example, [11-14].

A 'top-down' approach instead considers risks at the population level using twin and family studies and data on the importance of environmental factors in determining a trait. However, analysis of twin data is usually limited by the assumptions made in the classical twin study [15], including that: (i) there are no gene-gene interactions (epistasis); (ii) there are no gene-environment interactions; (iii) the effects of environmental factors shared by twins are independent of zygosity (the 'equal environments' assumption). These assumptions have all been individually explored and shown to be important in influencing the conclusions drawn from twin and family data [16-18]. In addition, the magnitude of any gene-environment interaction is critically important in determining the utility of targeting environmental interventions by genotype [19]. Although a general methodology to analyze twin data without making these assumptions has been developed, the algebra becomes intractable once multiple loci are involved [17]. This is problematic because, for common diseases, the impacts of multiple genetic variants, and potentially the whole genetic sequence, on disease susceptibility (here called 'genotypic risk') may be important.

The four-category model of population risks developed by Khoury and others [19] is a useful starting point for a top-down analysis of genetic prediction and prevention. It allows the merits of a targeted intervention strategy (which seeks to reduce the exposure of the high-risk genotype group only) to be explored, and can readily be extended to include more than four risk categories [10]. However, this model's use to date has been limited to bottom-up consideration of single genetic variants or to studying hypothetical examples of multiple variants. The four-category model is limited by the assumption of no confounders, which means it is applicable to only a subset of possible models of gene-gene and gene-environment interaction. However, situations where the 'no confounders' assumption is valid are arguably most likely to be of relevance to public health.

The aim of this paper is to combine the four-category model with population level data from twin, family and environmental studies, without adopting the classical twin model assumptions. This model of gene-gene and gene-environment interactions is then used to implement a 'top-down' approach to quantifying the utility of genetic 'prediction and prevention'.

Method

The four-category model

Consider a population divided into genotypic or environmental risk categories for a given trait (Figure 1a and 1b). The fraction of the population in the 'high environmental risk group' (designated by subscript e) is ε, and this subpopulation is at risk r_e. The remainder of the population is at risk r_{oe}. The fraction of the population in the 'high genotypic risk' group (designated by the subscript g) is γ, and this subpopulation is at risk r_g, with the remainder of the population at risk r_{og}. The total risk r_t for this trait in this population is then given by:

$$r_t = \gamma r_g + (1-\gamma) r_{og} \quad (1)$$

or by:

$$r_t = \varepsilon r_{\mathrm{e}} + (1-\varepsilon) r_{oe} \quad (2)$$

The same population can alternatively be divided into four categories, making a four-category model (Figure 1c)) with risks R_{oo}, R_{oe}, R_{go} and R_{ge}. Table 1 shows the risk categories in this model.

The risks are related to the previous definitions by:

$$r_g = \varepsilon R_{ge} + (1-\varepsilon)\, R_{go} \quad (3)$$

$$r_{og} = \varepsilon R_{oe} + (1-\varepsilon)\, R_{oo} \quad (4)$$

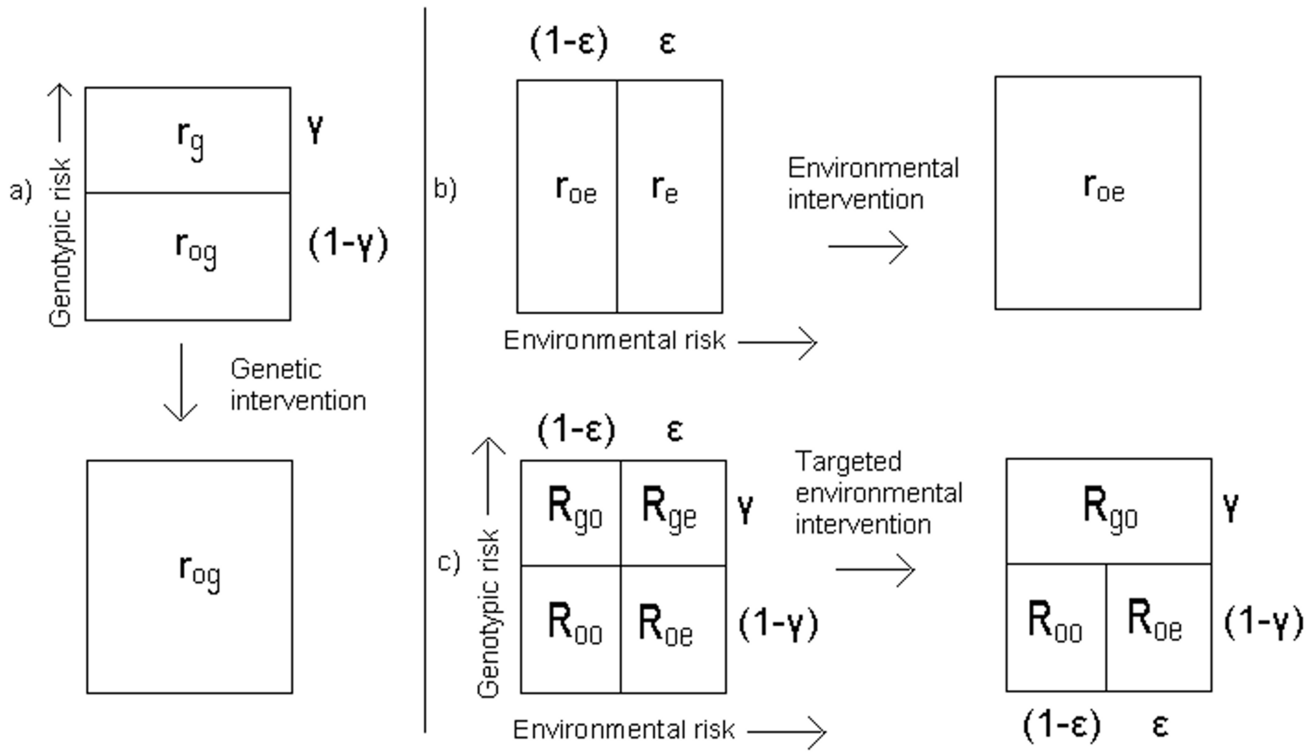

Figure 1
The four-category model. A population divided into: (a) high and low genotypic risk categories (r_g and r_{og}); (b) high and low environmental risk categories (r_e and r_{oe}); (c) four categories based on combined genotypic and environmental risk.

$r_e = \gamma R_{ge} + (1\text{-}\gamma)\, R_{oe}$ (5)

$r_{oe} = \gamma R_{og} + (1\text{-}\gamma)\, R_{oo}$ (6)

The category risks R remain constant in different populations (i.e. as ε and γ vary), provided there are no confounders. This assumption restricts the model to special cases of gene-gene and gene-environment interaction. Note that for a single genetic variant, r_g corresponds to the penetrance of the variant, and that in general (provided $R_{ge} \neq R_{go}$) this varies with the proportion of the population in the high exposure group, ε, as has been observed [20,21].

The total risk for the given trait is given by:

$r_t = \gamma\varepsilon R_{ge} + \gamma(1\text{-}\varepsilon)R_{go} + \varepsilon(1\text{-}\gamma)R_{oe} + (1\text{-}\varepsilon)(1\text{-}\gamma)R_{oo}$ (7)

The subpopulation of cases has different characteristics from the general population: for example, it contains a higher proportion of people from the 'ge' subgroup. The relative risk for a person drawn randomly from a subpopulation with the same genotypic and environmental characteristics as the cases, RR^{cases}, is given by the sum of the relative risks for each category shown in Table 1:

$$RR^{cases} = \frac{\gamma\varepsilon R_{ge}^2 + \gamma(1-\varepsilon)R_{go}^2 + \varepsilon(1-\gamma)R_{oe}^2 + (1-\varepsilon)(1-\gamma)R_{oo}^2}{r_t^2} \qquad (8)$$

Similarly, the relative risk for a person drawn randomly from a subpopulation with the same genotypic characteristics as the cases (but with the environmental characteristics of the general population) is:

$$RR_{gen}^{cases} = \frac{\gamma r_g^2 + (1-\gamma) r_{og}^2}{r_t^2} \qquad (9)$$

The relative risk for a person drawn randomly from a subpopulation with the same environmental characteristics as the cases (but with the genotypic characteristics of the general population) is:

$$RR_{env}^{cases} = \frac{\varepsilon r_e^2 + (1-\varepsilon) r_{oe}^2}{r_t^2} \qquad (10)$$

Table 1: The four category model: risks and cases for a population of size N.

Category	Risk of being in category	Number of people in category	Number of cases in category
ge (high-risk genotype/high-risk exposure)	R_{ge}	$\gamma\varepsilon N$	$\gamma\varepsilon\ R_{ge}N$
go (high-risk genotype/low-risk exposure)	R_{go}	$\gamma\ (1\text{-}\varepsilon)N$	$\gamma\ (1\text{-}\varepsilon)R_{go}N$
oe (low-risk genotype/high-risk exposure)	R_{oe}	$\varepsilon\ (1\text{-}\gamma)N$	$\varepsilon\ (1\text{-}\gamma)R_{oe}N$
oo (low-risk genotype/low-risk exposure)	R_{oo}	$(1\text{-}\varepsilon)\ (1\text{-}\gamma)N$	$(1\text{-}\varepsilon)\ (1\text{-}\gamma)R_{oo}N$
Total		N	r_tN

Population attributable fractions

Provided there are no confounders, the population attributable fraction (PAF^E_e) due to the presence of the high exposure (E) in the high exposure population subgroup (e) may be defined as:

$$PAF_e^E = \frac{\varepsilon(r_e - r_{oe})}{r_t} = \varepsilon\left\{\gamma(R_{ge} - R_{go}) + (1-\gamma)(R_{oe} - R_{oo})\right\}\big/ r_t \qquad (11)$$

If the trait is a disease, PAF^E_e is the proportion of cases that could be avoided if an environmental intervention (such as a lifestyle change or reduction in exposure) succeeds in moving everyone in the 'high environmental risk group' to the 'low environmental risk' category, as shown in Figure 1b.

The targeted population attributable fraction (PAF^E_{ge}) may be defined as the proportion of cases that could be avoided by targeting the same environmental intervention at the 'high genotypic + high environmental risk' subgroup only (the 'ge' subgroup), as shown in Figure 1c. Again assuming no confounders, it is given by:

$$PAF_{ge}^E = \varepsilon\gamma(R_{ge} - R_{go})/r_t \qquad (12)$$

Note that PAF^E_{ge} differs from PAF_{ge} as defined by Khoury & Wagener [19]. The latter implicitly assumes that both environmental and genetic risk factors are reduced and thus is inappropriate for assessing the merits of a targeted environmental intervention. PAF^E_{ge} as defined here is instead equivalent to the targeted attributable fraction (AF_T) defined by Khoury et al. [10]. To avoid confusion, the notation adopted here specifies both the nature of the intervention (environmental, denoted by superscript E) and the target subpopulation (the 'ge' subgroup, at both high genotypic and high environmental risk). Thus, the proportion of cases that would be avoided were it possible to move the 'high genotypic risk' subgroup to 'low genotypic risk' (as shown in Figure 1a) is written as PAF^G_g, given by:

$$PAF_g^G = \frac{\gamma(r_g - r_{og})}{r_t} = \gamma\left\{\varepsilon(R_{ge} - R_{oe}) + (1-\varepsilon)(R_{go} - R_{oo})\right\}\big/ r_t \qquad (13)$$

Although in practice it is not possible to change the genotype of the population, the parameter PAF^G_g is nevertheless useful in the calculations that follow.

Measures of utility

Khoury et al. [10] define the Population Impact (PI) as:

$$PI = PAF_{ge}^E \big/ PAF_e^E \qquad (14)$$

PI is one possible measure of the usefulness of targeting the environmental intervention (E) at the 'ge' subgroup. It measures the proportion of cases avoided by targeting the 'high genotypic + high environmental risk' subgroup (the 'ge' subgroup), compared to the proportion avoided by applying the environmental intervention to the whole 'high environmental risk' group. PI has the property:

$$0 \le PI \le 1 \qquad (15)$$

and has its maximum value when $PAF^E_{ge} = PAF^E_e$. However, as a measure of the utility of genotyping, PI has the disadvantage that it takes no account of the proportion of the population γ in the high genotypic risk group. This means PI = 1 when $\gamma = 1$ simply because the whole population is then in the high genotypic risk group, although using genotyping to target environmental interventions is more likely to be useful if PI = 1 and γ is also small.

Therefore, consider an alternative utility parameter U_{ge}, defined by:

$$U_{ge} = \frac{PAF_{ge}^E}{PAF_e^E} - \gamma = \frac{\gamma(1-\gamma)\left[(R_{ge} - R_{go}) - (R_{oe} - R_{oo})\right]}{\left[\gamma(R_{ge} - R_{go}) + (1-\gamma)(R_{oe} - R_{oo})\right]} \qquad (16)$$

which has the property

$$-\gamma \le U_{ge} \le (1\text{-}\gamma) \qquad (17)$$

U_{ge} tends to 1 only if PI = 1 and γ is also small. It is a measure of the utility of using genotyping to target the environmental intervention at the 'ge' subgroup, compared to randomly selecting the same proportion γ of the population to receive the intervention. U_{ge} is positive if those at high genotypic risk have *more to gain* than those at low genotypic risk from the intervention ((R_{ge}-R_{go}) ≥ (R_{oe}-R_{oo})) and negative if they have *less to gain* from the intervention. This reflects the fact that targeting those who have least to gain through an intervention is worse than using random selection in terms of its impact on population health.

Note that even if genotyping is better than random selection, other types of test that are more useful may be available [22]; a population-based approach still has the potential to reduce more cases of disease [9,19,23]; and such targeting also has broader psychological and social implications. Therefore a positive U_{ge} does not necessarily imply that genotyping is the best means of selecting a subpopulation to target, or that a targeted approach is necessarily effective or socially acceptable. Note also that the measure U_{ge} applies only to interventions that are considered applicable to the whole population (such as smoking cessation) and neglects other relevant issues such as cost-effectiveness and the burden of disease [24]. In addition, it is necessary to consider the magnitude of the Population Attributable Fraction, PAF^E_e before proposing this approach. This is because both PI and U_{ge} may tend to unity even if only a small proportion of cases can be avoided by means of environmental interventions.

Limits on parameters

Consider only populations where $r_g \geq r_{og}$ and $r_e \geq r_{oe}$ for all values of ε and γ. Then the risks in the four box model must be ordered such that:

$$1 \geq R_{ge} \geq R_{oe} \geq R_{oo} \geq 0 \qquad (18)$$

and

$$R_{ge} \geq R_{go} \geq R_{oo} \qquad (19)$$

Using the known relationships (Equations (11), (13) and (16)) between PAF^E_e, PAF^G_g, U_{ge} and the risks R_{oo}, R_{go}, R_{oe} and R_{ge}, leads to the limits on the utility parameter U_{ge} shown in Table 2. These conditions also ensure that PAF^E_e, PAF^G_g and PAF^E_{ge} are all positive. The two remaining inequalities ($R_{ge} \leq 1$ and $R_{oo} \geq 0$) are considered later, where they are used to derive limits on the proportion of the population in the 'high genotypic risk' group, γ. This step is not possible at this stage because PAF^E_e, PAF^G_g and PAF^E_{ge} are themselves dependent on γ.

The twin and familial risks model

Data from studies of monozygotic and dizygotic twins are commonly used to estimate the genetic and environmental variances V_g and V_e of a trait. Here, the aim is to use twin and other data to estimate the possible magnitudes of the population attributable fractions and measures of utility defined above. To do this it is necessary to estimate V_g, V_e and the variance due to gene-environment interaction, V_{ge}. The standard methodology for twin data analysis is inappropriate because it assumes V_{ge} = 0.

First note that we are interested in the extent to which relatives share *risk categories* (which may be either environmental or genotypic, or both), rather than a particular genetic variant. The probability that a relative of a proband is also a case depends on the extent to which their environmental and genotypic risks are correlated with those of the proband. Rather than adopting a specific form for the genetic model, define p^{rel}_g as the correlation in genotypic risk category (g) between relatives of type denoted by the superscript 'rel'. The parameter p^{rel}_g is the probability that the genotypic risk category (high or low) is identical by descent.

For monozygotic (MZ) twins, assumed to share their entire genome, p^{MZ}_g = 1. For dizygotic (DZ) twins and other siblings, who share half their genome, $p^{DZ}_g = p^{sib}_g$ = 1/2 for a single allele model (dominant Mendelian disorder) or an additive polygenic model. For a two allele model (recessive Mendelian disorder) or the dominance term of a polygenic model (in which multiple pairs of alleles interact), $p^{DZ}_g = p^{sib}_g$ = 1/4. Here, allowing for the possibility of multiple gene-gene interactions (epistasis), require only that:

$$1/2 \geq p_g^{DZ} \geq 0 \qquad (20)$$

The meaning of p^{DZ}_g and its relationship to the polygenic risk model first adopted by Ronald Fisher in 1918 is discussed further below.

Similarly, define p^{rel}_e as the correlation in environmental risk category (e) between relatives of type "rel", requiring only that:

$$1 \geq p_e^{rel} \geq 0 \qquad (21)$$

Assume that p^{rel}_g and p^{rel}_e are independent (so that there is no genotype-environment correlation) and that risks within a category are randomly distributed. The relative risk for a relative of type "rel" may then be written:

$$\lambda_{rel} = (1-p_g^{rel})(1-p_e^{rel}) + p_g^{rel}(1-p_e^{rel})RR_{gen}^{cases} + (1-p_g^{rel})p_e^{rel}RR_{env}^{cases} + p_g^{rel}p_e^{rel}RR^{cases} \qquad (22)$$

Substituting for the relative risks RR^{cases}_{gen}, RR^{cases}_{env} and RR^{cases} using Equations (8), (9) and (10) leads (after some algebra) to:

$$\lambda_{rel} - 1 = p_g^{rel}\frac{V_g}{r_t^2} + p_e^{rel}\frac{V_e}{r_t^2} + p_g^{rel}p_e^{rel}\frac{V_{ge}}{r_t^2} \qquad (23)$$

where

$$\frac{V_e}{r_t^2} = \frac{(1-\varepsilon)}{\varepsilon}\left[PAF_e^E\right]^2 \qquad (24)$$

$$\frac{V_g}{r_t^2} = \frac{(1-\gamma)}{\gamma}\left[PAF_g^G\right]^2 \qquad (25)$$

$$\frac{V_{ge}}{r_t^2} = \frac{(1-\varepsilon)}{\varepsilon\gamma(1-\gamma)}\left[U_{ge}PAF_e^E\right]^2 \qquad (26)$$

Note that if the G-E interaction component of the variance, V_{ge}, is zero, the utility of targeting the environmental intervention by genotype, U_{ge}, is also zero (Equation (26)), because those at high genotypic risk have no more to gain from the intervention than those at low genotypic risk (R_{ge}-R_{go} = R_{oe}-R_{oo}).

Equation (23) can also be derived more formally using matrix methods (Appendix A).

The gene-environment interaction factor and remaining inequalities

Without loss of generality, define the gene-environment interaction factor f_{ge} such that:

$$\frac{V_{ge}}{r_t^2} = f_{ge}^2 \frac{V_g}{r_t^2}.\frac{V_e}{r_t^2} \qquad (27)$$

and choose its sign so that (combining Equations (24), (25) and (26)):

$$U_{ge} = f_{ge}\sqrt{\gamma(1-\gamma)\frac{V_g}{r_t^2}} \qquad (28)$$

U_{ge} is zero if $f_{ge} = 0$ (i.e. for an additive G-E model, with no G-E interaction), but for a given γ and V_g, U_{ge} increases with increasing gene-environment interaction factor, f_{ge}. For a fixed f_{ge} and genetic variance component V_g, U_{ge} is maximum when $\gamma = 1/2$, i.e. when half the population is in the high genotypic risk group, provided solutions with $\gamma = 1/2$ exist (see also below: *cases where* $\gamma_{\text{maxge}} < 1/2$).

Using the definitions of V_e, V_g and V_{ge} (Equations (24), (25) and (26)) and the remaining inequalities, $R_{ge} \leq 1$ and $R_{oo} \geq 0$, two limits can be derived on the proportion of the population in the 'high genotypic risk' group, γ (see Table 2).

Scoping studies

The general system of equations represented by Equation (23) may be simplified where data exist from monozygotic twins, dizygotic twins and other siblings, such that $\lambda_{DZ} > \lambda_{sib}$. This implies that environmental risks are more strongly correlated in dizygotic twins than in other siblings, $p^e_{DZ} > p^e_{sib}$. Remembering that $p^{MZ}_g = 1$ and $p^{sib}_g = p^{DZ}_g$, three independent equations for the relative risk in monozygotic, dizygotic twins and siblings may then be written:

$$\lambda_{MZ} - 1 = \frac{V_g}{r_t^2} + p_e^{MZ}\frac{V_e}{r_t^2} + p_e^{MZ}\frac{V_{ge}}{r_t^2} \qquad (29)$$

$$\lambda_{DZ} - 1 = p_g^{DZ}\frac{V_g}{r_t^2} + p_e^{DZ}\frac{V_e}{r_t^2} + p_g^{DZ}p_e^{DZ}\frac{V_{ge}}{r_t^2} \qquad (30)$$

$$\lambda_{sib} - 1 = p_g^{DZ}\frac{V_g}{r_t^2} + p_e^{sib}\frac{V_e}{r_t^2} + p_g^{DZ}p_e^{sib}\frac{V_{ge}}{r_t^2} \qquad (31)$$

To solve, assume the recurrence risks λ are known (see Appendix B and [25]) and define:

$$R_{MD} = \frac{\lambda_{MZ} - 1}{\lambda_{DZ} - 1} \qquad (32)$$

$$R_{SD} = \frac{\lambda_{sib} - 1}{\lambda_{DZ} - 1} \qquad (33)$$

with

$$R_{MD} \geq 1 \qquad (34)$$

and

$$0 \leq R_{SD} \leq 1. \qquad (35)$$

Note that if $R_{SD} = 1$, Equations (30) and (31) are identical, $p^e_{DZ} = p^e_{sib}$, and more relatives are needed to obtain solutions, except in the special case where there is no environmental variance (see below: *no environmental variance*).

In addition, define the variable parameters (assumed unknown):

$$c_{MD} = \frac{p_e^{MZ}}{p_e^{DZ}} \qquad (36)$$

$$c_{SD} = \frac{p_e^{sib}}{p_e^{DZ}} \qquad (37)$$

with

$$c_{MD} \geq 1 \qquad (38)$$

Table 2: Constraints on model parameters

Condition	Limits on U_{ge}	Limits on γ	Limits on p^{DZ}_g	Limits on f_{ge}
$R_{oe} \geq R_{oo}$	$U_{ge} \leq (1 - \gamma)$	$\gamma \leq \gamma_{\max ge}$ where $\gamma_{\max ge} = \frac{1}{1+\frac{V_{ge}}{V_e}}$		
$R_{go} \geq R_{oo}$	$U_{ge} \leq (1-\gamma)\frac{PAF_g^G}{PAF_e^E}$		$p_g^{DZ} \leq p_{g\max}^{DZ}$	$f_{ge} \leq \frac{1}{PAF_e^E}$
$R_{ge} \geq R_{go}$	$U_{ge} \geq -\gamma$	$\gamma \geq \gamma_{neg}$ where $\gamma_{neg} = \frac{1}{1+\frac{V_e}{V_{ge}}}$		
$R_{ge} \geq R_{oe}$	$U_{ge} \geq -(1-\gamma)\frac{\varepsilon PAF_g^G}{(1-\varepsilon)PAF_e^E}$		$p_g^{DZ} \leq p_{gneg}^{DZ}$	$f_{ge} \geq -\frac{\varepsilon}{(1-\varepsilon)PAF_e^E}$
$R_{ge} \leq 1$		$\gamma \geq \gamma_{\min ge}$ where $\gamma_{\min ge} = \frac{1}{1+\frac{F_1^2}{(V_g / r_t^2)}}$		
$R_{oo} \geq 0$		$\gamma \leq \gamma_o$ where $\gamma_o = \frac{1}{1+F_2^2(V_g / r_t^2)}$		

and

$0 \leq c_{SD} \leq 1.$ (39)

For $\lambda_{DZ} > 1$ and $R_{SD} < 1$ the simultaneous Equations (29), (30) and (31) can then be solved to give:

$$\frac{V_g}{r_t^2} = \frac{(\lambda_{DZ} - 1)}{p_g^{DZ}} \cdot \frac{(R_{SD} - c_{SD})}{(1 - c_{SD})} \qquad (40)$$

$$\frac{V_e}{r_t^2} = \frac{(\lambda_{DZ} - 1)}{p_e^{DZ} c_{MD}(1 - p_g^{DZ})}\left[\frac{(c_{MD} - 1)(1 - R_{SD})}{(1 - c_{SD})} + (1 - p_g^{DZ} R_{MD})\right] \qquad (41)$$

$$\frac{V_{ge}}{r_t^2} = \frac{(\lambda_{DZ} - 1)}{p_e^{DZ} p_g^{DZ} c_{MD}(1 - p_g^{DZ})}\left[\frac{(1 - c_{MD} p_g^{DZ})(1 - R_{SD})}{(1 - c_{SD})} + (1 - p_g^{DZ} R_{MD})\right] \qquad (42)$$

provided $p_g^{DZ} \neq 0$, $p_e^{DZ} \neq 0$ and $c_{SD} \neq 1$ (see also below).

For situations in which a targeted intervention is under consideration, the population attributable fraction PAF^E_e and exposure ε are likely to be known, allowing V_e to be treated as an input variable. However, p^{DZ}_e is usually unknown, since environmental correlations are often difficult to measure. Therefore, it is useful to eliminate p^{DZ}_e from Equations (41) and (42), leading to:

$$\frac{V_{ge}}{V_e} = \frac{\left\{\frac{p_g^{DZ}}{p_{g\min}^{DZ}} - 1\right\}\frac{(R_{SD} - c_{SD})}{(1 - c_{SD})}}{p_g^{DZ} R_{MD}(p_{gtop}^{DZ} - p_{g\min}^{DZ})} \qquad (43)$$

where

$$p_{gtop}^{DZ} = \frac{1}{R_{MD}}\left\{1 + \frac{(c_{MD} - 1)(1 - R_{SD})}{(1 - c_{SD})}\right\} \qquad (44)$$

and

$$p^{DZ}_{g\min} = \frac{(R_{SD} - c_{SD})}{\{R_{MD}(1 - c_{SD}) - c_{MD}(1 - R_{SD})\}} \qquad (45).$$

Equations (27), (40) and (43) allow the gene-environment interaction factor f_{ge} to be written as:

$$f^2_{ge} = \frac{\left\{\dfrac{p^{DZ}_{g}}{p^{DZ}_{g\min}} - 1\right\}}{(\lambda_{DZ} - 1)R_{MD}(p^{DZ}_{gtop} - p^{DZ}_{g})} \qquad (46).$$

The parameter p^{DZ}_{g}, which defines the form of the genetic model, is then given by:

$$\frac{p^{DZ}_{g}}{p^{DZ}_{g\min}} = \frac{1 + f^2_{ge}(\lambda_{DZ} - 1)R_{MD}p^{DZ}_{gtop}}{1 + f^2_{ge}(\lambda_{DZ} - 1)R_{MD}p^{DZ}_{g\min}} \qquad (47).$$

For known R_{MD}, R_{SD} and λ_{DZ} a solution space can now be mapped, which includes all possible variances consistent with the data and with the inequalities derived above.

Requiring the variances to be positive leads to the additional conditions on p^{DZ}_{g} and c_{SD} shown in Table 3.

The limits on U_{ge} shown in Table 2 set limits on the range of gene-environment interaction models such that:

$$-\frac{\varepsilon}{(1-\varepsilon)PAF^{E}_{e}} \le f_{ge} \le \frac{1}{PAF^{E}_{e}} \qquad (48)$$

Noting that $f_{ge} = 0$ corresponds to $p^{DZ}_{g} = p^{DZ}_{g\min}$ (Equation (64)), this implies that, for $U_{ge} \ge 0$, the solution space may be defined by:

$$p^{DZ}_{g\min} \le p^{DZ}_{g} \le p^{DZ}_{g\max} \qquad (49)$$

where $p^{DZ}_{g\max}$ is given by Equation (47) with $f_{ge} = 1/PAF^{E}_{e}$.

For $U_{ge} \le 0$, the solution space may be defined by:

$$p^{DZ}_{g\min} \le p^{DZ}_{g} \le p^{DZ}_{gneg} \qquad (50)$$

where p^{DZ}_{gneg} is given by Equation (47) with $f_{ge} = -\varepsilon/(1-\varepsilon)PAF^{E}_{e}$.

The remaining limits on U_{ge} lead to the additional conditions on the range of γ values (the proportion of the population in the high risk group) shown in Table 2. These conditions on γ may be written:

$$\gamma_{\min} \le \gamma \le \gamma_{\max} \qquad (51)$$

where (noting that $\gamma_{maxge} = \gamma_{o}$ when $f_{ge} = 1$):

$$\gamma_{\max} = \begin{cases} \gamma_{\max ge} \ \text{for} \ f_{ge} \ge 1 \\ \gamma_{0} \ \text{for} \ f_{ge} \le 1 \end{cases} \qquad (52)$$

and (noting that $\gamma_{minge} = \gamma_{neg}$ when $f_{ge} = -r_t/(1-r_t)$):

$$\gamma_{\min} = \begin{cases} \gamma_{\min ge} \ \text{for} \ f_{ge} \ge -r_t/(1-r_t) \\ \gamma_{neg} \ \text{for} \ f_{ge} \le -r_t/(1-r_t) \end{cases} \qquad (53)$$

Two transition lines can therefore be defined such that $p^{DZ}_{g} = p^{DZ}_{gt}$ when $f_{ge} = 1$ and $p^{DZ}_{g} = p^{DZ}_{gnegt}$ when $f_{ge} = -r_t/(1-r_t)$. The values of p^{DZ}_{gt} and p^{DZ}_{gnegt} may be calculated using Equation (47).

The full range of gene-environment interaction models specified by f_{ge} (within the limits given by Equation (48)) and the corresponding range of γ values are summarized in Table 4. Note that the risk distribution associated with $f_{ge} = 1$ corresponds to a multiplicative model of gene-environment interaction. If $f_{ge} \ge 1$ solutions with population impact PI = 1 may exist (i.e. with $PAF^{E}_{ge} = PAF^{E}_{e}$), provided the proportion of the population in the high risk genotypic group takes the maximum value consistent with the data ($\gamma = \gamma_{maxge}$). For lower values of f_{ge}, solutions with PI = 1 cannot exist.

One additional condition is necessary for solutions to exist, namely:

$$\gamma_{\max} \ge \gamma_{\min} \qquad (54)$$

This condition is always met if

$$\lambda_{MD} \le \gamma_{e} + 1 \qquad (55)$$

where

$$\gamma_{e} = \begin{cases} F_1/f_{ge} \ \text{for} \ f_{ge} \ge 1 \\ F_1/F_2 \ \text{for} \ 1 \ge f_{ge} \ge -r_t/(1-r_t) \\ -F_2/f_{ge} \ \text{for} \ f_{ge} \le -r_t/(1-r_t) \end{cases} \qquad (56)$$

and F_1 and F_2 are given by:

$$F_1 = \frac{\left[\left(\dfrac{1-r_t}{r_t}\right) - \left(\dfrac{1-\varepsilon}{\varepsilon}\right)PAF^{E}_{e}\right]}{\left[1 + f_{ge}\left(\dfrac{1-\varepsilon}{\varepsilon}\right)PAF^{E}_{e}\right]} = \frac{(1-r_e)}{\left[r_t + f_{ge}(r_e - r_t)\right]} \qquad (57)$$

$$F_2 = \frac{\left(1 - PAF_e^E\right)}{\left(1 - f_{ge} PAF_e^E\right)} \qquad (58).$$

However, if λ_{MD} is greater than this, the requirement $\gamma_{max} \geq \gamma_{min}$ further restricts the values of c_{SD} that lie within the solution space (Table 3).

If V_e and ε are known, a solution space can be now be mapped for p^{DZ}_g and f_{ge} with known input data from twin and sibling studies (λ_{MZ}, λ_{DZ} and λ_{sib}), for a given c_{MD} and all values of c_{SD} within the assumed range. The boundaries of the solution space are determined by the limits on f_{ge} given by Equation (48), the condition $\gamma_{max} \geq \gamma_{min}$ (Equation (54)), and the requirement that p^{DZ}_g is less than or equal to 1/2 (Equation (20)) – no other condition on the genetic model is specified a priori. For each genetic risk model and gene-environment interaction model in the solution space, defined by p^{DZ}_g and f_{ge} respectively, the variances V_g and V_{ge} can then be calculated, as can γ_{max} and γ_{min}. For a chosen γ value in the allowed range, U_{ge} can then be calculated from Equation (28).

The model code is available as [Additional file 1: heritability12.xls].

Note that the condition on $p^{DZ}_g \leq 1/2$ may also be rewritten using Equation (47), so that:

$$p_g^{DZ} \leq 1/2 \Rightarrow \frac{\left(p_{g\min}^{DZ} - \frac{1}{2}\right)}{p_{g\min}^{DZ}} \leq R_{MD} f_e^2 \left(\lambda_{DZ} - 1\right)\left(1/2 - p_{gtop}^{DZ}\right) \qquad (59)$$

which is always met if

$$p_{gtop}^{DZ} \leq 1/2 \qquad (60).$$

Before mapping the solution space, first consider some special cases and a comparison of the model with the classical twin studies approach.

Special cases

1. No genetic variance

If $V_g = 0$, Equation (27) implies that $V_{ge} = 0$ also. Equations (29), (30) and (31) then give:

$$R_{SD} = c_{SD} \qquad (61)$$

and

$$R_{MD} = c_{MD} \qquad (62)$$

Under the usual assumption that $c_{MD} = 1$ (the 'equal environments' assumption), this is the well-known result that genetic variance can be zero only when the concordance in monozygotic and dizygotic twins is the same (leading to $R_{MD} = 1$). However, if the equal environments assumption is not met ($c_{MD} > 1$), values of R_{MD} greater than 1 do not necessarily imply that a genetic component to the variance exists (see, for example, [18]).

2. No environmental variance

If $V_e = 0$, Equation (27) implies that $V_{ge} = 0$ also. Equations (29), (30) and (31) then give:

$$R_{SD} = 1 \qquad (63)$$

and

$$R_{MD} = 1/p_g^{DZ} \qquad (64)$$

Table 3: Further constraints on model parameters

Condition	Limits on p^{DZ}_g	Limits on c_{SD}
$V_e \geq 0$	$p_g^{DZ} \leq p_{gtop}^{DZ}$	
$V_{ge} \geq 0$	$p_g^{DZ} \geq p_{g\min}^{DZ}$	
$V_g \geq 0$		$c_{SD} \leq R_{SD}$
$\gamma_{max} \geq \gamma_{min}$		If $\lambda_{MD} > \gamma_e + 1$ require: $c_{SD} \geq c_{SDm}$ where $c_{SDm} = 1 - \frac{(\lambda_{DZ}-1)(1-R_{SD})\left[c_{MD} + f_{ge}^2(\lambda_{DZ}-1)R_{MD} + \gamma_e f_{ge}^2 (c_{MD}-1)\right]}{\left[1+f_{ge}^2(\lambda_{DZ}-1)\right]\left[(\lambda_{DZ}-1)R_{MD} - \gamma_e\right]}$

Table 4: Limits on the gene-environment interaction factor (f_{ge}) and the proportion of the population in the high-genotypic risk group (γ).

Gene-environment interaction model	Interaction factor f_{ge}	Risk distribution		Utility U_{ge}	Fraction of population at high genotypic risk	
					Maximum γ_{max}	Minimum γ_{min}
Genetic effect in high-exposure group only	$1/PAF^E_e$	$\mathbf{R_{00}}$	$\mathbf{R_{ge}}$	Positive	γ_{maxge} (where $PAF^E_{ge} = PAF^E_e$; $PI = 1$; and $U_{ge} = 1-\gamma$).	γ_{minge} (where $R_{ge} = 1$).
		$\mathbf{R_{00}}$	$\mathbf{R_{0e}}$			
Multiplicative	1	$\mathbf{R_{g0}}$	$\mathbf{R_{g0}R_{0e}/R_{00}}$		$\gamma_{maxge} = \gamma_0$ (where $PAF^E_{ge} = PAF^E_e$; $R_{00} = 0$; and $PAF^G_g = 1$).	
		$\mathbf{R_{00}}$	$\mathbf{R_{0e}}$			
Additive	0	$\mathbf{R_{g0}}$	$\mathbf{R_{g0}+R_{0e}-R_{00}}$	Zero	γ_0 (where $R_{00} = 0$).	
		$\mathbf{R_{00}}$	$\mathbf{R_{0e}}$			
Reverse multiplicative	$-r_t/(1-r_t)$	$\mathbf{R_{g0}}$	$\mathbf{(1-R_{g0})\ (1-R_{0e})/(1-R_{00})}$	Negative		$\gamma_{neg} = \gamma_{minge}$ (where $PAF^E_{ge} = 0$ and $R_{ge} = 1$)
		$\mathbf{R_{00}}$	$\mathbf{R_{0e}}$			
Genetic effect in low-exposure group only	$-\varepsilon/(1-\varepsilon)PAF^E_e$	$\mathbf{R_{g0}}$	$\mathbf{R_{0e}}$			γ_{neg} (where $PAF^E_{ge} = 0$ and $PI = 0$).
		$\mathbf{R_{00}}$	$\mathbf{R_{0e}}$			

For a purely genetic model with no environmental variance, Equation (64) implies that if $R_{MD} > 2$, $p^{DZ}_g < 1/2$. This is consistent with Risch's finding [16] that neither an additive genetic model nor a single dominant gene model (both with $p^{DZ}_g = 1/2$) can fit the data for conditions such as schizophrenia (which has an R_{MD} value significantly greater than 2).

3. Classical twin study assumptions

Assuming no gene-environment interaction ($V_{ge} = 0$); an additive genetic risk model ($p^{DZ}_g = 1/2$); and the 'equal environments' assumption ($c_{MD} = 1$) in Equations (29), (30) and (31) gives:

$$\frac{V_g}{r_t^2} = 2(\lambda_{MZ} - \lambda_{DZ}) \qquad (65)$$

This is the classical twin study result, assuming the dominance term of the genetic variance is negligible. Note that, if $R_{MD} = 2$, the classical solution implies that the environmental variance terms in Equations (29) to (31) are zero and shared sibling risk is due to entirely to shared genes.

4. No correlation in genotypic risk in siblings ($p^{DZ}_g = 0$)

Equation (20) allows p^{DZ}_g to tend to zero. Substituting $p^{DZ}_g = 0$ in Equations (29), (30) and (31) and using the definition of the gene-environment interaction factor (Equation (28)) gives:

$$R_{SD} = c_{SD} \qquad (66)$$

and

$$\frac{V_g}{r_t^2} = \frac{(\lambda_{DZ} - 1)(R_{MD} - c_{MD})}{\left[1 + f_{ge}^2 c_{MD}(\lambda_{DZ} - 1)\right]} \qquad (67)$$

Note that, from Equations (30) and (31), $p^{DZ}_g = 0$ corresponds to a purely environmental explanation for shared sibling risks (although there may remain a genetic component to shared risks in monozygotic twins, from Equation (29)). The solution $p^{DZ}_g = 0$ may not exist in reality; however, the solution at this limit is of interest because low values of p^{DZ}_g are plausible.

Also, note that if $f_{ge} = 0$ (no gene-environment interaction) and $c_{MD} = 1$ (the 'equal environments' assumption), the genetic variance V_g given by Equation (67) is half the classical twin study result (Equation (65)).

5. Cases where $\gamma_{max} = \gamma_{min}$

If the line $\gamma_{max} = \gamma_{min}$ exists within the solution space, some special cases may arise with risk distributions of particular interest (including, for example, a solution with $R_{ge} = 1$ and all other risks zero). These special cases and the conditions that they meet are shown in Table 5.

6. Cases where $\gamma_{maxge} < 1/2$

Equation (27) shows that for a fixed gene-environment interaction factor f_{ge} and genetic variance component V_g, the utility U_{ge} is maximum when $\gamma = 1/2$, i.e. when half the population is in the high genotypic risk group, provided this solution exists. However, if $\gamma_{max} < 1/2$, utility is maximum when $\gamma = \gamma_{max}$. As a smaller proportion of the population is then targeted, these solutions are of particular interest. Because solutions with population impact PI = 1 may exist when $1 \le f_{ge} \le 1/PAF^E_e$ if $\gamma = \gamma_{maxge}$ (Table 4), it is of interest to identify the area of the solution space with

$\gamma_{maxge} < 1/2$. Maximum utility is then obtained when $\gamma = \gamma_{maxge}$ (where PI = 1 and $U_{ge} = 1-\gamma_{maxge}$). For the condition

$$\gamma_{\max ge} < 1/2 \Rightarrow p_g^{DZ} > p_{gx}^{DZ} \qquad (68)$$

where p^{DZ}_{gx} is given by:

$$R_{MD}(1-c_{SD})(p_{gx}^{DZ})^2 + [(1-c_{SD})(R_{MD}-1)-(2c_{MD}-1)(1-R_{SD})]p_{gx}^{DZ} - (R_{SD}-c_{SD}) = 0 \qquad (69)$$

solving for p^{DZ}_{gx} allows the region of the solution space where $\gamma_{maxge} < 1/2$ to be defined.

7. Cases where the 'equal environments' assumption holds (c_{MD} = 1)

In the special case where the 'equal environments' assumption holds ($c_{MD} = 1$, and hence $p^{DZ}_{gtop} = 1/R_{MD}$), Equation (63) simplifies to give $R_{MD} \geq 2$. Equation (62) also simplifies to give:

$$p_g^{DZ} \leq 1/2 \Rightarrow c_{SD} \geq c_1 \qquad (70)$$

where

$$c_1 = 1 - \frac{(1-R_{SD})\left[1+f_{ge}^2(\lambda_{DZ}-1)(2-R_{MD})\right]}{(2-R_{MD})\left[1+f_{ge}^2(\lambda_{DZ}-1)\right]} \qquad (71)$$

Meeting the condition $p^{DZ}_g \leq 1/2$ at $c_{SD} = 0$ then requires:

$$R_{MD} \geq 2 - \frac{(1-R_{SD})}{\left[1+f_{ge}^2(\lambda_{DZ}-1)R_{SD}\right]} \qquad (72).$$

It follows that if $c_{MD} = 1$, solutions with $p^{DZ}_g = 1/2$ (an additive genetic model) and positive utility exist only when the following condition holds for R_{MD}:

$$R_{MD} \leq 2 - \frac{(1-R_{SD})}{\left[1+(\lambda_{DZ}-1)R_{SD}/(PAF_e^E)^2\right]} \qquad (73).$$

Further, all three classical twin study assumptions (c_{MD} = 1, $p^{DZ}_g = 1/2$ and $f_{ge} = 0$) can be met only for values of R_{MD} that are low enough to satisfy:

$1 + R_{SD} \geq R_{MD} > 1$ (74).

If R_{MD} lies within this range, the classical twin study gives one possible solution; however, other solutions also exist. All alternative solutions favour a less 'genetic' and more 'environmental' explanation for shared sibling risks (i.e. they have higher values of c_{SD}). If R_{MD} is greater than $1+R_{SD}$, all three assumptions of the classical twin study cannot be met simultaneously.

Comparison with the classical twins approach

Table 6 summarizes the differences between the classical twin studies approach and the method adopted here.

A central feature of the model is that it abandons Fisher's assumption [26] that genes act as risk factors for common traits in a manner necessarily dominated by an additive polygenic term. In his historic 1918 paper, Fisher synthesized Mendelian inheritance with Darwin's theory of evolution by showing that the genetic variance of a continuous trait could be decomposed into additive and non-additive components [26,27]. Following Fisher, the classical twin study analysis depends on writing the genetic component of a trait as a convergent series of terms, consisting of an additive term (the sum of contributions of individual alleles at each locus) plus a smaller dominance term (the sum of contributions from pairs of alleles at each locus) and – usually neglected – epistatic terms (involving potentially multiple interactions between alleles at multiple loci) [15]. Often the additive term is assumed to dominate the series (equivalent to assuming $p^{DZ}_g = 1/2$).

Fisher saw his polygenic model as "*abandon* [ing] *the strictly Mendelian mode of inheritance, and treat* [ing] *Galton's 'particulate inheritance' in almost its full generality*" [26]. However, it can be argued that Fisher's model is flawed in so far as it fails to distinguish between the function of alleles and the properties of traits [4,28]. In particular, epistasis (although referred to here as 'gene-gene interaction') is not strictly an interaction between genes, but can be shown to depend on the structure and interdependence of metabolic pathways [28].

The alternative model adopted here is based on correlations in *risk categories* for a trait (which may be either environmental or genetic, or both), rather than single or multiple genetic variants. Adopting Porteous' critique [28], there is no *a priori* biological reason why the parameter p^{DZ}_g (the probability that the genotypic risk category of a dizygotic twin pair is identical by descent) cannot take any value between 1/2 (its value if the additive model holds) and zero. Low p^{DZ}_g can then be understood to mean either a situation in which Fisher's polygenic model [26] is dominated by negative (synergistic) epistatic terms (for example, $p^{DZ}_g = 1/2^n$ implies that interactions between n deleterious alleles are necessary to produce a phenotypic effect), or, more meaningfully, a situation in which human phenotypes are *biologically robust* to individual genetic variants [29]. Thus, in the extreme case where numerous genetic variants combine to influence a trait through the interdependence of metabolic pathways, the trait may be highly correlated in monozygotic twins (who share all the genetic variants) but not correlated at all ($p^{DZ}_g = 0$) in dizygotic twins or siblings (who share only

Table 5: Special cases with $\gamma_{max} = \gamma_{min}$for $U_{ge} \geq 0$

Special cases with $\gamma_{max} = \gamma_{min}$				**Special cases with $\gamma_{max} = \gamma_{min}$ and specific G-E interaction models**				**Special cases with $\gamma_{max} = \gamma_{min}$ and all risks all 0 or 1**			
Risk distribution		**Conditions**	**Population impact and Utility**	**Risk distribution**		**Conditions**	**Population impact and Utility**	**Risk distribution**		**Conditions**	**Population impact and Utility**
								1	1	$r_t = 1$ $PAF_e = 0$	Undefined ($PAF_{ge} = 0$)
				R_{00}	1	$\gamma_{minge} = \gamma_{maxge}$ ($R_{ge} = 1$ and $PAF_{ge} = PAF_e$) $f_{ge} = 1/PAF_e$	PI = 1 $U_{ge} = 1-\gamma$	1	1		
R_{g0}	1	$\gamma_{minge} = \gamma_{maxge}$ ($R_{ge} = 1$ and $PAF_{ge} = PAF_e$) $f_{ge} \geq 1$	PI = 1 $U_{ge} = 1-\gamma$	R_{00}	R_{00}			0	1	$r_t = \gamma\varepsilon$ $PAF_e = 1$	PI = 1 $U_{ge} = 1-\gamma$
R_{00}	R_{00}			R_{g0}	1	$\gamma_{minge} = \gamma_0 = \gamma_{maxge}$ ($R_{ge} = 1$; $R_{00} = 0$; $PAF_{ge} = PAF_e$) $f_{ge} = 1$	PI = 1 $U_{ge} = 1-\gamma$	0	0		
R_{g0}	1	$\gamma_{minge} = \gamma_0$ ($R_{ge} = 1$; $R_{00} = 0$) $0 \leq f_{ge} \leq 1$	0 = PI = 1 $U_{ge} = PI-\gamma$	0	0			1	1	$r_t = \gamma$ $PAF_e = 0$	Undefined ($PAF_{ge} = 0$)
0	R_{0e}			$1-R_{0e}$	1	$\gamma_{minge} = \gamma_0$ ($R_{ge} = 1$; $R_{00} = 0$) $f_{ge} = 0$	PI = γ $U_{ge} = 0$	0	0		
				0	R_{0e}			0	1	$r_t = \varepsilon$ $PAF_e = 1$	PI = γ $U_{ge} = 0$
								0	1		

Table 6: Comparison with classical twin study

	Classical twin study	**Twins + siblings model**
Genetic model	Additive and dominance terms only: $V^{DZ}_g = 1/2V_A + 1/4V_D$	Variable: $V^{DZ}_g = p^{DZ}_g V_g$ with $0 <= p^{DZ}_g <= 1/2$
Shared twin environments	Equal environments assumption: $c_{MD} = 1$	Variable: $1 <= c_{MD} <= R_{MD}$ $c_{MD} = R_{MD}$ implies $V_g = 0$
Shared sibling environments	Siblings not included.	Variable: $0 <= c_{SD} <= R_{SD}$ Familial aggregation may be due to genes ($c_{SD} = 0$) or environment ($c_{SD} = R_{SD}$).
Gene-environment interactions	None	Variable: $V_{ge} = f^2_{ge} \cdot V_g \cdot V_e/r^2_t$ $-\varepsilon/(1-\varepsilon)PAF_e <= f_{ge} <= 1/PAF_e$
Gene-environment correlations	None	None
Method	Total phenoptypic variance given by: $V_P = V_g + V_e$ V_P is input and a single solution for V_e and V_g calculated. Heritabilities are given by: $H^2 = V_g/V_P$ $h^2 = V_A/V_P$	V_e and ε are input and V_g and V_{ge} calculated, for a chosen c_{MD} and all possible values of f_{ge} and p^{DZ}_g. Method is not valid if $R_{SD} = 1$.

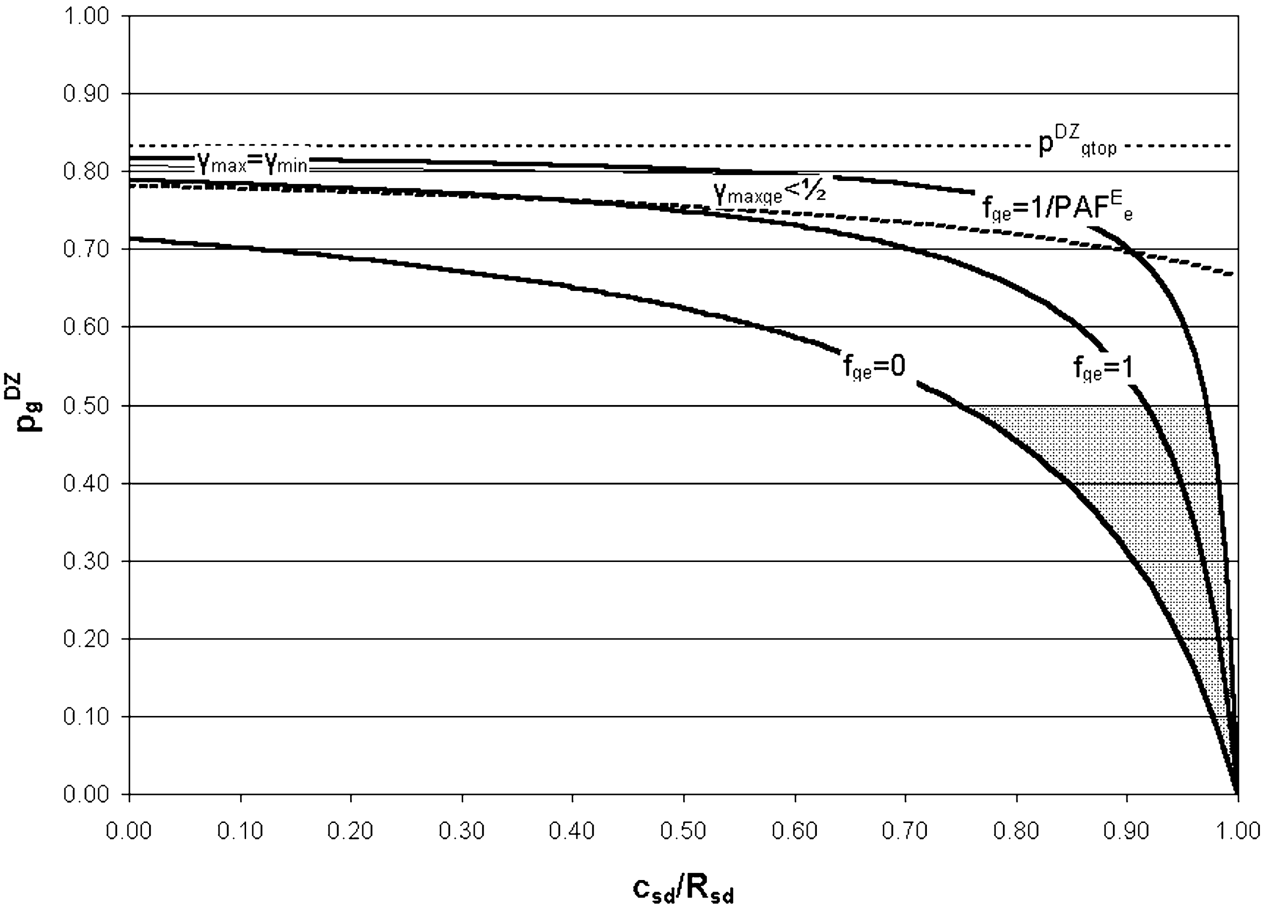

Figure 2

Example model solution space with $R_{MD} < 1+R_{SD}$ and $U_{ge} \geq 0$. Input parameters: $\lambda_{MZ} = 3.4$, $\lambda_{DZ} = 3$, $\lambda_{sib} = 2$, $\varepsilon = 0.2$, $PAF^E_e = 0.5$, $c_{MD} = 1$, $r_t = 0.1$. Hence $R_{MD} = 1.2$, $R_{SD} = 0.5$.

half the relevant variants by descent). Although $p^{DZ}_g = 0$ may not be realistic, low values of p^{DZ}_g are plausible, and may even be typical of complex diseases.

The classical twin study assumptions (see above) allow a single solution to be calculated from the under-determined system of simultaneous Equations (29), (30) and (31). However, in the absence of prior knowledge about the form of the genetic model, the presence or absence of gene-environment interactions, and the validity of the 'equal environments' assumption, the approach adopted here is more rigorous.

Results

General model solutions

First consider the behaviour of the model when the 'equal environments' assumption holds and hence $c_{MD} = 1$ (as described above).

Figures 2, 3 and 4 show the possible solution spaces for an arbitrary set of plausible input parameters satisfying the requirement $R_{MD} > 1+R_{SD}$ necessary for the classical twin study solution to exist. In Figure 2 the gene-environment interaction factor f_{ge} and hence utility, U_{ge}, are both positive and in Figure 3 they are negative. The horizontal axis shows c_{SD}/R_{SD}, which is zero if shared sibling risk is due to shared genetic factors only and 1 if shared sibling risk is due to shared environmental factors only. The vertical axis shows p^{DZ}_g, which is 1/2 if the additive genetic model holds, but may reduce to zero if epistasis dominates and the phenotype is robust to genetic variation. The three curved solid lines represent three models of gene-environment (G-E) interaction: an additive G-E model (i.e. no gene-environment interaction, $f_{ge} = 0$); a multiplicative G-E model ($f_{ge} = 1$); and maximum G-E interaction ($f_{ge} = 1/PAF^E_e$). The possible solution spaces are shaded grey. Each point in each shaded solution space corresponds to a

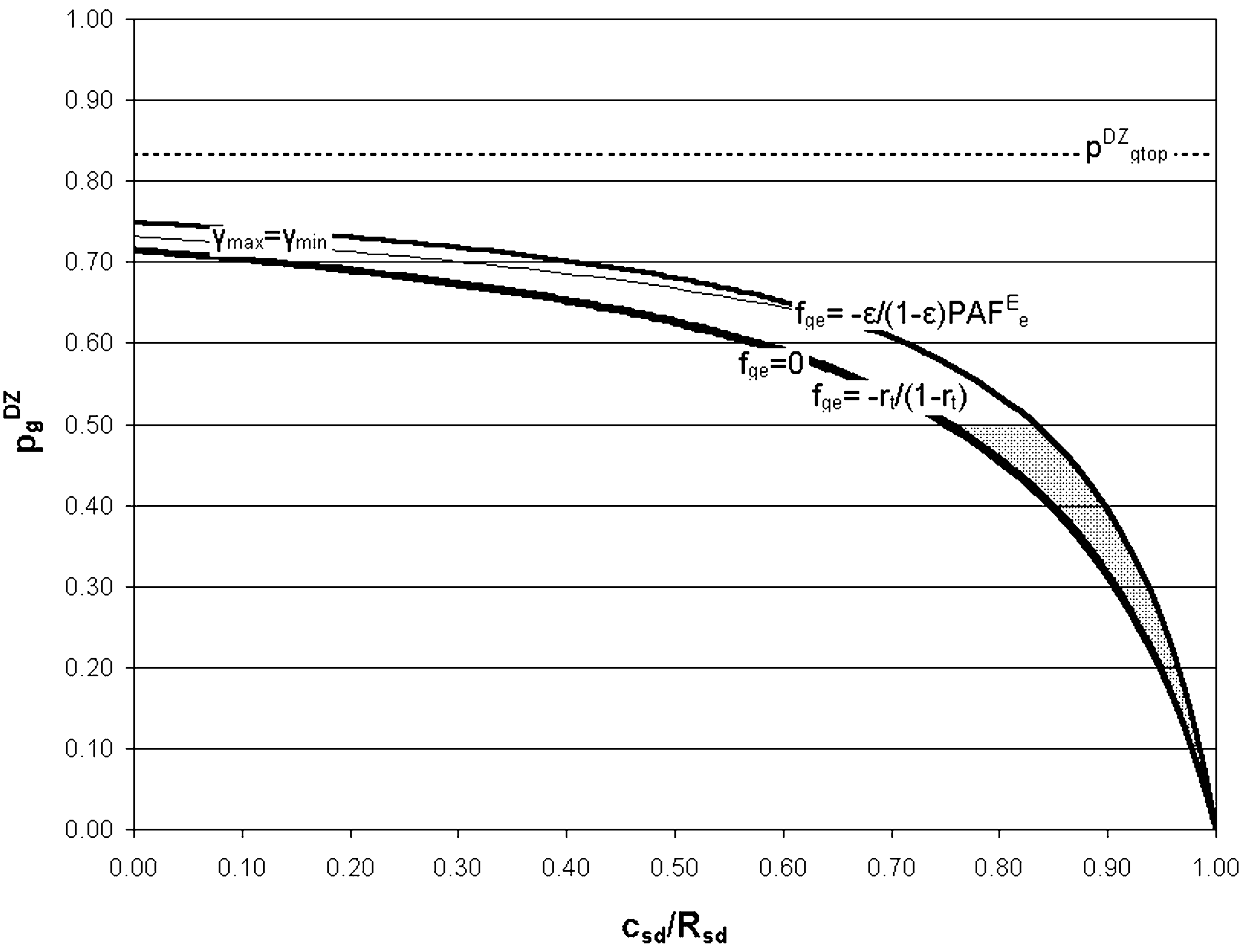

Figure 3
Example model solution space with $R_{MD} < 1+R_{SD}$ and $U_{ge} \leq 0$. Input parameters as for Figure 2.

given genetic model (defined by p^{DZ}_g) and a given G-E interaction model (defined by f_{ge}). Figure 4 plots the entire solution space (including both negative and positive utility) by transforming the horizontal axis to represent the G-E interaction parameter, f_{ge}. Although the classical twin model can fit the data, an infinite number of other solutions corresponding to different genetic and gene-environment interaction models also exist. In this example, the line $\gamma_{max} = \gamma_{min}$ lies outside the solution space and no solutions exist with $\gamma_{maxge} < 1/2$.

For lower values of RMD, the curves defining the solution space are shifted downwards [see Additional files 2 to 9], so that the line fge = 0 (corresponding to no gene-environment interaction) lies entirely below the line pDZg = 1/2 (corresponding to an additive genetic model). The classical twin study solution does not exist, but many other combinations of genetic and gene-environment interaction models may fit the data.

When cMD > 1, lines of constant fge no longer decrease monotonically to zero, and are also shifted upwards, so that solutions with strong G-E interactions are no longer possible [see Additional files 10 to 12].

Example applications using twin, sibling and environmental data

Input values

Consider example applications of the model for male lung cancer, female breast cancer and schizophrenia. The model input variables used are shown in Table 7.

The recurrence risks, λ, and total risks, r_t, for breast and lung cancer are those calculated by Risch [30], based on

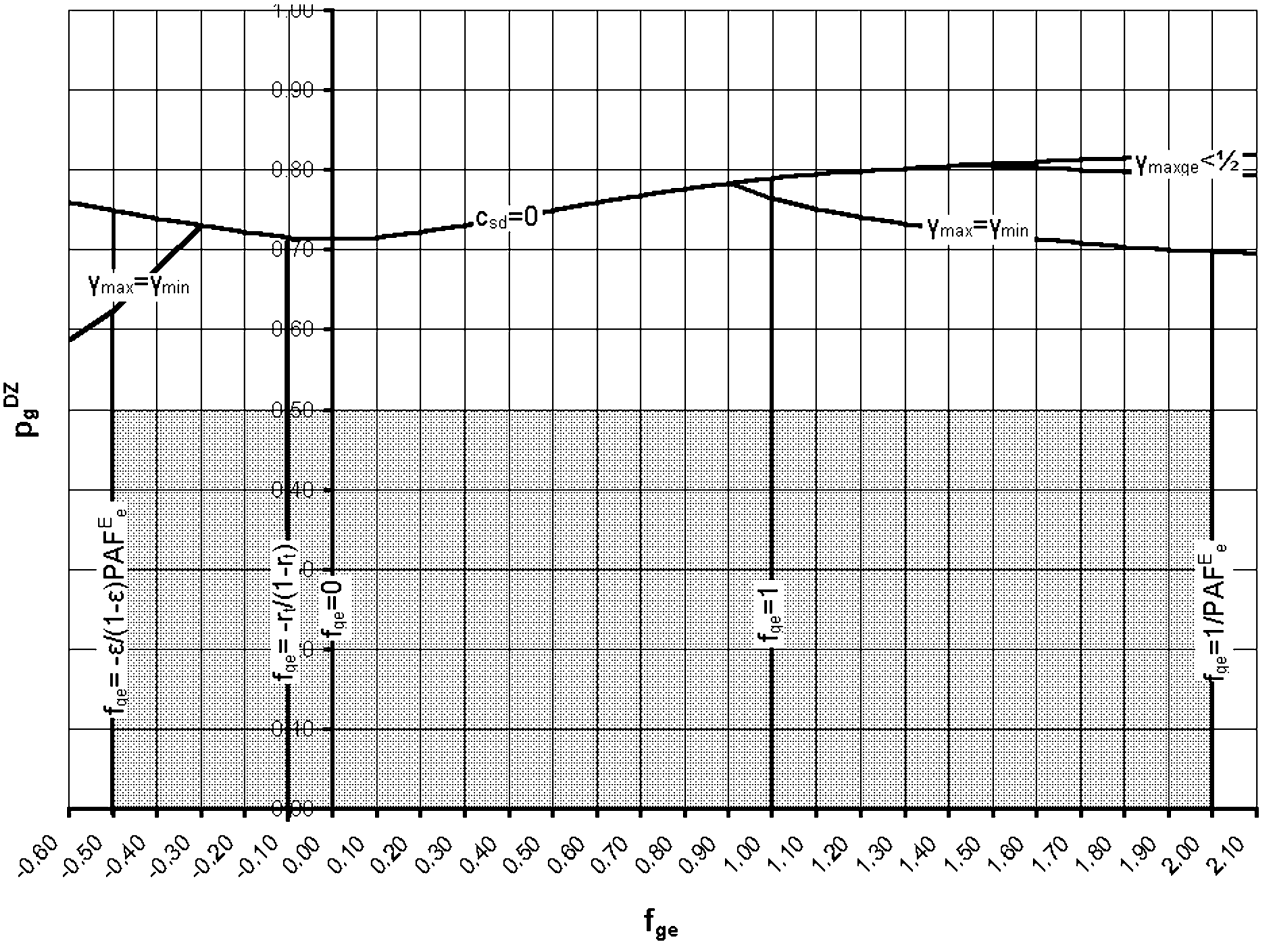

Figure 4
Example full model solution space with $R_{MD} < 1+R_{SD}$. Input parameters as for Figure 2, with the solution space transformed so that f_{ge} is on the horizontal axis.

Scandinavian twin data reported by Lichtenstein et al. [31] (involving more than 44,000 twin pairs) and Swedish familial data reported by Doug and Hemminki [32] (involving more than 2 million families). The proportion of the population exposed, ε, and population attributable fraction, PAF^E_e, for breast cancer are taken from those reported by Rockhill et al. [33] for a US population. Although strictly speaking these values may not be appropriate for a Scandinavian population, and include a component due to family history that may be (at least partly) genetic, they give a low V_e, consistent with the known environmental risk factors for breast cancer, and results are not sensitive to these input values (because V_e is so small). For lung cancer, it is assumed that 15% of the Scandinavian population smokes and that 86% of lung cancer cases could be avoided if they did not (giving a risk of lung cancer in smokers of 10%).

The recurrence risks λ, and total risk, r_t, for schizophrenia are those used by Risch [16], based on European data summarized by McGue et al. [34]. More recent twin studies for schizophrenia have given variable results and this example should be treated as illustrative only. Further, environmental exposures and population attributable fractions are unknown for schizophrenia. Two exploratory sets of results are therefore reported, using data consistent with a low environmental variance (based on the values used for breast cancer), and high environmental variance (based on the values used for smoking and lung cancer).

Detailed results for the three diseases are shown in [Additional file 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. The key findings are outlined below.

Table 7: Input variables

Condition	λ_{MZ}	λ_{DZ}	λ_{sib}	ε	PAF^{E}_{e}	r_t
Breast cancer	4.09	2.51	2.01	0.62	0.15	0.036
Lung cancer	6.27	6.14	3.16	0.15	0.86	0.017
Schizophrenia	52.1	14.2	8.6	0.62	0.15	0.01
				0.15	0.86	

Breast cancer results

For breast cancer, the PAF^{E}_{e} associated with known environmental factors is low. The value of the model is therefore less in calculating the utility of targeted environmental interventions than in exploring the solution space for a complex disease with R_{MD} close to 2.

Although strictly speaking the classical twin study solution (with an additive genetic model, $p^{DZ}_{g} = 1/2$, and an additive G-E model, $f_{ge} = 0$) does not exist as a solution, it might lie within the margin of error of the data. However, an infinite number of other models also could also fit the data. The classical twin model result always overestimates the genetic component of the variance, which reduces as the gene-environment interaction factor f_{ge} increases, and also as p^{DZ}_{g} decreases (i.e. as epistatic terms begin to dominate the genetic model). These alternative models imply that shared environmental factors may partially explain familial aggregation of breast cancer. This contrasts with the classical twin method result (see earlier), which for $R_{MD} = 2$ leads to the inevitable conclusion that shared sibling risk must be due solely to shared genes [35].

In theory, a model with $p^{DZ}_{g} = 0$ (where shared sibling risk is due entirely to shared environmental factors) could fit the data. However, for breast cancer the existence of known mutations that significantly increase risk (particularly mutations in the BRCA1 and BRCA2 genes, which are relatively common) rules out this solution. Although it is not possible to subtract out the effect of these mutations from the model, it is possible to show that they could be sufficient to explain the twin data if a G-E interaction also exists. For example, one possible solution consistent with the data could involve one or more dominant genes ($p^{DZ}_{g} = 1/2$), a strong G-E interaction ($f_{ge} = 1/PAF^{E}_{e}$), but a largely environmental explanation for shared sibling risk (say $c_{SD}/R_{SD} = 0.9$). This solution implies that the genetic component of the variance is less than a fifth of the classical twin study result, which could be low enough to be explained by mutations in the BRCA1 and BRCA2 genes alone [35]. If this model were correct it would have important implications for women with such mutations, but would not contribute significantly to reducing the incidence of breast cancer in the population as a whole, because the affected proportion of the population γ would be rather small. Other solutions, involving different genetic models with lower p^{DZ}_{g}, and/or less gene-environment interaction, are also possible.

The line $\gamma_{max} = \gamma_{min}$ does not occur within the solution space for breast cancer; however, in some circumstances the lines γ_{max} and γ_{min} may be rather close together. This suggests that, although as expected there is always a trade-off between selecting a small proportion (γ_{min}) of the population with a high Positive Predictive Value (PPV), or a larger proportion of the population (γ_{max}) with a higher Population Impact (PI) [19], some possible solutions could exist for breast cancer where the PPV and PI are both relatively high. Further, γ_{max} is often less than 1/2, so that, in these regions of the possible solution space, maximum utility might be obtained by targeting less than 50% of the population. However, known environmental factors for breast cancer are often not amenable to intervention and other possible solutions, with low, zero or negative utility, also exist.

Lung cancer results

For lung cancer, all the possible solutions imply that shared sibling risk is largely due to shared environmental factors (smoking) because solutions occur only when c_{SD}/R_{SD} is close to 1. Unlike for breast cancer, the line $\gamma_{max} = \gamma_{min}$ lies outside the solution space, even for negative f_{ge}, as does the area of solutions with $\gamma_{maxge} < 1/2$. However, the classical twin study solution, with $f_{ge} = 0$ and $p^{DZ}_{g} = 1/2$, clearly lies within the solution space.

Although the classical twin model again provides an upper limit to the genetic component of the variance, even the classical result indicates that the risk of lung cancer is dominated by smoking in this population and the variance has at most a small genetic component.

Unlike the breast cancer example, γ_{max} and γ_{min} are always far apart, suggesting a strong trade off between high Postive Predictive Value (R_{ge}) for a genotypic test and a high Population Impact (PI) for a targeted intervention. This means that a genotypic test that predicts which smokers will get lung cancer cannot exist. To predict all cases of lung cancer in smokers (i.e. to obtain PI = 1), 95% or more of the population would have to be in the high genotypic risk group, and the predictive value of such a test would be very low.

Because the genetic component of the variance is so small, it follows that the utility of genetic 'prediction and prevention' (measured by U_{ge}) is also small (from Equation (28)). Utility is maximum when $\gamma = 1/2$, but even then values are low. The maximum utility of genotyping occurs when about 60% of cases could be prevented by targeting the 50% of smokers at high genotypic risk. However, other possible solutions have zero or negative utility.

Schizophrenia results

For schizophrenia, the classical twin study solution (with $f_{ge} = 0$ and $p^{DZ}{}_{g} = 1/2$ and $c_{MD} = 1$) cannot not fit the data. If the 'equal environments' assumption holds, neither a single dominant gene ($p^{DZ}{}_{g} = 1/2$), nor additive polygenic model (also with $p^{DZ}{}_{g} = 1/2$), nor single recessive gene ($p^{DZ}{}_{g} = 1/4$) can explain the twin and family data, consistent with Risch's 1990 findings [16]. This may suggest that the genetic model for schizophrenia is likely to be dominated by epistatic terms. However, if gene-environment interactions are important, it is also possible that a recessive gene, combined with at least multiplicative G-E interaction ($p^{DZ}{}_{g} = 1/4$ and $f_{ge} = 1$ or higher), could explain the data.

The possible solution spaces include purely genetic explanations for shared sibling risk (at $c_{SD}/R_{SD} = 0$), or purely environmental ones (at $c_{SD}/R_{SD} = 1$, applicable if $p^{DZ}{}_{g} = 0$).

Assuming a small environmental component to the variance, there is no region of the solution space for which $\gamma_{maxge} < 1/2$, suggesting that the utility of targeted environmental interventions under these assumptions is likely to be low. However, if the environmental component of the variance is assumed to be much larger, the available solution space changes dramatically, because the line $\gamma_{max} = \gamma_{min}$ now constrains the solution space to a much smaller area, which excludes solutions with no G-E interaction ($f_{ge} = 0$). Special solutions may exist along the line $\gamma_{max} = \gamma_{min}$, as shown in Table 5. Because the environmental factors contributing to schizophrenia are unknown, it is impossible to draw any conclusions about the potential benefits of targeting environmental interventions at those at high genotypic risk.

Because prenatal development is thought to be important in schizophrenia, it is plausible that monozygotic twins are more likely to share environmental risk factors than dizygotic twins are. Breaking the 'equal environments' assumption changes the shape of the solution space significantly, and, assuming a small environmental component to the variance, only limited G-E interactions are now possible (the multiplicative G-E model, $f_{ge} = 1$, lies largely outside the solution space). The utility of targeting environmental interventions by genotype is then likely to be low. However, in these circumstances it is possible that an additive genetic model ($p^{DZ}{}_{g} = 1/2$) with some G-E interaction, or a recessive gene ($p^{DZ}{}_{g} = 1/4$) with no G-E interaction, could explain the data.

Discussion

If Fisher's polygenic model [26] is abandoned, along with the usual twin study assumption that there are no gene-environment interactions, the four-category model developed by Khoury and others can be combined with twin, family and environmental data to implement a 'top down' approach to assessing the utility of targeting environmental/lifestyle interventions by genotype. Scoping studies, valid when $R_{SD} \neq 1$, provide a first step to modelling the health of populations [23].

Abandoning Fisher's assumption that the polygenic model is necessarily dominated by an additive term can be justified by the growing evidence that phenotypic effects can result from the synergistic action of alleles in many genes [36]. For example, Bardet-Biedl Syndrome, historically assumed to be a recessive trait, has been shown to involve three interacting mutations at two loci in some patients (implying that $p^{DZ}{}_{g} = 1/8$) and, more recently, an additional locus has been identified that can also interact to change disease severity and symptoms [37]. Both positive and negative gene-environment interactions have also been observed in human diseases, although there are difficulties in confirming their statistical validity [38,39].

The model also allows the impact of the much criticised 'equal environments' assumption to be explored.

A number of conclusions can be drawn about the merits of the classical twin study and the utility of genetic 'prediction and prevention'.

Firstly, the model confirms that the classical twin study solution is not always valid and gives at best an upper limit to the genetic component of the variance of a trait. The importance of the 'equal environments' assumption and of gene-environment interactions have previously been recognised [17,18]; however, less attention has been paid to the potential role of gene-gene interactions (epistasis). For larger values of R_{MD} (greater than $1+R_{SD}$), observed for conditions such as schizophrenia, the model generalizes Risch's findings [16] to show that the three assumptions of the classical twin model cannot all be satisfied simultaneously. For intermediate R_{MD} values, observed for conditions such as breast cancer (for which R_{MD} is approximately 2), the model illustrates that the conclusion drawn from classical twin studies, that familial aggregation is due entirely to shared genetic factors, may be erroneous. This raises the possibility – previously rejected on the basis of twin study results [35] – that genetic variants are important in determining risk only for the relatively rare familial forms of cancer. If so, genetic models of familial aggregation (for example [40]) may be incorrect and the hunt for additional susceptibility genes could be largely fruitless. Existing published findings might then reflect prevailing bias, rather than true associations [14].

Secondly, the model confirms that the potential for reducing the incidence of common diseases using environmental/lifestyle interventions targeted by genotype may be limited [7] by:

(i) the low importance of genetic differences in determining the risk of some conditions (for example, lung cancer);

(ii) the complexity of gene-gene and gene-environment interactions and/or lack of knowledge of environmental factors (for example, schizophrenia).

Targeting environmental/lifestyle interventions at those at 'high genotypic risk' can be of high utility only in specific circumstances. The utility of targeting environmental interactions by genotype (compared to randomly selecting the same number of people from the population) is zero if there is no gene-environment interaction. Utility can also be negative in the presence of a negative interaction (i.e. if the people at high genotypic risk have *less to gain* by the intervention than people at low genotypic risk). The finding that utility increases with gene-environment interaction is consistent with Khoury and Wagener [19] but the relationship is considerably clarified by the adoption here of different measures of the population attributable fraction associated with a targeted intervention (PAF^{E}_{ge}) and of utility (U_{ge}). Further, by formally introducing constraints on the model (for example, that risks are positive and do not exceed 100%), it is possible to demonstrate that both the gene-environment interaction factor and utility have maximum values, which cannot be exceeded for a given data set.

The lung cancer example is apparently trivial but also of critical importance. The R_{MD} value for lung cancer is close to 1, and neither the Scandinavian data used here [31], nor earlier US studies [41], have identified a significant heritable component. It follows from Equation (27) that if the genetic component of the variance, V_g, is zero, V_{ge} (the G-E component of the variance) is also zero and using genotyping to target an intervention such as smoking cessation is therefore of zero utility (no better than randomly selecting the same number of individuals). This approximate conclusion is confirmed by the results presented for lung cancer, which show extremely low utility. The detailed calculations may at first sight seem unnecessary, particularly because smoking causes multiple diseases and targeting smoking cessation on the basis of lung cancer risk alone is therefore ill-advised. However, the idea that a genetic test will one day predict which smokers get lung cancer has been widely promoted in the literature and has driven much research aimed at identifying the supposed 'genes for lung cancer' [42]. The results presented here strongly suggest that there will never be a genetic or genotypic test that predicts which smokers will get lung cancer, because the genetic component of the variance is not high enough.

Finally, the model illustrates the argument of Terwilliger and Weiss [11] that the potential for population biobanks to quantify risks for complex disease is limited by a 'multiple testing' problem caused by the large number of genetic and gene-environment interaction models that could fit existing data. Each point in each solution space described above represents a different combination of a genetic risk model (defined by p^{DZ}_{g}) and a G-E interaction model (defined by f_{ge}). Further, any given value of p^{DZ}_{g} may be obtained by an infinite number of different combinations of different alleles acting through multiple biological pathways. Because the number of hypotheses that could be tested is essentially infinite, sample sizes necessary to quantify the risks (R_{oo}, R_{go}, R_{oe} and R_{ge}) could "*plausibly be larger than the number of people that have ever lived*" [11].

The model has several limitations. Measurements of shared sibling risk (λ_{sib}) are needed from the same population as twin data, and the scoping studies are only valid for $\lambda_{DZ} > \lambda_{sib}$, implying that environmental risks are more strongly correlated in dizygotic twins than other siblings. Some data exist to support this assumption for smoking [43] but for other exposures its validity is usually unknown. However, the model does not reduce to the classical twin study solution if this condition is not met: instead, data from more relatives are needed. In principle the model could, and should, be expanded to include data from more relatives, other data (such as migration study data), more risk categories and error terms. However, the number of unknown parameters will then increase, unless more data are available to quantify exposures (which change from generation to generation) and to estimate the extent to which environments are correlated between different types of relative.

Treating exposure and environmental variance (or population attributable fraction) as input data is also problematic when the effects of environmental factors on risk are often unknown. Further, the simple nature of the model (with one environmental axis) cannot adequately represent the complexity of environmental (including socio-economic) causes of disease. However, if targeting environmental interventions by genotype is to be considered, this implies that at least something is known (or expected to be learned) about environmental factors, such as particular exposures, that are amenable to intervention.

The assumption of no gene-environment correlation will often hold (for example it is rather implausible that the same genes strongly influence both lung cancer risk and nicotine addiction), but is not necessarily always true. Adult lactose intolerance is an example of a condition with a strong gene-environment interaction where targeted intervention to avoid drinking milk may be of high utility. However, the model is invalid for lactose intolerance unless exposures are applied equally to the population studied because, in general, people who are lactose intolerant may be less likely to drink milk (a gene-environment correlation) owing to the unpleasant symptoms.

A more fundamental problem is caused by the assumptions that: (i) the risks R_{oo}, R_{go}, R_{oe} and R_{ge} are inherent properties of a given trait within a given population (with a given γ and ε) and that there are therefore no confounders; and (ii) risks are randomly distributed within these categories.

These assumptions, although often made, are implausible in many situations. The assumption of no confounders means that the model can only represent a subset of the potential models of gene-gene and gene-environment interaction described by more complex models (for example [17]). It is unlikely to be met if multiple genetic factors interact with multiple environmental ones [44]. Although this may well render the results presented here invalid, such complexity is likely to reduce the utility of targeting by genotype, rather than enhance it. Hence, situations where the 'no confounders' assumption at least approximately holds are those most likely to be of relevance to public health.

The second assumption neglects the fact that for most exposures there is a gradient in risk, with higher exposure meaning higher risk, and that the same may also be true of genetic factors. This means that increasing the number of categories in the model will increase V_e (see [45]) and perhaps V_g. Further, these subcategories may be differently correlated between relatives (for example, the twin of a heavy smoker may be more likely to be a heavy smoker than a light one). If so, a relative of a proband may not be representative of their allocated risk category in the four-category model and Equation (22) then becomes invalid.

More broadly, these assumptions make the model, like the classical twin model, essentially deterministic: it assumes that all the factors contributing to correlations in risk between relatives are perfectly known and are either environmental or genetic. Retention of these assumptions here may be problematic and could limit the applicability of the results. Nevertheless, all the other questionable assumptions of the classical twin model have been simultaneously removed.

Conclusion

The model shows that the potential for reducing the incidence of common diseases using environmental interventions targeted by genotype may be limited, except in special cases. The model also confirms that the importance of an individual's genotype in determining their risk of complex diseases tends to be exaggerated by the classical twin studies method, owing to the 'equal environments' assumption and the assumption of no gene-environment interaction. In addition, if phenotypes are genetically robust, because of epistasis, a largely environmental explanation for shared sibling risk is plausible, even if the classical heritability is high. The model therefore highlights the possibility – previously rejected on the basis of twin study results – that inherited genetic variants are important in determining risk only for the relatively rare familial forms of diseases such as breast cancer. If so, genetic models of familial aggregation may be incorrect and the hunt for additional susceptibility genes could be largely fruitless.

Competing interests

The author(s) declare that they have no competing interests.

Appendix A: formal derivation of equation (31)

Equation (23) may be derived more formally by extending the matrix method of Li and Sacks [46].

Define the probability that an affected proband is in genotypic risk category z and environmental risk category w as P_{zw} and assume that risks are randomly distributed within categories. Using the definitions of the four category model given in Table 1, a vector **P** may be defined:

$$\mathbf{P} = \begin{pmatrix} P_{oo} \\ P_{oe} \\ P_{go} \\ P_{ge} \end{pmatrix} = \begin{pmatrix} (1-\varepsilon)(1-\gamma)R_{oo}/r_t \\ \varepsilon(1-\gamma)R_{oe}/r_t \\ \gamma(1-\varepsilon)R_{go}/r_t \\ \gamma\varepsilon R_{ge}/r_t \end{pmatrix} \qquad (\text{A1})$$

A risk vector **R** may also be defined:

$$\mathbf{R} = \begin{pmatrix} R_{oo} \\ R_{oe} \\ R_{go} \\ R_{ge} \end{pmatrix} \qquad (\text{A2})$$

Now define G_{xy} as the conditional probability P(relative is in genotypic risk category y|proband is in genotypic risk category x). Similarly, define E_{xy} as the conditional probability P(relative is in environmental risk category y|proband is in environmental risk category x). Using the definitions of p^{rel}_g and p^{rel}_e given in Section 2.5, matrices **G** and **E** may be written such that:

$$\mathbf{G}^{rel} = \begin{pmatrix} G_{oo} & G_{og} \\ G_{go} & G_{gg} \end{pmatrix} = \begin{pmatrix} p_g^{rel} + (1-\gamma)(1-p_g^{rel}) & \gamma(1-p_g^{rel}) \\ (1-\gamma)(1-p_g^{rel}) & p_g^{rel} + \gamma(1-\gamma)p_g^{rel} \end{pmatrix} \qquad (\text{A3})$$

$$\mathbf{E}^{rel} = \begin{pmatrix} E_{oo} & E_{oe} \\ E_{eo} & E_{ee} \end{pmatrix} = \begin{pmatrix} p_e^{rel} + (1-\varepsilon)(1-p_e^{rel}) & \varepsilon(1-p_e^{rel}) \\ (1-\varepsilon)(1-p_e^{rel}) & p_e^{rel} + \varepsilon(1-\varepsilon)p_e^{rel} \end{pmatrix} \qquad (\text{A4})$$

Finally, define $X_{ab\text{-}cd}$ as the conditional probability P(relative is in risk category cd|proband is in risk category ab), where the risk categories are as defined in Table 1 (for example risk categorgy 'ge' implies high-genotypic and high-environmental risk). Provided p^{rel}_g and p^{rel}_e are independent (there are no gene-environment correlations), the gene-environment interaction matrix $\mathbf{M}^{rel}_{ge}$ may be written as:

$$\mathbf{M}_{ge}^{rel} = \begin{pmatrix} X_{oo-oo} & X_{oo-oe} & X_{oo-go} & X_{oo-ge} \\ X_{oe-oo} & X_{oe-oe} & X_{oe-go} & X_{oe-ge} \\ X_{go-oo} & X_{go-oe} & X_{go-go} & X_{go-ge} \\ X_{ge-oo} & X_{ge-oe} & X_{ge-go} & X_{ge-ge} \end{pmatrix} = \begin{pmatrix} G_{oo}E_{oo} & G_{oo}E_{oe} & G_{og}E_{oo} & G_{og}E_{oe} \\ G_{oo}E_{eo} & G_{oo}E_{ee} & G_{og}E_{eo} & G_{og}E_{ee} \\ G_{go}E_{oo} & G_{go}E_{oe} & G_{gg}E_{oo} & G_{gg}E_{oe} \\ G_{go}E_{eo} & G_{go}E_{ee} & G_{gg}E_{eo} & G_{gg}E_{ee} \end{pmatrix} \qquad (\text{A5})$$

Then the risk in a relative of the proband is given by:

$$\lambda_{rel} r_t = \mathbf{P}.\left(\mathbf{M}_{ge}^{rel}\mathbf{R}\right) \qquad (\text{A6})$$

After some algebra, this yields equation (23).

Appendix B: calculating recurrence risks for twins

The sibling recurrence risk λ_{sib} is often available directly from familial studies. For twins the recurrence risks, if not reported, may be calculated from the case-wise concordance (Cc):

$$\lambda_{MZ} = Cc_{MZ}/r_t \qquad (\text{B1})$$

$$\lambda_{DZ} = Cc_{DZ}/r_t \qquad (\text{B2})$$

where, if there is complete ascertainment of all affected twins in a population,

$$Cc = 2C/(2C + D) \qquad (\text{B3})$$

and C is the number of concordant and D the number of discordant pairs [25].

Additional material

Additional File 1

Gene-gene and gene-environment interaction model. Contains the Visual Basic macro (Twincal), input and output datasheets and charts used to calculate the solutions described in the text. The program is run by entering parameters in the 'Inputs' sheet and clicking on the 'Run' button. Note that for the final chart ('fe') the number of categories on the horizontal axis changes depending on the environmental input parameters ε and PAF^E_e. If these parameters are changed it is therefore necessary to delete the lower part of the output sheet prior to running the model and, after the run, to redraw the chart using the source data option from the chart. All other charts are drawn automatically. The line $\gamma_{max} = \gamma_{min}$ is calculated exactly for the chart 'fe' but is approximated in the charts 'pgdz' and 'pgdzneg' using Newton's method and an initial guess for f_{ge} (f0) and step (fet). For some input parameters it may be necessary to change these values by editing the Visual Basic code (Twincal) to obtain a valid solution.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S1.xls]

Additional File 2

Supplementary Figure 1: Example model solution space with $R_{MD} = 1.7$ and $U_{ge} \geq 0$. Model solution space with $U_{ge} \geq 0$ for the same input parameters as Figure 2, apart from $\lambda_{MZ} = 4.4$.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S2.bmp]

Additional File 3

Supplementary Figure 2: Example model solution space with $R_{MD} = 1.8$ and $U_{ge} \geq 0$. Model solution space with $U_{ge} \geq 0$ for the same input parameters as Figure 2, apart from $\lambda_{MZ} = 4.6$.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S3.bmp]

Additional File 4

Supplementary Figure 3: Example model solution space with $R_{MD} = 1.95$ and $U_{ge} \geq 0$. Model solution space with $U_{ge} \geq 0$ for the same input parameters as Figure 2, apart from $\lambda_{MZ} = 4.9$.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S4.bmp]

Additional File 5

Supplementary Figure 4: Example model solution space with $R_{MD} = 2.1$ and $U_{ge} \geq 0$. Model solution space with $U_{ge} \geq 0$ for the same input parameters as Figure 2, apart from $\lambda_{MZ} = 5.2$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S5.bmp]

Additional File 6

Supplementary Figure 5: Example full solution space with $R_{MD} = 1.7$. Full model solution space for the same input parameters as Figure 5, transformed so that f_{ge} is on the horizontal axis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S6.bmp]

Additional File 7

Supplementary Figure 6: Example full solution space with $R_{MD} = 1.8$. Full model solution space for the same input parameters as Figure 6, transformed so that f_{ge} is on the horizontal axis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S7.bmp]

Additional File 8

Supplementary Figure 7: Example full solution space with $R_{MD} = 1.95$. Full model solution space for the same input parameters as Figure 7, transformed so that f_{ge} is on the horizontal axis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S8.bmp]

Additional File 9

Supplementary Figure 8: Example full solution space with $R_{MD} = 2.1$. Full model solution space for the same input parameters as Figure 8, transformed so that f_{ge} is on the horizontal axis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S9.bmp]

Additional File 10

Supplementary Figure 9: Example model solution with $c_{MD} > 1$ and $U_{ge} \geq 0$. Input parameters: $\lambda_{MZ} = 5.2$, $\lambda_{DZ} = 3$, $\lambda_{sib} = 2$, $\varepsilon = 0.2$, $PAF^E_e = 0.5$, $c_{MD} = 2$, $r_t = 0.1$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S10.bmp]

Additional File 11

Supplementary Figure 10: Example model solution with $c_{MD} > 1$ and $U_{ge} \geq 0$. Input parameters as for Figure 13.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S11.bmp]

Additional File 12

Supplementary Figure 11: Example full solution space with $c_{MD} > 1$. Full model solution space for the same parameters as Figure 13, transformed so that f_{ge} is on the horizontal axis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S12.bmp]

Additional File 13

Supplementary Figure 12: Breast cancer solution space with $U_{ge} \geq 0$. Input parameters are as shown in Table 5, with $c_{MD} = 1$. The solution space is shown (shaded) for positive f_{ge}, assuming the 'equal environments' assumption holds ($c_{MD} = 1$). The darker shaded area shows the part of the solution space for which $\gamma_{maxge} < 1/2$. Utility U_{ge} is at its maximum when $\gamma = 1/2$ except within this darker shaded area.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S13.bmp]

Additional File 14

Supplementary Figure 13: Breast cancer variances with $f_{ge} = 0$. Input parameters as for Figure 16. Additive model of G-E interaction ($f_{ge} = 0$). Variance components are genetic (V_g) or environmental (V_e).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S14.bmp]

Additional File 15

Supplementary Figure 14: Breast cancer variances with $f_{ge} = 1$. Input parameters as for Figure 16. Multiplicative G-E interaction model ($f_{ge} = 1$). Variance components are genetic (V_g), environmental (V_e) or due to gene-environment interaction (V_{ge}).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S15.bmp]

Additional File 16

Supplementary Figure 15: Breast cancer variances with $f_{ge} = 1/PAF^E_e$. Input parameters as for Figure 16. Maximum G-E interaction model ($f_{ge} = 1/PAF^E_e$). Variance components are genetic (V_g), environmental (V_e) or due to gene-environment interaction (V_{ge}).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S16.bmp]

Additional File 17

Supplementary Figure 16: Breast cancer γ values with $f_{ge} = 0$. Input parameters as for Figure 16. The proportion of the population in the 'high genotypic risk' group, γ, may take any value in the shaded area. γ_{min} occurs when $R_{ge} = 1$, i.e. when the Positive Predictive Value (PPV) of being in the 'ge' subgroup is 100%. γ_{max} occurs when $R_{oo} = 1$ for an additive G-E model and solutions with a Population Impact of 100% (PI = 1) cannot exist.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S17.bmp]

Additional File 18

Supplementary Figure 17: Breast cancer γ values with $f_{ge} = 1$. Input parameters as for Figure 16. The proportion of the population in the 'high genotypic risk' group, γ, may take any value in the shaded area. A solution with a Population Impact of 100% (PI = 1) may exist if $\gamma = \gamma_{max}$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S18.bmp]

Additional File 19

Supplementary Figure 18: Breast cancer γ values with $f_{ge} = 1/PAF^{E}_{e}$. Input parameters as for Figure 16. The proportion of the population in the 'high genotypic risk' group, γ, may take any value in the shaded area. A solution with a Population Impact of 100% (PI = 1) may exist if $\gamma = \gamma_{max}$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S19.bmp]

Additional File 20

Supplementary Figure 19: Breast cancer solution space with $U_{ge} \leq 0$. Input parameters are as for Figure 16. The solution space is shown for negative f_{ge} (where the utility of targeting environmental interventions at the high genotypic risk group is negative, $U_{ge} \leq 0$). Solutions exist only in the shaded area where $\gamma_{max} \geq \gamma_{min}$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S20.bmp]

Additional File 21

Supplementary Figure 20: Breast cancer: full solution space. Input parameters are as for Figure 16. The same solution space as Figures 16 and 23 is shown (shaded), transformed so that the G-E interaction factor is plotted on the horizontal axis. Again, each point in the shaded solution space represents a genetic model defined by p^{DZ}_{g} and a G-E interaction model defined by f_{ge}. The area of solutions with $\gamma_{maxge} < 1/2$ is highlighted with darker shading. The classical twin study solution lies on the vertical axis ($f_{ge} = 0$) at the point $p^{DZ}_{g} = 1/2$, and is slightly outside the solution space.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S21.bmp]

Additional File 22

Supplementary Figure 21: Lung cancer solution space with $U_{ge} \geq 0$. Input parameters are as shown in Table 5, with $c_{MD} = 1$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S22.bmp]

Additional File 23

Supplementary Figure 22: Lung cancer variances with $f_{ge} = 0$. Input parameters as for Figure 25. Note that the horizontal axis has been expanded to show high values of c_{SD}/R_{SD} only.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S23.bmp]

Additional File 24

Supplementary Figure 23: Lung cancer variances with $f_{ge} = 1$. Input parameters as for Figure 25. Note that the horizontal axis has been expanded to show high values of c_{SD}/R_{SD} only.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S24.bmp]

Additional File 25

Supplementary Figure 24: Lung cancer variances with $f_{ge} = 1/PAF^{E}_{e}$. Input parameters as for Figure 25. Note that the horizontal axis has been expanded to show high values of c_{SD}/R_{SD} only.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S25.bmp]

Additional File 26

Supplementary Figure 25: Lung cancer γ values for $f_{ge} = 1$. Input parameters as for Figure 25. The proportion of the population in the 'high genotypic risk' group, γ, may take any value in the shaded area.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S26.bmp]

Additional File 27

Supplementary Figure 26: Lung cancer γ values for $f_{ge} = 1/PAF^{E}_{e}$. Input parameters as for Figure 25. The proportion of the population in the 'high genotypic risk' group, γ, may take any value in the shaded area.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S27.bmp]

Additional File 28

Supplementary Figure 27: Lung cancer U_{ge} values for $f_{ge} = 1$. Input parameters as for Figure 25. The utility parameter, U_{ge}, may take any value in the shaded area, but is maximum when $\gamma = 1/2$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S28.bmp]

Additional File 29

Supplementary Figure 28: Lung cancer U_{ge} values for $f_{ge} = 1/PAF^{E}_{e}$. Input parameters as for Figure 25. The utility parameter, U_{ge}, may take any value in the shaded area, but is maximum when $\gamma = 1/2$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S29.bmp]

Additional File 30

Supplementary Figure 29: Lung cancer: full solution space. Input parameters as for Figure 25.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S30.bmp]

Additional File 31

Supplementary Figure 30: Schizophrenia $U_{ge} \geq 0$, small environmental variance and $c_{MD} \geq 1$. Input parameters are as shown in Table 5, with $\varepsilon = 0.62$, $PAF^{E}_{e} = 0.15$ and $c_{MD} = 1$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S31.bmp]

Additional File 32

Supplementary Figure 31: Schizophrenia $U_{ge} \geq 0$, small environmental variance and $c_{MD} > 1$. Input parameters are as shown in Table 5, with $\varepsilon = 0.62$, $PAF^{E}_{e} = 0.15$ and $c_{MD} = 3.8$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S32.bmp]

Additional File 33

Supplementary Figure 32: Schizophrenia $U_{ge} \geq 0$, large environmental variance and $c_{MD} = 1$. Input parameters are as shown in Table 5, with $\varepsilon = 0.15$, $PAF^{E}_{e} = 0.86$ and $c_{MD} = 1$.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4682-3-35-S33.bmp]

Acknowledgements

The author is grateful to the Joseph Rowntree Charitable Trust for funding the completion of this work.

References

1. Collins FS: **Shattuck Lecture – medical and societal consequences of the Human Genome Project.** *New Engl J Med* 1999, **341:**28-37.
2. Bell J: **The new genetics in clinical practice.** *BMJ* 1998, **316(7131):**618-620.
3. Collins FS, McKusick VA: **Implications of the Human Genome Project for medical science.** *J Am Med Assoc* 2001, **285:**540-544.
4. Strohman RC: **The coming Kuhnian revolution in biology.** *Nat Biotechnol* 1997, **15:**194-200.
5. Holtzman NA, Marteau TM: **Will genetics revolutionize medicine?** *New Engl J Med* 2000, **343:**141-144.
6. Vineis P, Schulte P, McMichael AJ: **Misconceptions about the use of genetic tests in populations.** *Lancet* 2001, **357:**709-712.
7. Baird P: **The Human Genome Project, genetics and health.** *Community Genet* 2001, **4:**77-80.
8. Cooper RS, Psaty BM: **Genetics and medicine: distraction, incremental progress, or the dawn of a new age?** *Ann Intern Med* 2003, **138:**576-580.
9. Vineis P, Ahsan H, Parker M: **Genetic screening and occupational and environmental exposures.** *Occup Environ Med* 2004, **62:**657-662.
10. Khoury MJ, Yang Q, Gwinn M, Little J, Flanders WD: **An epidemiologic assessment of genetic profiling for measuring susceptibility to common diseases and targeting interventions.** *Genet Med* 2004, **6(1):**38-47.
11. Terwilliger JD, Weiss KM: **Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'.** *Ann Med* 2003, **35:**532-544.
12. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ionnidis DG: **Replication validity of genetic association studies.** *Nat Genet* 2001, **29:**306-309.
13. Cordell HJ, Clayton DG: **Genetic association studies.** *Lancet* 2005, **366:**1121-1131.
14. Ioannidis J: **Why most published research findings are false.** *PloS Med* 2005, **2(8):**e124. DOI: 10.137/journal.pmed.0020124
15. Layzer D: **Heritability analyses of IQ scores: science or numerology?** *Science* 1974, **183:**1259-1266.
16. Risch N: **Linkage strategies for genetically complex traits. I. Multilocus models.** *Am J Hum Genet* 1990, **46:**222-228.
17. Guo S-W: **Gene-environment interaction and the mapping of complex traits: some statistical models and their interpretation.** *Hum Hered* 2000, **50:**286-303.
18. Hopper JL: **Why 'common environmental effects' are so uncommon in the literature.** In *Advances in twin and sib-pair analysis* Edited by: Spector TD, Sneider H, MacGregor AJ. London: Greenwich Medical Media Ltd; 2000.
19. Khoury MJ, Wagener DK: **Epidemiological evaluation of the use of genetics to improve the predictive value of disease risk factors.** *Am J Hum Genet* 1995, **56:**835-844.
20. Lewis SJ, Brunner EJ: **Methodological problems in genetic association studies of longevity – the apolipoprotein E gene as an example.** *Int J Epidemiol* 2004, **33:**962-970.
21. Tryggvadottir L, Sigvaldason H, Olafsdottir GH, Jonasson JG, Jonsson T, Tulinius H, Eyfjord JE: **Population-based study of changing breast cancer risk in Icelandic BRCA2 mutation carriers, 1920–2000.** *J Natl Cancer Inst* 2006, **98(2):**116-122.
22. Humphries S, Ridker PM, Talmud PJ: **Genetic testing for cardiovascular disease susceptibility: a useful clinical management tool or possible misinformation?** *Arterioscler Thromb Vasc Biol* 2004, **24:**628-636.
23. Rose G: **Sick individuals and sick populations.** *Int J Epidemiol* 1985, **14:**32-38.
24. Khoury MJ, Jones K, Grosse SD: **Quantifying the health benefits of genetic tests: The importance of a population perspective.** *Genet Med* 2006, **8(3):**191-195.
25. MacGregor AJ: **Practical approaches to account for bias and confounding in twin data.** In *Advances in twin and sib-pair analysis* Edited by: Spector TD, Sneider H, MacGregor AJ. London: Greenwich Medical Media Ltd; 2000.
26. Fisher RA: **The correlation between relatives on the supposition of Mendelian inheritance.** *Trans R Soc Edinb* 1918, **52:**399-433.
27. Hopper JL: **Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health.** *Stat Methods Med Res* 1993, **2:**199-223.
28. Porteous JW: **A rational treatment of Mendelian genetics.** *Theor Biol Med Model* 2004, **1:**6. DOI: 10.1186/1742-4682-1-6
29. Azevedo RBR, Lohaus R, Srinivasan S, Dang KK, Burch CL: **Sexual reproduction selects for robustness and negative epistasis in artificial gene networks.** *Nature* 2006, **440:**87-90.
30. Risch N: **The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches.** *Cancer Epidemiol Biomarkers Prev* 2001, **10:**733-741.
31. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer.** *New Engl J Med* 2000, **343:**78-85.
32. Dong C, Hemminki K: **Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families.** *Int J Cancer* 2001, **92:**144-150.
33. Rockhill B, Weinberg CR, Newman B: **Population attributable fraction estimation for established breast cancer risk factors: considering the issues of high prevalence and unmodifiability.** *Am J Epidemiol* 1998, **147(9):**826-833.
34. McGue M, Gottesman II, Rao DC: **The transmission of schizophrenia under a multifactorial threshold model.** *Am J Hum Genet* 1983, **35:**1161-1178.
35. Easton DF: **How many more breast cancer predisposition genes are there?** *Breast Cancer Res* 1999, **1(1):**14-17.
36. Badano JL, Katsanis N: **Beyond Mendel: an evolving view of human genetic disease transmission.** *Nat Rev Genet* 2002, **3:**779-789.
37. Badano JL, Leitch CC, Ansley SJ, May-Simera H, Lawson S, Lewis RA, Beales PL, Dietz HC, Fisher S, Katsanis N: **Dissection of epistasis in oligogenic Bardet-Biedl syndrome.** *Nature* 2006, **439:**326-330.
38. Taioli E, Zocchetti C, Garte S: **Models of interaction between metabolic genes and environmental exposure in cancer susceptibility.** *Environ Health Perspect* 1998, **106(2):**67-70.
39. Hunter DJ: **Gene-environment interactions in human diseases.** *Nat Rev Genet* 2005, **6:**287-298.
40. Antoniou AC, Pharoah PDP, McMullan G, Day NE, Stratton MR, Peto J, Ponder BJ, Easton DF: **A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes.** *Br J Cancer* 2002, **86:**76-83.
41. Braun MM, Caporaso NE, Page WF, Hoover RN: **A cohort study of twins and cancer.** *Cancer Epidemiol Biomarkers Prev* 1995, **4(5):**469-473.
42. Hall W, Madden P, Lynskey M: **The genetics of tobacco use: methods, findings and policy implications.** *Tob Control* 2002, **11:**119-124.
43. Vink JM, Willemsen G, Boomsma DI: **The association of current smoking behavior with the smoking behavior of parents, siblings, friends and spouses.** *Addiction* 2003, **98:**923-931.
44. Taioli E, Garte S: **Covariates and confounding in epidemiologic studies using metabolic gene polymorphisms.** *Int J Cancer* 2002, **100:**97-100.
45. Guo S: **The behaviors of some heritability estimators in the complete absence of genetic factors.** *Hum Hered* 1999, **49(4):**215-228.

46. Li CC, Sacks L: **The derivation of joint distribution and correlation between relatives by the use of stochastic matrices.** *Biometrics* 1954, **10:**347-360.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."
Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral