A Better Coefficient of Determination for Genetic Profile Analysis

Sang Hong Lee,[1,2]* Michael E Goddard,[3,4] Naomi R Wray,[1,2] and Peter M Visscher[1,2,5]

[1]*Queensland Institute of Medical Research, Brisbane Queensland, Australia*
[2]*Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia*
[3]*Biosciences Research Division, Department of Primary Industries, Victoria, Australia*
[4]*Department of Agriculture and Food Systems, University of Melbourne, Melbourne, Australia*
[5]*University of Queensland Diamantina Institute University of Queensland, Princess Alexandra Hospital, Brisbane Queensland, Australia*

Genome-wide association studies have facilitated the construction of risk predictors for disease from multiple Single Nucleotide Polymorphism markers. The ability of such "genetic profiles" to predict outcome is usually quantified in an independent data set. Coefficients of determination (R^2) have been a useful measure to quantify the goodness-of-fit of the genetic profile. Various pseudo-R^2 measures for binary responses have been proposed. However, there is no standard or consensus measure because the concept of residual variance is not easily defined on the observed probability scale. Unlike other nongenetic predictors such as environmental exposure, there is prior information on genetic predictors because for most traits there are estimates of the proportion of variation in risk in the population due to all genetic factors, the heritability. It is this useful ability to benchmark that makes the choice of a measure of goodness-of-fit in genetic profiling different from that of nongenetic predictors. In this study, we use a liability threshold model to establish the relationship between the observed probability scale and underlying liability scale in measuring R^2 for binary responses. We show that currently used R^2 measures are difficult to interpret, biased by ascertainment, and not comparable to heritability. We suggest a novel and globally standard measure of R^2 that is interpretable on the liability scale. Furthermore, even when using ascertained case-control studies that are typical in human disease studies, we can obtain an R^2 measure on the liability scale that can be compared directly to heritability. *Genet. Epidemiol.* 36:214–224, 2012.

Key words: coefficient of determination; risk predictors; genetic profiles; goodness-of-fit; genome-wide association studies

Supporting Information is available in the online issue at wileyonlinelibrary.com.
*Correspondence to: Sang Hong Lee, Queensland Brain Institute, University of Queensland, Brisbane, QLD 4072, Australia. E-mail: Hong.Lee@uq.edu.au
Received 5 September 2011; Revised 6 December 2011; Accepted 9 December 2011
Published online 14 March 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21614

INTRODUCTION

The discovery of multiple genetic loci that are associated with disease and other complex traits has sparked an interest in making individual risk predictions from genetic data [Demirkan et al., 2011; Evans et al., 2009; Kraft et al., 2009; Lyssenko et al., 2008; Pharoah et al., 2008; Purcell et al., 2009; The International Multiple Sclerosis Genetics Consortium, 2010; Wray et al., 2007, 2010]. The genetic risk of healthy individuals can be predicted from their measured genotype at multiple loci, and, since the total (phenotypic) risk is correlated with genetic risk, a prediction can be made of total risk from genetic data. Typically, the effects of measured genotypes on disease are estimated in one or more discovery samples, and those estimated effects are then combined with the genotypes at one or more independent validation samples that contain affected and unaffected individuals. For each individual in the validation sample, a genetic profile is calculated and these profiles are correlated with outcome (affected/unaffected) to quantify the precision of the genetic risk predictor.

What is a good measure to quantify the goodness-of-fit of the genetic profile? Importantly, there is prior information on genetic predictors because for most traits there are estimates of the proportion of variation in risk in the population due to all genetic factors, the heritability. From the heritability, we can calculate the maximum precision of a genetic profile, i.e. the precision if all causal variants were known and their effect sizes known without error [Wray et al., 2007, 2010]. Therefore, we have a natural benchmark in that we can compare the fit of the genetic profile in the validation sample to the heritability. It is this useful ability to benchmark that makes the choice of a measure of goodness-of-fit in genetic profiling different from that of nongenetic predictors.

Coefficients of determination (R^2) for binary responses have been used in measuring the goodness-of-fit of models containing genetic predictors of human disease [Baneshi et al., 2010; Barrett et al., 2008, 2009, de Cid et al., 2009; Demirkan et al., 2011; Gharavi et al., 2011; Janssens et al., 2011; Labruna et al., 2011; Lyssenko et al., 2008; Painter et al., 2011; Purcell et al., 2009; Richards et al., 2011; Sarafidis

et al., 2007; Shea et al., 2010; Study, 2010; Tassone et al., 2000; The International Multiple Sclerosis Genetics Consortium, 2010; Vaidya et al., 2010; Witte and Hoffmann, 2011]. In the statistical literature, various pseudo-R^2 measures for binary responses have been proposed [Cox and Snell, 1989; DeMaris, 2002; Efron, 1978; McFadden, 1974; McKelvey and Zavoina, 1975; Nagelkerke, 1991; Veall and Zimmermann, 1996]. However, there is no standard or consensus measure because the concept of residual variance is not easily defined on the observed disease scale [Menard, 2000; Nagelkerke, 1991]. Most of the pseudo-R^2 measures use the likelihood functions from logistic or probit models that are based on the observed disease scale. This causes the obtained R^2 to be different from its value on the underlying liability scale, and therefore obscures comparisons with heritability, since the latter is usually expressed on the scale of liability. Furthermore, most case-control studies from which the precision of genetic profiles are estimated are ascertained, and traditional R^2 measures are not invariant with respect to ascertainment because they are based on goodness-of-fit statistics of the actual (ascertained) data. This complicates comparisons with other studies or inference about the population. In the literature, the effect of ascertainment has been poorly addressed or ignored [Barrett et al., 2008, 2009; Cubiella et al., 2010; Kochi et al., 2010; Peel et al., 2006, 2007].

In addition to an R^2 statistic to measure the goodness-of-fit of a genetic profile, the area under the curve (AUC) of receiver-operator characteristic (ROC) is frequently used to assess the precision with which a genetic predictor can correctly classify individuals into those that will become affected and those that will not. ROC curves have an advantage that they are not affected by ascertainment of the sample in which the goodness-of-fit of the genetic predictor is tested. Although the AUC can be interpreted as an R^2 on the liability scale [Wray et al., 2010], the AUC statistic does not provide a direct measure of how well the predictor performs relative to capturing all genetic variation or relative to the maximum value it can attain from genetic data [Wray et al., 2010].

In this study, we use a liability threshold model to establish the relationship between the probability of disease on the observed scale and an underlying scale of liability. We propose a novel measure of R^2 that is based upon a transformation between the observed probability scale and underlying liability scale. The R^2 on the liability scale can be obtained from a linear, logit, or probit model. Furthermore, we are interested in obtaining an R^2 at the population level even when the validation sample is ascertained. We used a modified version of the transformation between the observed and liability scale that corrects for bias due to ascertainment in case-control studies. Therefore, we obtain an R^2 measure on the scale of liability that can be compared directly to heritability.

MATERIALS AND METHODS

RELATIONSHIP BETWEEN THE OBSERVED PROBABILITY SCALE AND THE PROBIT OR LOGIT LIABILITY SCALE

Liability of disease is assumed to be the sum of environmental and additive genetic factors that are sampled from independent normal distributions. The model for liability can be written as,

$$l_i = \mu + g_i + e_i, \tag{1}$$

where l_i is the liability "phenotype" for the ith individual μ is the overall mean, g_i the random genetic effect on the liability scale, and e_i is the residual. For most of the theoretical derivations and simulations, we make the distributional assumptions that g and e are independently normally distributed with variances σ_g^2 and σ_e^2. For the theoretical validation and analyses of simulated data, we used g_i as an explanatory variable in linear, probit, or logistic models to validate the relationship between the observed disease scale and underlying liability scale. For real data, where g_i is not observed, we can use its estimate generated from genetic marker data and effect sizes estimated from independent data [Baneshi et al., 2010; Barrett et al., 2009; Gail, 2008; Lyssenko et al., 2008; Purcell et al., 2009; Wacholder et al., 2010]. In the Discussion section, we discuss the consequences of estimating g with error. Liability l is $\sim N(0,1)$, and the proportion of variance on the liability scale due to the genetic profile is $h_l^2 = \sigma_g^2$. In the liability threshold model, all affected individuals have a liability phenotype exceeding a certain threshold value t. This leads to observations (y) that are 0 or 1 for unaffected and affected individuals, with a Bernoulli distribution with a probability, p, i.e., $y \sim Bern(p)$.

For analysis of data, a generalized linear model can be used to link probabilities to effects on a linear scale. Using a logit link, the probability of disease p_i for individual i can be written as a function of linear predictors as,

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \mu_{\text{logit}} + b_{\text{logit}} g_i,$$

where g_i is an explanatory variable and a measure of genetic value for an individual; for theoretical validation and analyses of simulated data, we used the g_i defined in (1) above. For real data, where g_i is not observed, g_i is the estimated genetic predictor (profile score). The term μ_{logit} and b_{logit} are regression coefficients for the mean and genetic effects estimated in a logistic regression. The probability $p_i = \frac{1}{1+\exp(-\mu_{\text{logit}} - b_{\text{logit}} g_i)}$. The log-likelihood for the logistic regression is,

$$\ln L = \sum_{i=1}^{N_{\text{case}}} \ln \frac{1}{1+\exp(-\mu_{\text{logit}} - b_{\text{logit}} g_i)} + \sum_{i=1}^{N_{\text{control}}} \ln\left(1 - \frac{1}{1+\exp(-\mu_{\text{logit}} - b_{\text{logit}} g_i)}\right). \tag{2}$$

Similarly, a probit model is,

$$p_i = \Phi(\mu_{\text{probit}} + b_{\text{probit}} g_i),$$

where Φ is the standard cumulative density function, and μ_{probit} and b_{probit} are regression coefficients for the mean and genetic effects estimated in a probit model. The log-likelihood for the probit model is,

$$\ln L = \sum_{i=1}^{N_{\text{case}}} \ln[\Phi(\mu_{\text{probit}} + b_{\text{probit}} g_i)] + \sum_{i=1}^{N_{\text{control}}} \ln[1 - \Phi(\mu_{\text{probit}} + b_{\text{probit}} g_i)]. \quad (3)$$

Under the null model without the genetic effects, the probability can be expressed as $p_i = \bar{y} = \hat{K}$, where $\hat{K}$ is the mean proportion of cases in the sample, and the likelihood for the null model is,

$$\ln L_{\text{null}} = \sum_{i=1}^{N_{\text{case}}} \ln(K) + \sum_{i=1}^{N_{\text{control}}} \ln(1-K) = K \cdot N \cdot \ln(K) + (1-K) \cdot N \cdot \ln(1-K). \quad (4)$$

In the classical liability threshold model (1), the probability of an individual being affected given his or her genetic value can be derived using normal distribution theory [Dempster and Lerner, 1950], hence assuming that total liability follows a normal distribution. The relationship with the probit estimates of the parameters is,

$$p_i = \frac{1}{\sqrt{2\pi\sigma_e^2}} \int_{t-g_i}^{\infty} e^{-x^2/2\sigma_e^2} dx = \Phi(\mu_{\text{probit}} + b_{\text{probit}} g_i).$$

Dempster and Lerner [1950] showed that the additive genetic value expressed as a probability on the observed disease scale can be written as a linear function of the additive genetic value on the liability scale,

$$p_i = c + \hat{u}_i = \mu_{\text{obs}} + b_{\text{obs}} g_i = \mu_{\text{obs}} + z g_i, \quad (5)$$

where c is a constant, u is the genetic value on the observed scale, and μ_{obs} and b_{obs} are regression coefficients for the mean and genetic effects on the observed scale estimated in a linear model with 0, 1 observations. According to the Robertson transformation (1950), the regression coefficient for the genetic effects is the same as the probability density at the threshold t, i.e. $b_{\text{obs}} = \text{cov}(y, g)/\text{var}(g) = [E(y \cdot g) - E(y)E(g)]/h_l^2 = Km = z$, where m is the mean liability for cases and z is the height of a normal density curve at the point that truncates the proportion K in the upper tail. Therefore, the likelihood (2) and (3) can be approximated as,

$$\ln L \cong \sum_{i=1}^{N_{\text{case}}} \ln(\mu_{\text{obs}} + z g_i) + \sum_{i=1}^{N_{\text{control}}} \ln[1 - (\mu_{\text{obs}} + z g_i)]. \quad (6)$$

PSEUDO-R^2 MEASURES BASED ON THE LIKELIHOOD FUNCTION

The linear approximation of the likelihood function (6) implies that the likelihood function is based on probabilities on the observed disease scale, and not based on the logit or probit liability scales. This explains why pseudo-R^2 based on the likelihood [Cox and Snell, 1989; McFadden, 1974; Nagelkerke, 1991] do not give an appropriate interpretation when measuring the goodness-of-fit of the linear predictor for the logit ($\mu_{\text{logit}} + b_{\text{logit}} g_i$) or probit model ($\mu_{\text{probit}} + b_{\text{probit}} g_i$). Since observations and underlying explanatory factors are not on the same scale for binary traits, it has been observed that the pseudo-R^2 based on the likelihood never reach one even when a model has a perfect fit [Cox and Snell, 1989; Nagelkerke, 1991]. For example, the R^2 proposed by Cox and Snell (C&S) is,

$$R^2_{\text{C\&S}} = 1 - \left[\prod_{i=1}^{N_{\text{case}}} \left(\hat{K} \Big/ \frac{1}{1 + \exp(-\hat{\mu}_{\text{logit}} - \hat{b}_{\text{logit}} g_i)} \right) \times \prod_{i=1}^{N_{\text{control}}} \left((1-K) \Big/ \frac{\exp(-\hat{\mu}_{\text{logit}} - \hat{b}_{\text{logit}} g_i)}{1 + \exp(-\hat{\mu}_{\text{logit}} - \hat{b}_{\text{logit}} g_i)} \right) \right]^{2/N} \cong 1 - \left[\prod_{i=1}^{N_{\text{case}}} \left(\frac{\hat{K}}{\hat{\mu}_{\text{obs}} + \hat{b}_{\text{obs}} g_i} \right) \prod_{i=1}^{N_{\text{control}}} \left(\frac{1-\hat{K}}{1 - (\mu_{\text{obs}} + \hat{b}_{\text{obs}} g_i)} \right) \right]^{2/N}.$$

This equation shows that $1 - R^2_{\text{C\&S}}$ is the mean squared ratio of the probability explained by random chance (i.e. the numerators) over that explained by random chance plus additional genetic factors (i.e. the denominators), which are obviously on the observed probability scale. In a linear model with 0, 1 responses on the observed probability scale, the numerators are analogous to the mean squared errors in the full model, i.e. $y = \mu + g + e$, and the denominators are analogous to the mean squared errors in the null model, i.e. $y = \mu + e$. Therefore, $R^2_{\text{C\&S}}$ can be approximated as,

$$R^2_{\text{C\&S}} \cong 1 - \frac{\sum_i^N (y_i - \hat{y})^2/N}{\sum_i^N (y_i - \bar{y})^2/N} = 1 - \frac{\hat{\sigma}_e^2}{\hat{K}(1-\hat{K})},$$

where $\hat{\sigma}_e^2$ is estimated residual variance on the observed probability scale that is a proportion of the total variance unexplained by the genetic factor. If liability is normally distributed then the variance on the observed scale removed by the genetic profile is approximately equal to $z^2 h_l^2$ and the total variance on the observed scale is $K(1-K)$, so that the residual on that scale is the difference between the two. Therefore, an approximation of the expectation of $R^2_{\text{C\&S}}$ can be written as,

$$E\left(R^2_{\text{C\&S}}\right) \cong 1 - \frac{K(1-K) - z^2 h_l^2}{K(1-K)} = \frac{z^2 h_l^2}{K(1-K)}.$$

This expression shows why Cox and Snell's R^2 is approximately equal to R^2 on the observed scale in a linear model, although the difference increases with extremely high heritability [Cox and Wermuth, 1992] (also see Table I). Nagelkerke [1991] tried to correct Cox and Snell's R^2 by scaling it by the maximum value it can ever attain, i.e. $R^2_N = R^2_{\text{C\&S}}/R^2_{\max}$ where $R^2_{\max} = 1 - K^{2K} \cdot (1-K)^{2(1-K)}$ from (4). However, this adjustment is not appropriate if the aim is to measure the goodness-of-fit of models on the scale of liability.

TABLE I. Brief description of R^2 measures used in this study and their theoretical expectation

Brief description	Notation and formula	Expectation
R^2 on the observed scale	$R^2_o = 1 - \frac{\sum_i^N (y_i - \hat{y})^2}{\sum_i^N (y_i - \bar{y})^2}$	$h^2_l \frac{z^2}{K(1-K)}$
Cox and Snell's R^2 on the observed scale	$R^2_{C\&S} = 1 - \left\{ \frac{\text{Likelihood}_{\text{null}}}{\text{Likelihood}_{\text{full}}} \right\}^{2/N}$	$h^2_l \frac{z^2}{K(1-K)}$
Nagelkerke's R^2 on the observed scale	$R^2_N = \frac{R^2_{C\&S}}{1-(\text{Likelihood}_{\text{null}})^{2/N}}$	$\frac{R^2_{C\&S}}{1-K^{2K} \cdot (1-K)^{2(1-K)}}$
R^2 on the liability scale	$R^2_l = R^2_o \frac{\hat{K}(1-\hat{K})}{z^2}$	h^2_l
R^2 on the probit liability scale	$R^2_{\text{probit}} = \frac{\text{var}(\hat{b}_{\text{probit}} g_i)}{\text{var}(\hat{b}_{\text{probit}} g_i)+1}$	h^2_l
R^2 on the logit liability scale	$R^2_{\text{logit}} = \frac{\text{var}(\hat{b}_{\text{logit}} g_i)}{\text{var}(\hat{b}_{\text{logit}} g_i)+3.29}$	h^2_l
R^2 on the liability scale using AUC	$R^2_{\text{AUC}} = \frac{2Q^2}{(m_2-m)^2+Q^2 m(m-t)+m_2(m_2-t)}$	h^2_l
R^2 on the liability scale when using ascertained case-control studies	$R^2_{l_{cc}} = \frac{R^2_{o_{cc}} C}{1+R^2_{o_{cc}} \theta C}$	h^2_l

y, observations that are 0 or 1 for unaffected and affected individuals; h^2_l, heritability on the liability scale, in this context the proportion of variance on the liability scale explained by the genetic profile; K, population prevalence; z, the height of a normal density curve at the point according to K; g, the sum of all additive genetic factors in the estimated genetic predictor; b, regression coefficient from generalized linear model; m, the mean liability for cases; m_2, the mean liability for controls; t, the threshold on the normal distribution that truncates the proportion of disease prevalence K; Q, the inverse of the cumulative density function of the normal distribution up to values of AUC; C and θ, correcting factors for ascertainment.

R^2 ON THE LIABILITY SCALE

In order to derive R^2 on the liability scale, we first obtain the R^2 on the observed scale using linear regression. In a linear model with 0, 1 observation, the R^2 on the observed probability scale can be written as,

$$R^2_o = 1 - \left(\frac{\text{Likelihood}_{\text{null}}}{\text{Likelihood}_{\text{full}}} \right)^{2/N} = 1 - \frac{\sum_i^N (y_i - \hat{y})^2}{\sum_i^N (y_i - \bar{y})^2} = \frac{\text{var}(\hat{b}_{\text{obs}} g_i)}{\hat{K}(1-\hat{K})}, \quad (7)$$

where $\text{var}(\hat{b}_{\text{obs}} g_i)$ (or $\text{var}(z g_i)$) is the variance due to the explanatory variable (genetic variance) on the observed probability scale. Hence, R^2_o measures a portion of the total variance explained by the genetic factor on the observed probability scale. This proportion can be transformed to that on the liability scale using the Robertson transformation [Dempster and Lerner, 1950],

$$R^2_l = R^2_o \frac{\hat{K}(1-\hat{K})}{z^2} = \frac{\text{var}(g)}{\text{var}(l)} = \text{var}(g). \quad (8)$$

This concept of R^2 on the liability scale can be simply extended to probit or logit models. In a probit model, the R^2 on the probit liability scale can be directly obtained as the variance explained by linear predictors as a proportion of the total variance on the probit liability scale, that is,

$$R^2_{\text{probit}} = \frac{\text{var}(\hat{b}_{\text{probit}} g_i)}{\text{var}(\hat{b}_{\text{probit}} g_i) + \text{var}(e)}, \quad (9)$$

where the residual variance is defined as $\text{var}(e) = 1$ in the probit model. Since the assumption of normality on the scale of liability is assumed for both (8) and (9), their expectations are identical. Equation (8) is based upon an analysis on the 0–1 scale followed by a transformation, whereas Equation (9) is based upon a generalized linear model analysis.

Similarly, assuming that the liability has a logistic distribution, R^2 on the logit liability scale can be obtained with the residual variance of $\text{var}(e) = \pi^2/3 = 3.29$ as,

$$R^2_{\text{logit}} = \frac{\text{var}(\hat{b}_{\text{logit}} g_i)}{\text{var}(\hat{b}_{\text{logit}} g_i) + \text{var}(e)}. \quad (10)$$

McKelvey and Zavoina [1975] were the first to propose an R^2 measure expressed on an underlying latent scale using a generalized linear model. Equation (9) implements this for the probit link function and Equation (10) for the logit link function, and this R^2 is widely used. The derivation for the liability threshold model has not been considered previously (Equation (8)). Although the threshold model and the generalized linear model are equivalent, the formulae for R^2 in Equations (9) and (10) are different to that in Equation (8). R^2 values from Equations (9) and (10) are based on estimated linear predictors in logit or probit models, whereas that from Equation (8) was based on a transformation from R^2 on the observed scale that was based on the likelihood. Importantly, the transformation in Equation (8) can be modified to correct bias in ascertained case-control studies (see next sections).

R^2 ON THE LIABILITY SCALE FROM AUC

AUC is a useful statistic of the precision of predicting the genetic risk of disease [Janssens et al., 2006; Wray et al., 2010]. Using estimated AUC, R^2 on the liability scale can be obtained [Wray et al., 2010]. Estimation in this approach is independent of the relative proportion of cases and controls even if there is ascertainment in the case-control study. However, the estimation become less accurate for high heritabilities [Wray et al., 2010]. Given K and AUC, R^2 can be obtained as,

$$R^2_{\text{AUC}} = \frac{2Q^2}{(m_2 - m)^2 + Q^2 m(m - t) + m_2(m_2 - t)}, \quad (11)$$

where $Q = \Phi^{-1}(\text{AUC})$, and $m_2 = -mK/(1 - K)$.

R^2 ON THE LIABILITY SCALE FOR ASCERTAINED CASE-CONTROL STUDIES

In genetic epidemiology studies, case-control designs are widely used where cases are usually oversampled relative to the prevalence in the population. In this situation, there is no R^2 measure that is estimated on the liability scale and corrects for ascertainment. We consider the same liability model (1) but when samples are ascertained in a case-control study. According to (5), the genetic value on the observed scale (u_{cc}) for an individual in a case-control study is,

$$u_{cc} = c + b_{cc} g_{cc}, \quad (12)$$

where $b_{cc} = \text{cov}(y_{cc}, g_{cc})/\text{var}(g_{cc}) = z\frac{P(1-P)}{K(1-K)}\frac{\sigma^2_g}{\sigma^2_{g_{cc}}}$, where P is the proportion of cases in the case-control sample. The details of this derivation are in Appendix A. From (12), the variance explained by the genetic factor on the observed scale is [Lee et al., 2011],

$$\begin{aligned}\sigma^2_{u_{cc}} &= b^2_{cc}\sigma^2_{g_{cc}} = \left[z\frac{P(1-P)}{K(1-K)}\frac{\sigma^2_g}{\sigma^2_{g_{cc}}}\right]^2 \sigma^2_{g_{cc}} \\ &= \left[z\frac{P(1-P)}{K(1-K)}\right]^2 \frac{\sigma^2_g}{\sigma^2_{g_{cc}}}\sigma^2_g. \end{aligned} \quad (13)$$

R^2_o is a proportion of the total variance explained by the genetic factor on the observed probability scale (7), and we define $R^2_{o_{cc}}$ as that proportion for an ascertained case-control study. Therefore,

$$R^2_{o_{cc}} = \frac{\sigma^2_{u_{cc}}}{P(1-P)} = \left[z\frac{\sqrt{P(1-P)}}{K(1-K)}\right]^2 \frac{\sigma^2_g}{\sigma^2_{g_{cc}}}\sigma^2_g. \quad (14)$$

Finally, we express the proportion of the total variance explained by the genetic factor on the liability scale (8), corrected for ascertainment. This parameter $R^2_{l_{cc}}$ can be derived from (14) as (Appendix A),

$$R^2_{l_{cc}} = \sigma^2_g = \frac{R^2_{o_{cc}} C}{1 + R^2_{o_{cc}}\theta C}, \quad (15)$$

where $C = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}$ and $\theta = m\frac{P-K}{1-K}(m\frac{P-K}{1-K} - t)$.

To summarize the theoretical sections, we have derived an expression for the proportion of variation in liability explained by the genetic profile in the population, using an estimate of the proportion of variation explained by the profile on the observed 0–1 scale in an ascertained case-control sample. The expression (Equation (15)) uses the goodness-of-fit R^2 on the 0–1 scale and then transforms it to the liability scale whilst adjusting for ascertainment.

SIMULATION STUDY

In a simulation study, genetic (g) and residual values (e) were independently generated from random normal distributions with means of zero and variances of σ^2_g and σ^2_e, respectively. The value for σ^2_e was chosen such that the desired proportion of variation in liability due to the genetic profile was obtained. Liability for each individual was $l = g + e$. Disease status for each individual was determined by comparing l with the threshold of liability determined by the population prevalence. In this study, a population prevalence $K = 0.5$, 0.1, or 0.01 was used with 10,000 individual observations. Therefore, for the case-control designs, we had samples of 5,000, 1,000, and 100 cases and 5,000, 9,000, and 9,900 controls for $K = 0.5$, 0.1, and 0.01, respectively. When testing ascertained samples, cases were oversampled such that the number of cases and controls was approximately equal, i.e. $P = 0.5$. So for the ascertained case-control designs, we had samples of 5,000 cases and 5,000 controls for $K = 0.1$ or 0.01. In an alternative simulation, genetic and residual values were independently generated from logistic distributions with a means of zero and variances of σ^2_g and $\sigma^2_e = \pi^2/3$, respectively. The value for σ^2_g was chosen such that the desired proportion of variation in liability due to the genetic profile was obtained. Disease status for each individual was decided given his or her liability and the threshold determined by the population prevalence according to the logistic distribution.

For analyses of the simulated data, we used a linear, logit, or probit model where disease status was used as 0, 1 observations (y), and genetic values on the liability scale were used as explanatory variables [Cox and Wermuth, 1992). Using those models, we obtained several kinds of R^2 measures (Table I). First, we obtained R^2_o from a linear regression using (7). Second, we used Cox and Snell's method [Cox and Snell, 1989]. Third, we used Nagelkerke's scale method [Nagelkerke, 1991]. Fourth, we transformed R^2_o to R^2_l on the liability scale using (8). Fifth and sixth, we used the variance explained by linear predictors proportional to the total variance on the probit (9) and logit scale (10), respectively. Seventh, we obtained R^2_{AUC} from AUC estimated from a probit model (a linear or logit model gave the same results) using (11). These methods and notations are briefly described in Table I, and pseudo-R codes for them are shown in the Appendix B. In testing ascertained case-control samples, all the same methods were applied except that $R^2_{o_{cc}}$ was transformed to $R^2_{l_{cc}}$ using (15).

RESULTS

ESTIMATED R^2 MEASURES USING SEVERAL METHODS WITH SIMULATED DATA

In Figure 1, R^2 values on the observed scale ($R^2_{\text{C\&S}}$ and R^2_N) and liability scale (R^2_{probit},R^2_{logit}, R^2_l, and R^2_{AUC}) are

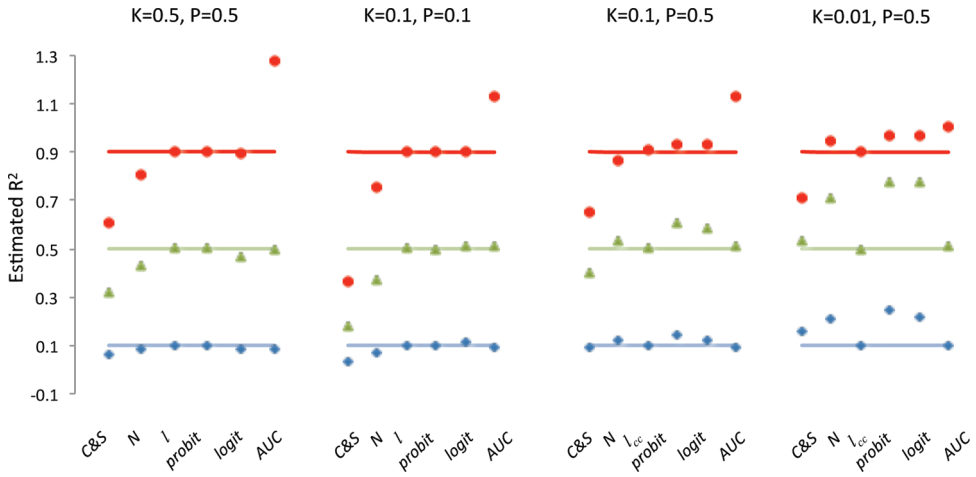

Fig. 1. Estimated coefficients of determination using simulated data. The true proportion of variance explained by the genetic profile was simulated as 0.1, 0.5, and 0.9 under a normal distribution. Various combinations of population prevalence (*K*) and the proportion of cases in the case-control study (*P*) were simulated. The first two figures are from simulations without ascertainment (i.e. $K = P$) and the last two figures are from simulations with ascertainment (i.e. $K < P$). Several R^2 measures were used and compared, i.e. Cox and Snell's R^2 on the observed scale (C&S), Nagelkerke's R^2 on the observed scale (*N*), R^2 on the liability scale transformed from linear model (*l* for population samples and l_{cc} for ascertained case-control samples), R^2 on the probit liability scale (probit), R^2 on the logit liability scale (logit), and R^2 on the liability scale using AUC (AUC). The three horizontal lines represent the true values. The colored lines are for heritability of 0.1 (blue), 0.3 (green), and 0.9 (red). Closed symbols represent the average of the estimated R^2 from each method.

presented when using population prevalence $K = 0.5$ and 0.1 without ascertainment, or $K = 0.1$ and 0.01 with ascertainment under a normal distribution. The results for R^2_o are not presented because the values for R^2_o and $R^2_{C\&S}$ are very similar unless heritability is very high, as expected from Equation (6). In Supplementary Material Tables SI–SIV, we report the results from seven kinds of R^2 measures for $K = 0.5$, 0.1, and 0.01 without ascertainment, and for $K = 0.1$ and 0.01 with ascertainment, both under a normal or logistic distribution of liability.

Under a normal distribution without ascertainment (i.e. $K = P$ in Fig. 1), the values for R^2_N were higher than $R^2_{C\&S}$; however, they were still much lower than the true values on the scale of liability. However, the values for R^2_l on the liability scale were unbiased and close to the true values (Fig. 1). For R^2_{probit} on the probit liability scale, the values were very similar to R^2_l and close to the true values. For R^2_{logit} on the logit liability scale, the values were similar to the true values with slight bias. The values for R^2_{AUC} were close to the true values with low heritabilities; however, they were overestimated with high heritabilities (Fig. 1).

Under a normal distribution with ascertainment (i.e., $K < P$ in Fig. 1), only $R^2_{l_{cc}}$ values gave unbiased and correct values that were transformed from $R^2_{o_{cc}}$ using (15) (Fig. 1). We note that R^2_{AUC} values with ascertained samples were very similar to those without ascertained samples, showing that R^2_{AUC} is not affected by ascertainment, because AUC, on which it is based, is known not to be affected by ascertainment. It was shown that ascertained samples resulted in biased estimation for the values for R^2_{probit} and R^2_{logit} that gave correct values when using unascertained population samples. We show in the Appendix C that a weighted probit model produces unbiased estimates for the normal distribution, although the weighted scheme does not fully use all information.

Under a logistic distribution without ascertainment (Table SIII in Supplementary Material), R^2 values on the observed scale ($R^2_{C\&S}$ and R^2_N) did not agree with the true values, as expected. The values for R^2_l and R^2_{AUC} were different from the true values. This was due to the fact that the transformation based on a normal distribution is not valid for a logistic distribution. The values for R^2_{probit} were slightly biased because the normality assumption in the probit model is violated when the actual distribution is logistic. Only R^2_{logit} values were unbiased and close to the true values (Fig. 1). Using a logistic distribution, we also tested and estimated R^2 values for ascertained case-control studies (Table SIV in Supplementary Material). In this situation, no method gave correct estimates. Again, a weighting scheme in a logistic model can be used to produce unbiased estimates for the logistic distribution (Appendix C).

COMPARING OBSERVED AND EXPECTED VALUES

The expected value for each R^2 can be obtained as described in Table I that gives approximate relationships between the R^2 values. Table II shows the ratio of the observed estimated value over its expectation under a normal distribution of liability. Without ascertainment ($K = P$), the ratios for R^2_l and R^2_{probit} were close to one, indicating that the observations and expectations agreed very well. The observations and expectations for $R^2_{C\&S}$, R^2_N, and R^2_{AUC} agreed approximately unless the true heritability was high. The ratio for R^2_{logit} deviated from one, probably because the liability had a normal distribution, not a logistic distribution (Table II).

When using ascertained case-control studies ($K < P$), the patterns for the ratio of observed and expected value for R^2 were similar to those without ascertainment except

TABLE II. The ratio of estimated R^2 over its expectation using several methods under a normal distribution of liability

	Observed scale		Liability scale			
True h^2[a]	$R^2_{C\&S}$	R^2_N	R^2_l	R^2_{probit}	R^2_{logit}	R^2_{AUC}
$K = P = 0.5$						
0.1	1.02	1.02	1.02	1.02	0.83	0.85
	(0.06)	(0.06)	(0.06)	(0.06)	(0.05)	(0.06)
0.5	1.01	1.01	1.00	1.00	0.94	1.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
0.9	1.06	1.06	1.00	1.00	1.00	1.42
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
$K = P = 0.1$						
0.1	1.00	1.01	1.00	1.01	1.12	0.94
	(0.10)	(0.10)	(0.11)	(0.10)	(0.11)	(0.10)
0.5	1.04	1.03	1.00	1.00	1.03	1.02
	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.04)
0.9	1.17	1.18	1.00	1.00	1.00	1.26
	(0.03)	(0.01)	(0.03)	(0.01)	(0.01)	(0.02)
$K = 0.1, P = 0.5$						
0.1	1.00	1.00	1.00	1.43	1.19	0.94
	(0.06)	(0.06)	(0.06)	(0.09)	(0.08)	(0.06)
0.5	1.00	1.01	1.01	1.22	1.18	1.03
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
0.9	1.03	1.03	1.01	1.03	1.04	1.25
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
$K = 0.01, P = 0.5$						
0.1	1.00	1.00	1.00	2.49	2.15	0.99
	(0.05)	(0.05)	(0.04)	(0.12)	(0.11)	(0.06)
0.5	0.98	0.98	1.00	1.55	1.55	1.02
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
0.9	0.96	0.96	1.00	1.07	1.08	1.11
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)

[a]The true proportion of variance explained by the genetic profile.
The expectation was obtained as described in Table I.
Standard deviation over 30 replicates is in the bracket.

R^2_{probit}. The observations and expectations for $R^2_{l_{cc}}$ agreed well (Table II). With low and moderate heritability, the observed and expected values for $R^2_{C\&S}$, R^2_N, and R^2_{AUC} agreed (Table II). However, the ratio of observed and expected values for R^2_{probit} and R^2_{logit} substantially deviated from one. This was due to the fact that R^2_{probit} and R^2_{logit} were not corrected for ascertainment bias. We report complete results for the ratio of observed and expected values in Supplementary Material Tables SV–SVI.

DISCUSSION

It is reasonable to assume that there is underlying liability for complex disease (Falconer and Mackay, 1996], and recent empirical findings from genome-wide association studies are consistent with highly polygenic models for common disease [Antoniou and Easton, 2003; Pharoah et al., 2002; Purcell et al., 2009; The International Multiple Sclerosis Genetics Consortium, 2010; Witte and Hoffmann, 2011]. If this assumption is valid, it is desirable to have coefficients of determination on the same scale as liability because then the goodness-of-fit can be compared across studies and traits. We showed that pseudo-R^2 statistics based on the likelihood function (e.g. $R^2_{C\&S}$) are on the observed probability scale. This is the reason why such R^2 are inappropriate to measure the goodness-of-fit of models, i.e. it never reaches unity even when there is a perfect model fit. Nagelkerke R^2 is adjusted for the maximum value so that it may reach unity; however, the adjustment is inappropriate to measure the goodness-of-fit of models on the liability scale. We derived and showed the relationship between the observed probability scale and the underlying liability scale. R^2 is a proportion of variance explained by explanatory genetic factors, and can be transformable between the observed and the liability scale. Given the simulation results, the R^2 values on the liability scale were much more appropriate in measuring the goodness-of-fit of models and interpreting model parameters. The concept of R^2 as a proportion of total variance explained by explanatory factors on the liability scale was suggested previously [McKelvey and Zavoina, 1975]. We explicitly show the relationship between R^2 on the observed and liability scale, and justified that R^2 on the liability scale is globally valid and comparable. Moreover, when samples were ascertained, an unbiased estimate of R^2 on the liability scale could be obtained, corrected for ascertainment bias using a modified version of the transformation.

The transformation of R^2 values on the liability scale depends on the distribution of underlying liability. The assumption of a normal distribution in obtaining R^2_l, $R^2_{l_{cc}}$, R^2_{probit}, and R^2_{AUC} was violated when the true liability had a logistic distribution (Fig. 1). If liability is the sum of many multiple independent random genetic and environmental factors, then the central limit theorem predicts that its distribution will tend to normality [Falconer and Mackay, 1996; Gibson, 2009; Valentin, 1999; Wray et al., 2010]. For this reason, the assumption of a normal distribution for liability to common disease seems reasonable.

In practice, with real data, genetic values on the liability scale are not observed (g_i) but estimated ($\hat{g}_i$). For example, $\hat{g}_i$ can be created from validated genome-wide significant Single Nucleotide Polymorphism (SNPs) or else from a large number of SNPs with effect sizes estimated in an independent discovery sample in a "profile scoring" approach [Chen et al., 2011; Meigs et al., 2008; Morrison et al., 2007; Purcell et al., 2009; The International Multiple Sclerosis Genetics Consortium, 2010; Wacholder et al., 2010; Wray et al., 2007]. In these examples, effect sizes are estimated in a fixed effects framework, and the resulting predictor will be estimated with error (ε_i), such that $\hat{g}_i = g_i + \varepsilon_i$. The effect of such errors on the R^2 values investigated in this study is that the "heritability" of the predictor is lower than if it were estimated without error, and so the R^2 values will be lower. However, all calculations and simulations results are still valid if the variation in liability explained by $\hat{g}$ is substituted for the heritability of liability. If the predictor is estimated from random effects models and unbiased in the sense that the regression of g_i on $\hat{g}_i$ is unity [Goddard et al., 2009] then the R^2 values will be unbiased and will reflect the proportion of variance of liability explained by g_i.

We suggest that R^2 values on the liability scale should be used to measure the goodness-of-fit of models in which genetic profiles are used. They are consistent with the underlying scale, independent of population parameters such

as K and P, globally comparable between analytical models and methods and can be compared to heritability. Particularly, R^2_l and R^2_{probit} values are easily interpretable in relation to true heritability on the underlying liability scale when using population samples. When using ascertained case-control studies, $R^2_{l_{cc}}$ values, adjusted for ascertainment bias, is a useful measure with desirable properties.

ACKNOWLEDGMENTS

We acknowledge funding from the Australian National Health and Medical Research Council (grants 613672, 613601, and 1011506) the Australian Research Council (grants DP0770096, DP1093502, and FT0991360).

REFERENCES

Antoniou AC, Easton DF. 2003. Polygenic inheritance of breast cancer: implications for design of association studies. Genet Epidemiol 25(3):190–202.

Baneshi MR, Warner P, Anderson N, Edwards J, Cooke TG, Bartlett JMS. 2010. Tamoxifen resistance in early breast cancer: statistical modelling of tissue markers to improve risk prediction. Br J Cancer 102(10):1503–1510.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS. 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41(6):703–707.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot J-P, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40(8):955–962.

Chen H, Poon A, Yeung C, Helms C, Pons J, Bowcock AM, Kwok P-Y, Liao W. 2011. A genetic risk score combining ten psoriasis risk loci improves disease prediction. PLoS One 6(4):e19454.

Cox DR, Snell EJ. 1989. The analysis of binary data. London: Chapman and Hall.

Cox DR, Wermuth N. 1992. A comment on the coefficient of determination for binary responses. Am Stat 46:1–4.

Cubiella FJ, Nunez CL, Gonzalez VE, Garcia GMJ, Alves PMT, Martinez SI, Fernandez SJ. 2010. Risk factors associated with the development of ischemic colitis. World J Gastroenterol 16:4564–4569.

Daetwyler HD, Villanueva B, Woolliams JA. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3(10):e3395.

de Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, Dannhauser EN, Giardina E, Stuart PE, Nair R, Helms C, Escaramis G, Ballana E, Martin-Ezquerra G, Heijer Md, Kamsteeg M, Joosten I, Eichler EE, Lazaro C, Pujol RM, Armengol L, Abecasis G, Elder JT, Novelli G, Armour JAL, Kwok P-Y, Bowcock A, Schalkwijk J, Estivill X. 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet 41(2):211–215.

DeMaris A. 2002. Explained variance in logistic regression. Sociol Methods Res 31(1):27–74.

Demirkan A, Penninx BWJH, Hek K, Wray NR, Amin N, Aulchenko YS, van Dyck R, de Geus EJC, Hofman A, Uitterlinden AG, Hottenga JJ, Nolen WA, Oostra BA, Sullivan PF, Willemsen G, Zitman FG, Tiemeier H, Janssens ACJW, Boomsma DI, van Duijn CM, Middeldorp CM. 2011. Genetic risk profiles for depression and anxiety in adult and elderly cohorts. Mol Psychiatry 16:773–783.

Dempster ER, Lerner IM. 1950. Heritability of threshold characters. Genetics 35:212–236.

Efron B. 1978. Regression and ANOVA with zero-one data: measures of residual variation. J Am Stat Assoc 73:113–121.

Evans DM, Visscher PM, Wray NR. 2009. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet. 18:3525–3531.

Falconer DS, Mackay TFC. 1996. Introduction to quantitative genetics. Harlow, Essex, UK: Longman.

Gail MH. 2008. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. J Natl Cancer Inst 100(14):1037–1041.

Gharavi AG, Kiryluk K, Choi M, Li Y, Hou P, Xie J, Sanna-Cherchi S, Men CJ, Julian BA, Wyatt RJ, Novak J, He JC, Wang H, Lv J, Zhu L, Wang W, Wang Z, Yasuno K, Gunel M, Mane S, Umlauf S, Tikhonova I, Beerman I, Savoldi S, Magistroni R, Ghiggeri GM, Bodria M, Lugani F, Ravani P, Ponticelli C, Allegri L, Boscutti G, Frasca G, Amore A, Peruzzi L, Coppo R, Izzi C, Viola BF, Prati E, Salvadori M, Mignani R, Gesualdo L, Bertinetto F, Mesiano P, Amoroso A, Scolari F, Chen N, Zhang H, Lifton RP. 2011. Genome-wide association study identifies susceptibility loci for IgA nephropathy. Nat Genet 43:321–327.

Gibson G. 2009. Decanalization and the origin of complex disease. Nat Rev Genet 10(2):134–140.

Goddard ME, Wray NR, Verbyla K, Visscher PM. 2009. Estimating effects and making predictions from genome-wide marker data. Statist Sci 24:517–529.

Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, van Duijn CM. 2006. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 8(7):395–400. DOI: 10.1097/01.gim.0000229689.18263.f4.

Janssens ACJW, Ioannidis JPA, Bedrosian S, Boffetta P, Dolan SM, Dowling N, Fortier I, Freedman AN, Grimshaw JM, Gulcher J, Gwinn M, Hlatky MA, Janes H, Kraft P, Melillo S, O'Donnell CJ, Pencina MJ, Ransohoff D, Schully SD, Seminara D, Winn DM, Wright CF, van Duijn CM, Little J, Khoury MJ. 2011. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. Eur J Hum Genet 19: doi:10.1038/ejhg.2011.27.

Kochi Y, Okada Y, Suzuki A, Ikari K, Terao C, Takahashi A, Yamazaki K, Hosono N, Myouzen K, Tsunoda T, Kamatani N, Furuichi T, Ikegawa S, Ohmura K, Mimori T, Matsuda F, Iwamoto T, Momohara S, Yamanaka H, Yamada R, Kubo M, Nakamura Y, Yamamoto K. 2010. A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. Nat Genet 42(6):515–519.

Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DJ, Chanock S. 2009. Beyond odds ratios [mdash] communicating disease risk based on genetic profiles. Nat Rev Genet 10(4):264–269.

Labruna G, Pasanisi F, Nardelli C, Caso R, Vitale DF, Contaldo F, Sacchetti L. 2011. High leptin/adiponectin ratio and serum triglycerides are associated with an "at-risk" phenotype in young severely obese patients. Obesity. 19:1492–1496.

Lee SH, Wray NR, Goddard ME, Visscher PM. 2011. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88:294–305.

Lyssenko V, Jonsson A, Almgren P, Pulizzi N, Isomaa B, Tuomi T, Berglund G, Altshuler D, Nilsson P, Groop L. 2008. Clinical risk factors, DNA variants, and the development of type 2 diabetes. New Engl J Med 359(21):2220–2232.

McFadden D. 1974. The measurement of urban travel demand. J Public Econ 3:303–328.
McKelvey RD, Zavoina W. 1975. A statistical model for the analysis of ordinal level dependent variables. J Math Sociol 4:103–120.
Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PWF, D'Agostino RB, Cupples LA. 2008. Genotype score in addition to common risk factors for prediction of type 2 diabetes. New Engl J Med 359(21):2208–2219.
Menard S. 2000. Coefficients of determination for multiple logistic regression analysis. Am Stat 54(1):17–24.
Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, Kane JP, Pankow JS, Devlin JJ, Willerson JT, Boerwinkle E. 2007. Prediction of coronary heart disease risk using a genetic risk score: the atherosclerosis risk in communities study. Am J Epidemiol 166(1):28–35.
Nagelkerke NJD. 1991. A note on a general definition of the coefficient of determination. Biometrika 78(3):691–692.
Painter JN, Anderson CA, Nyholt DR, Macgregor S, Lin J, Lee SH, Lambert A, Zhao ZZ, Roseman F, Guo Q, Gordon SD, Wallace L, Henders AK, Visscher PM, Kraft P, Martin NG, Morris AP, Treloar SA, Kennedy SH, Missmer SA, Montgomery GW, Zondervan KT. 2011. Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. Nat Genet 43:51–54.
Peel NM, McClure RJ, Hendrikz JK. 2006. Health-protective behaviours and risk of fall-related hip fractures: a population-based case-control study. Age Ageing 35(5):491–497.
Peel NM, McClure RJ, Hendrikz JK. 2007. Psychosocial factors associated with fall-related hip fractures. Age Ageing 36(2):145–151.
Pharoah PDP, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BAJ. 2002. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31:33–36.
Pharoah PDP, Antoniou AC, Easton DF, Ponder BAJ. 2008. Polygenes, risk prediction, and targeted prevention of breast cancer. New Engl J Med 358(26):2796–2803.
Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460(7256):748–52.
Richards AL, Jones L, Moskvina V, Kirov G, Gejman PV, Levinson DF, Sanders AR, Purcell S, Visscher PM, Craddock N, Owen MJ, Holmans P, O'Donovan MC. 2011. Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. Mol Psychiatry doi: 10.1038/mp.2011.11.
Sarafidis PA, Lasaridis AN, Nilsson PM, Pikilidou MI, Stafilas PC, Kanaki A, Kazakos K, Yovos J, Bakris GL. 2007. Validity and reproducibility of HOMA-IR, 1/HOMA-IR, QUICKI and McAuley's indices in patients with hypertension and type II diabetes. J Hum Hypertens 21(9):709–716.
Shea JL, Loredo-Osti JC, Sun G. 2010. Association of RBP4 gene variants and serum HDL cholesterol levels in the Newfoundland population. Obesity 18(7):1393–1397.
Study TIHC. 2010. The Major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science 330(6010):1551–1557.
Tassone F, Hagerman RJ, Taylor AK, Gane LW, Godfrey TE, Hagerman PJ. 2000. Elevated levels of FMR1 mRNA in carrier males: a new mechanism of involvement in the Fragile-X syndrome. Am J Hum Genet 66(1):6–15.
The International Multiple Sclerosis Genetics Consortium. 2010. Evidence for polygenic susceptibility to multiple sclerosis—the shape of things to come. Am J Hum Genet 86(4):621–625.
Vaidya VS, Ozer JS, Dieterle F, Collings FB, Ramirez V, Troth S, Muniappa N, Thudium D, Gerhold D, Holder DJ, Bobadilla NA, Marrer E, Perentes E, Cordier A, Vonderscher J, Maurer G, Goering PL, Sistare FD, Bonventre JV. 2010. Kidney injury molecule-1 outperforms traditional biomarkers of kidney injury in preclinical biomarker qualification studies. Nat Biotech 28(5):478–485.
Valentin J. 1999. Risk estimation for multifactorial diseases. Oxford, UK: ICRP by Elsevier Science Ltd.
Veall MR, Zimmermann KF. 1996. Pseudo-R^2 measures for some common limited dependent variable models. J Econ Surveys 10(3):241–259.
Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P, Rosner B, Berg CD, Brinton LA, Lissowska J, Sherman ME, Chlebowski R, Kooperberg C, Jackson RD, Buckman DW, Hui P, Pfeiffer R, Jacobs KB, Thomas GD, Hoover RN, Gail MH, Chanock SJ, Hunter DJ. 2010. Performance of common genetic variants in breast-cancer risk models. New Engl J Med 362(11):986–993.
Witte JS, Hoffmann TJ. 2011. Polygenic modeling of genome-wide association studies: an application to prostate and breast cancer. OMICS J Integr Biol 15(6):393–398.
Wray NR, Goddard ME, Visscher PM. 2007. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17:1520–1528.
Wray NR, Yang J, Goddard ME, Visscher PM. 2010. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6(2):e1000864.

APPENDIX A

TRANSFORMATION CORRECTED FOR ASCERTAINMENT

In ascertained case-control studies, the mean and variance for case-controls disease status (y_{cc}), disease liability (l_{cc}), and genetic liability (g_{cc}) following quantitative genetic theory [Falconer and Mackay, 1996] are,

$E(y_{cc}) = P$, which is the proportion of cases in the sample,

$\text{var}(y_{cc}) = P(1-P)$, which is the phenotypic variance on the observed scale in the case-control sample,

$$\begin{aligned} E(l_{cc}) &= Pm + (1-P)m_2 = Pm - (1-P)m[K/(1-K)] \\ &= m[(P-K)/(1-K)]. \end{aligned}$$

$$\begin{aligned} \text{var}(l_{cc}) &= E(l_{cc}^2) - E(l_{cc})^2 = P(1+mt) + (1-P)(1+m_2 t) \\ &\quad - m^2[(P-K)/(1-K)]^2 \\ &= 1 + Pmt - (1-P)tmK/(1-K) \\ &\quad - m^2[(P-K)/(1-K)]^2 \\ &= 1 - m[(P-K)/(1-K)]\{m[(P-K)/(1-K)] - t\} \\ &= 1 - \theta, \end{aligned}$$

where m is the mean liability for cases, $m_2 = -m[K/(1-K)]$ is the mean liability for controls, t is the threshold on the normal distribution truncating the proportion of disease prevalence K, and $\theta = m[(P-K)/(1-K)]\{m[(P-K)/(1-K)] - t\}$.

The mean of genetic liability depends on the mean liability phenotype of the cases and the heritability of liability,

$$\begin{aligned} E(g_{cc}) &= h_l^2 E(l_{cc}) = h_l^2[Pm + (1-P)m_2] \\ &= h_l^2 m[(P-K)/(1-K)]. \end{aligned}$$

The variance for genetic liability can be derived as [Daetwyler et al., 2008],

$$var(g_{cc}) = E\left(g_{cc}^2\right) - E(g_{cc})^2 = h_l^2\left(1 - h_l^2\theta\right).$$

For Equation (12) in main text, the regression of phenotype on the observed risk scale on genetic liability in the case-control study is,

$$\begin{aligned} b_{cc} &= \mathrm{cov}(y_{cc}, g_{cc})/\mathrm{var}(g_{cc}) \\ &= [E(y_{cc} \cdot g_{cc}) - E(y_{cc})E(g_{cc})]/\mathrm{var}(g_{cc}) \\ &= [P \cdot h_l^2 m - P \cdot h_l^2 m\{(P-K)/(1-K)\}]/var(g_{cc}) \\ &= \left[P \cdot h_l^2 m \left(1 - \frac{(P-K)}{(1-K)}\right)\right] \frac{1}{\sigma^2_{g_{cc}}} \\ &= z\frac{P(1-P)}{K(1-K)} \frac{\sigma^2_g}{\sigma^2_{g_{cc}}}. \left(\text{Note } m = \frac{z}{K}\right). \end{aligned}$$

The term $\frac{P(1-P)}{K(1-K)} \frac{\sigma^2_g}{\sigma^2_{g_{cc}}}$ quantifies the change of the regression coefficient in a regression of phenotype on the observed risk scale on genetic factors on the scale of liability due to ascertainment in a case-control study. In the absence of ascertainment ($P = K$), this term is 1.

From Equation (14), $R^2_{l_{cc}}$ can be derived as,

$$\begin{aligned} R^2_{l_{cc}} = \sigma^2_g &= R^2_{o_{cc}} \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)} \frac{\sigma^2_{g_{cc}}}{\sigma^2_g} \\ &= R^2_{o_{cc}} \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)} \frac{h_l^2\left(1 - h_l^2\theta\right)}{\sigma^2_g}. \end{aligned}$$

Since

$$R^2_{l_{cc}} = h_l^2 = \sigma^2_g,$$

and

$$C = \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)},$$

$$R^2_{l_{cc}} = R^2_{o_{cc}} C(1 - h_l^2\theta) = \frac{R^2_{o_{cc}} C}{1 + R^2_{o_{cc}} \theta C}.$$

APPENDIX B

PSEUDO-*R* CODE FOR THE METHODS DESCRIBED IN THE PAPER

```
nt = total number of the sample
ncase = number of cases
ncont = number of controls
thd = the threshold on the normal distribution
   which truncates the proportion of disease
    prevalence
K = population prevalence
P = proportion of cases in the case-control
   samples
thd = -qnorm(K,0,1)
zv = dnorm(thd) #z (normal density)
mv = zv/K #mean liability for case
mv2 = -mv*K/(1-K) #mean liability for controls
library(Design) #to call lrm
library(pROC) #to get AUC values
# R2 on the observed scale using a liner model
lmv = lm(y~g) # linear model
R2 = var(lmv$fitted.values)/(ncase/nt*ncont/nt)
# Cox & Snell R2
logf = logLik(glm(y~g,family = binomial(logit)))
logn = logLik(glm(y~1,family = binomial(logit)))
R2 = 1-exp((logn-logf)*(2/nt))
# Nagelkerke R2
lrmv2 = lrm(y~g) # a logistic model to get
   Nagelkerke's R2
R2 = lrmv2$stats[10]
# R2 on the probit liability scale using
   a probit model
pmv = glm(y~g,family = binomial(probit)) #
   probit model
R2 = var(pmv$linear.predictors)/(var(pmv$linear.
   predictors)+1)
# R2 on the logistic liability scale
lrmv = glm(y~g,family = binomial(logit)) #
   logistic model
R2 = var(lrmv$linear.predictors)/(var(lrmv$
   linear.predictors)+pi^2/3)
# R2 on the liability scale using AUC
aucv = auc(y,pmv$linear.predictors)
qv = qnorm(aucv[1]) #Q in equation (11)
R2 = 2*qv^2/((mv2-mv)^2+qv^2*mv*(mv-thd)+
   mv2*(mv2-thd))
# R2 on the liability scale using the
   transformation
lmv = lm(y~g) #linear model
R20 = var(lmv$fitted.values)/(ncase/nt*ncont/nt)
   #R2 on the observed scale
theta = mv*(P-K)/(1-K)*(mv*(P-K)/(1-K)-thd) #θ in
equation (15)
cv = K*(1-K)/zv^2*K*(1-K)/(P*(1-P)) #C in
   equation (15)
R2 = R20*cv/(1+R20*theta*cv)
```

APPENDIX C

WEIGHTED GLM ESTIMATES IN MEASURING R^2 FOR GENETIC PROFILE ANALYSIS USING ASCERTAINED CASE-CONTROL SAMPLES

When ascertained case-control samples are used, estimates from a probit model are biased because the normality assumption is violated (3). A weighted probit model can be used to obtain unbiased estimations, weighting cases for the proportion of ascertainment, i.e. $(1-P)K/[P(1-K)]$. Even after obtaining unbiased $\hat{b}_{\text{probit}}$, the R^2_{probit} from Equation (9) is still biased because there are still oversampled cases in the estimation. We used a proportion of cases, i.e. $(1-P)K/[P(1-K)]$, such that the number of cases and controls matched the population incidence, and estimated R^2_{probit} using Equation (9). The pseudo-*R* code for these processes is in the following.

```
# Weighted R2 from a weighted probit model
wv = (1-P)K/[P(1-K)] #weighting factor
```

```
wt = y+1
wt[wt = = 2] = wv #weighting array
pmv = glm(y~g,weights = wt, family = binomial
   (probit)) #weighted probit model
vr = runif(nt,0,1) #uniform random values
vsel = pmv$linear.predictors[y = = control
   | vr<wv] #select controls and a proportion
   (wv) of cases
R2 = var(vsel)/(var(vsel)+1)
```

In contrast to a probit model, estimates from a logistic model are not affected by sample ascertainment because the estimate is a function of odds ratios (Equation (2)). Therefore, we could obtain unbiased $\hat{b}_{\text{logit}}$ from a standard logistic regression. However, R^2_{logit} from Equation (10) is biased because of oversampled cases in the estimation. We used the same strategy as above to use a proportion of cases, i.e. $(1 - P)K/[P(1 - K)]$, and estimated R^2_{logit} using Equation (10). The pseudo-*R* code for obtaining weighted R^2_{logit} is in the following.

```
# Weighted R2 from a logistic model
lrmv = glm(y~g,family = binomial(logit))
   # logistic model
vr = runif(nt,0,1) #uniform random values
vsel = lrmv$linear.predictors[y = = control
   | vr<wv] #select controls and a proportion
   (wv) of cases
R2 = var(vsel)/(var(vsel)+pi^2/3)
```

Weighted R^2 was estimated and compared to the transformation method (Tables SVII and SVIII in Supplementary Material). For simulations using a normal distribution of liability, weighted R^2 from a weighted probit model was estimated (Table SVII in Supplementary Material). For simulations using a logistic distribution, weighted R^2 from a standard logistic model was obtained (Table SVIII in Supplementary Material).

Table SVII in Supplementary Material shows that weighted R^2 from the weighted probit estimates was much better than that from the unweighted probit estimates when using simulations of a normal distribution of liability. The performance of the weighted probit estimates was similar to that of transformation method (R^2_{lcc}) (Table I main text), although the standard deviations for the weighted R^2 were larger. When using simulations of a logistic distribution of liability, the weighted R^2 from a logistic model gave unbiased estimates, which were close to the true values (Table SVIII in Supplementary Material).