Dudbridge, F (2013) Power and predictive accuracy of polygenic risk scores. PLoS genetics, 9 (3). e1003348. ISSN 1553-7390 DOI: https://doi.org/10.1371/journal.pgen.1003348

Downloaded from: http://researchonline.lshtm.ac.uk/748772/

DOI: 10.1371/journal.pgen.1003348

Usage Guidelines

Please refer to usage guidelines at http://researchonline.lshtm.ac.uk/policies.html or alternatively contact researchonline@lshtm.ac.uk.

Available under license: http://creativecommons.org/licenses/by/2.5/

OPEN ACCESS Freely available online

PLOS | GENETICS

Power and Predictive Accuracy of Polygenic Risk Scores

Frank Dudbridge*

Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

Abstract

Polygenic scores have recently been used to summarise genetic effects among an ensemble of markers that do not individually achieve significance in a large-scale association study. Markers are selected using an initial training sample and used to construct a score in an independent replication sample by forming the weighted sum of associated alleles within each subject. Association between a trait and this composite score implies that a genetic signal is present among the selected markers, and the score can then be used for prediction of individual trait values. This approach has been used to obtain evidence of a genetic effect when no single markers are significant, to establish a common genetic basis for related disorders, and to construct risk prediction models. In some cases, however, the desired association or prediction has not been achieved. Here, the power and predictive accuracy of a polygenic score are derived from a quantitative genetics model as a function of the sizes of the two samples, explained genetic variance, selection thresholds for including a marker in the score, and methods for weighting effect sizes in the score. Expressions are derived for quantitative and discrete traits, the latter allowing for case/control sampling. A novel approach to estimating the variance explained by a marker panel is also proposed. It is shown that published studies with significant association of polygenic scores have been well powered, whereas those with negative results can be explained by low sample size. It is also shown that useful levels of prediction may only be approached when predictors are estimated from very large samples, up to an order of magnitude greater than currently available. Therefore, polygenic scores currently have more utility for association testing than predicting complex traits, but prediction will become more feasible as sample sizes continue to grow.

Citation: Dudbridge F (2013) Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet 9(3): e1003348. doi:10.1371/journal.pgen.1003348

Editor: Naomi R. Wray, Queensland Institute of Medical Research, Australia

Received May 24, 2012; **Accepted** January 16, 2013; **Published** March 21, 2013

Copyright: © 2013 Frank Dudbridge. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the Medical Research Council, grant number G1000718 (www.mrc.ac.uk). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: frank.dudbridge@lshtm.ac.uk

Introduction

Although individually significant markers in genome-wide association scans (GWAS) explain limited heritability of complex traits, evidence has been accruing that a considerable proportion of phenotypic variation can be explained by the ensemble of markers not achieving significance. Thus, while most of the specific genes underlying complex traits have yet to be identified, it is likely that many are represented on current genotyping products and specific identification is largely a matter of study size [1]. Polygenic score analysis has recently generated much interest for assessing the explanatory power of an ensemble of markers. A GWAS is conducted on an initial training sample, and the markers are ranked by their evidence for association, usually their *P*-values. An independent replication sample is then analysed by constructing, for each subject, a polygenic score consisting of the weighted sum of its trait-associated alleles, for some subset of top ranking markers. Two related but distinct applications of this score are then possible. Firstly, testing for association between the score and the trait in the replication sample can determine whether associated markers reside within those contributing to the score. Secondly and perhaps more usefully, the polygenic score can be used to predict individual trait values or risks of disease [2], potentially giving a predictor with better discrimination properties than one based on established markers only. Different considerations apply for these two applications, as the size of the replication sample has a direct bearing on the power of association testing, whereas the accuracy of individual predictions depends only on the size of the training sample.

The first successful application of polygenic score analysis to GWAS data was in schizophrenia [3], in which few individual markers were significant and the common disease common variant hypothesis remained in question. It was shown that a large mass, up to half, of all markers in one GWAS could be jointly associated with disease in a second sample, implying a polygenic component to disease risk that justified larger study sizes [4]. Furthermore, markers from schizophrenia GWAS could together be associated with bipolar disorder, and vice versa, establishing a common polygenic basis to those conditions, whereas such cross-prediction was not achieved with clinically distinct conditions such as cardiovascular disease. This common basis has further been exploited to discriminate sub-types of bipolar disorder [5].

Similar results using a large mass of markers have been obtained for other complex traits including multiple sclerosis [6], height [7], cardiovascular risk [8], rheumatoid arthritis [9] and body mass index [10]. In addition, several studies have demonstrated association of a score based on a limited number of top ranking markers [11–13]. In some cases, however, the polygenic association is less clear: studies of breast and prostate cancers have been inconclusive, owing in part to technical aspects in analysis but also, potentially, to their sample sizes [14,15]. An aim of the present work is to determine whether negative results from those studies could be explained by their sample size, or whether a true lack of polygenic effect is the more likely explanation.

Author Summary

Recently there has been much interest in combining multiple genetic markers into a single score for predicting disease risk. Even if many of the individual markers have no detected effect, the combined score could be a strong predictor of disease. This has allowed researchers to demonstrate that some diseases have a strong genetic basis, even if few actual genes have been identified, and it has also revealed a common genetic basis for distinct diseases. These analyses have so far been performed opportunistically, with mixed results. Here I derive formulae based on the heritability of disease and size of the study, allowing researchers to plan their analyses from a more informed position. I show that discouraging results in some previous studies were due to the low number of subjects studied, but a modest increase in study size would allow more successful analysis. However, I also show that, for genetics to become useful for predicting individual risk of disease, hundreds of thousands of subjects may be needed to estimate the gene effects. This is larger than most existing studies, but will become more common in the near future, so that gene scores will become more useful for predicting disease than has appeared to date.

Applications of polygenic scores to individual disease prediction have so far been less successful, although proof of concept has been established through simulations [2]. Several studies have shown that a limited number of top ranking markers can discriminate disease cases from unaffected subjects, but the degree of discrimination falls short both of clinical utility and the maximum achievable from genetic data [16–18]. The use of a mass of markers across the whole genome has been explored, but to date has not yielded a noticeable improvement in discrimination [14,19].

Polygenic scores must be estimated from a finite training sample, and their effectiveness for association testing and risk prediction depends on the precision of this estimation as well as the proportion of variation explained by the polygenic score. The role of the sample size has not been thoroughly considered in this context. Several authors have expressed sensitivity and specificity in terms of the genetic variance of a predictor [17,20–22], but they did not distinguish the variance explained by an estimated predictor from that of the true predictor, that is the one that would be estimated from an infinitely large sample. While large samples lead to small sampling variance on individual marker effects, the errors accumulate across multiple markers such that the effect of sampling variation on the polygenic score can be considerable. Wray et al [2] used simulations to study the predictive accuracy of scores estimated from finite case/control studies, but did not obtain an explicit relation between sample size and accuracy. Similarly, the International Schizophrenia Consortium (ISC) [3] used simulations to show empirical relations between sample size and accuracy under several genetic models. Daetwyler et al [23] considered the effect of sampling variation on the correlation between polygenic score and total genetic value. Their results can be adapted to prediction of phenotypes rather than genetic values, and also to other measures of power and accuracy, but their conclusions are limited by an assumption that all the markers have effects and are included in the score.

In this work, statistical properties of polygenic score analyses are derived from a quantitative genetics model as a function of the explained genetic variance and sample sizes in discovery and replication samples. A range of options for constructing the score is considered, including estimation of the score from a different trait to the one predicted, selection of markers according to their *P*-values, and different methods for weighting markers in the score. The power is obtained for testing a polygenic score for association in a replication sample, and the correlation, mean square error, and area under the receiver-operator characteristic curve (AUC) are obtained for a predictor estimated from a finite training sample. These results are used to assess some recent studies and to discuss prospects for the future utility of polygenic score analyses for the prediction of complex traits.

Results

Analytic power and accuracy

In the framework considered here, a set of genetic markers is genotyped on an initial training sample and each marker is tested for association to a trait. Effect sizes are estimated for each marker and used to construct a polygenic score for each subject in an independent replication sample. The score is tested for association in the replication sample, in which the tested trait may differ from that in the training sample. The correlation and mean square error between the polygenic score and the tested trait are calculated. If the traits are binary, the AUC is obtained.

More precisely, consider a pair of traits $\mathbf{Y}=(Y_1,Y_2)'$ expressed as a linear combination of m genetic effects and an error term that includes environmental and unmodelled genetic effects:

$$\mathbf{Y}=\boldsymbol{\beta}'\mathbf{G}+\mathbf{E}=\left(\sum_{i=1}^{m}\beta_{i1}G_i+\mathrm{E}_1,\sum_{i=1}^{m}\beta_{i2}G_i+\mathrm{E}_2\right)' \qquad (1)$$

where $\boldsymbol{\beta}$ is a $m\times 2$ matrix of coefficients, $\mathbf{G}$ is a m-vector of coded genetic markers, and $\mathbf{E}$ is a pair of random errors that are independent of $\mathbf{G}$. Now suppose that the genetic effects on Y_1 are estimated from a sample of size n_1 and used to construct a polygenic score to be tested for association to Y_2 in an independent sample of size Y_1. Define the polygenic score to be

$$\hat{S}=\sum_{i=1}^{m}\hat{\beta}_{i1}G_i$$

Some important statistical properties of $\hat{S}$ can be expressed in terms of $\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})$ and $\mathrm{var}(\hat{\beta}_{i1})$, expressions for which are derived in the Methods. The coefficient of determination for the polygenic score on the second trait is

$$R^2_{\hat{S},Y_2}=\frac{m\,\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})^2}{\mathrm{var}(\hat{\beta}_{i1})\mathrm{var}(Y_2)} \qquad (2)$$

which is the squared correlation between the score and the trait. The prediction mean square error is

$$E\left[(\hat{S}-Y_2)^2\right]=m\,\mathrm{var}(\hat{\beta}_{i1})-2m\,\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})+E(Y_2^2) \qquad (3)$$

The asymptotic non-centrality parameter of the χ^2 test for association of $\hat{S}$ with Y_2 is

$$\lambda=\frac{n_2R^2_{\hat{S},Y_2}}{1-R^2_{\hat{S},Y_2}} \qquad (4)$$

on 1df, and the power of the two-tailed test of association at significance level α is

$$1-\Phi(\Phi^{-1}(1-\frac{\alpha}{2})-\sqrt{\lambda})+\Phi(\Phi^{-1}(\frac{\alpha}{2})-\sqrt{\lambda}) \quad (5)$$

Binary traits are assumed to arise from a liability threshold model [24] leading to calculation of the AUC also in terms of $\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})$ and $\mathrm{var}(\hat{\beta}_{i1})$, with the expressions given in the Methods. For binary traits the coefficient of determination in equation (2) may be transformed to the liability scale for more satisfactory interpretation [25], the details also given in the Methods.

The expressions for power and accuracy are derived in terms of the parameters listed in Table 1. Estimates of marker effects $\hat{\beta}_{i1}$ are either obtained from linear regression or set to a signed constant, which corresponds to the common approach of counting risk alleles across markers. A proportion of markers is assumed to have no effect, and markers may be selected by thresholding on their *P*-values.

Equation 4 suggests an estimating equation for any parameter of the quantitative model, given the association test between $\hat{\boldsymbol{S}}$ and Y_2. Write $R^2_{\hat{\boldsymbol{S}},Y_2}(\theta;\psi)$ explicitly as a function of some parameter θ in Table 1, treating all other parameters ψ as fixed and known. For example, θ might be the variance of marker effects in the training sample σ^2_1, from which $m\sigma^2_1$ is the explained genetic variance of the marker panel. Alternatively θ might be the covariance between marker effects in the two samples σ_{12}, assuming fixed values for the explained variances, and so on. Equation 4 is the squared coefficient of the linear regression of Y_2 on $\hat{\boldsymbol{S}}$, scaled by its sampling variance. The sampling distribution of that coefficient is normal, with mean the square root of equation 4. Therefore applying normal theory an estimator $\hat{\theta}$ is the solution to the equation

$$T^2=\frac{n_2R^2_{\hat{\boldsymbol{S}},Y_2}(\hat{\theta};\psi)}{1-R^2_{\hat{\boldsymbol{S}},Y_2}(\hat{\theta};\psi)} \quad (6)$$

where T^2 is the observed χ^2 association statistic. An approximate 95% confidence interval for θ is given by $(\hat{\theta}_L,\hat{\theta}_H)$ where $\hat{\theta}_L$ is the solution of

$$\left[\Phi^{-1}(.025;T)\right]^2=\frac{n_2R^2_{\hat{\boldsymbol{S}},Y_2}(\hat{\theta}_L;\psi)}{1-R^2_{\hat{\boldsymbol{S}},Y_2}(\hat{\theta}_L;\psi)}$$

and $\hat{\theta}_H$ is the solution of

$$\left[\Phi^{-1}(.975;T)\right]^2=\frac{n_2R^2_{\hat{\boldsymbol{S}},Y_2}(\hat{\theta}_H;\psi)}{1-R^2_{\hat{\boldsymbol{S}},Y_2}(\hat{\theta}_H;\psi)}$$

Published data: Association testing

The ISC was the first to demonstrate the utility of testing polygenic scores [3]. In their main result, odds ratios for 74062 nearly independent SNPs were estimated in 3322 cases and 3587 controls and used to construct a polygenic score that was tested in 2687 cases and 2656 controls of the Molecular Genetics of Schizophrenia study [26]. The score was more strongly associated as higher *P*-value thresholds were used for including SNPs, with the most significant reported association having $P=2\times10^{-28}$ with an inclusion threshold of $P<0.5$.

Assuming a prevalence of 1%, equation 5 gives a power of 80% at nominal significance if the explained genetic variance in liability is 7.2%, rising to 99% if the explained genetic variance is 11.7%. Assuming a heritability of 80% [3] this shows that the test was well powered if the marker panel explains about 10% of the heritability, which seems reasonable. The observed result of $P=2\times10^{-28}$ can be used in equation 6 to give an estimated explained genetic variance of 28.7% (95% CI: 23.6%–33.7%), which is 36% of the heritability, assuming that all SNPs have effects that are identical in the two samples. The estimate reduces only to 26.9% if 99% of the SNPs are assumed null. These results are similar to a recent estimate using mixed modelling of the same data [27].

In the ISC report, the *P*-value of the polygenic score decreases as the SNP inclusion threshold increases. This seems to suggest that a large number of associated markers lie within the mass of individually non-significant SNPs. In Figure 1 and Figure 2, the expected *P*-value of the polygenic score is shown as a function of

Table 1. Parameters and notation of polygenic model.

n_1	Training sample size
n_2	Replication sample size
m	Number of markers in genotyping panel
σ^2_1	Variance of marker effects in training sample
σ^2_2	Variance of marker effects in replication sample
σ_{12}	Covariance of marker effects between training and replication samples
π_0	Proportion of markers with no effect in either sample
p_0	Lower bound on *P*-value in the training sample for a marker to be included in polygenic score
p_1	Upper bound on *P*-value in the training sample for a marker to be included in polygenic score
K_1	Prevalence of binary trait in training sample
K_2	Prevalence of binary trait in replication sample
P_1	Sampling proportion of cases of binary trait in training sample
P_2	Sampling proportion of cases of binary trait in replication sample

doi:10.1371/journal.pgen.1003348.t001

the inclusion threshold, with the explained genetic variance set to its estimated value of 28.7% and other parameters as stated above. The figures show that this trend could be observed when as many as 90% of SNPs have no effects, and for the linear regression estimator the significance of the score continues to improve until the whole marker panel is included. Only for a very high proportion of null SNPs is there an optimal inclusion threshold less than 1. The allele count estimator has an optimum threshold less than 1 for all scenarios, but it is consistently less significant. Thus, in this dataset with high power, decreasing *P*-values are consistent with a range of polygenic models including those with a high proportion of null markers.

In this analysis the replication sample was smaller than the training sample, and we may ask what balance of sample sizes is optimal. Given the total sample size of 3322+2687 = 6009 cases and 3587+2656 = 6243 controls, the non-centrality parameter can be numerically maximised over the proportion of subjects allocated to the training sample. It is found that the optimal split is close to one-half regardless of the proportion of null SNPs or the *P*-value threshold, and the non-centrality parameter is roughly symmetrical around one-half. This suggests that given two samples of different size, it matters little which is chosen for training and which for testing. Furthermore, given an initial sample to be split into training and replication subsets, an obvious rule of thumb is to make an even split. Similar properties are seen under different genetic models (results not shown). Note that these results apply to association testing and not to individual prediction, which is discussed in the next subsection. For association testing there is a balance to be made between the precision of estimating the score in the training subset, and the power of testing the score in the replication subset. For prediction, however, the size of the replication subset does not affect the accuracy, only how precisely it is estimated; thus a larger training subset is more desirable in the prediction context.

The ISC further tested the schizophrenia-derived score against bipolar disorder, to test for a common genetic basis to those conditions. Their strongest result was with the Wellcome Trust Case-Control Consortium (WTCCC) sample of 1829 cases and 2935 controls, obtaining $P = 1 \times 10^{-12}$ with an inclusion threshold of $p_1 = 0.5$. Assume similar heritability for bipolar disorder as for schizophrenia [28] and the same genetic variance explained by the markers, estimated above to be 28.7%. Then using equation 5, the study had 80% power at nominal significance if the correlation is 28% between genetic effects on schizophrenia and bipolar disorder. Using equation 6, the estimated correlation given the observed association statistic is 70.6% (95%CI: 51.3%–89.7%)

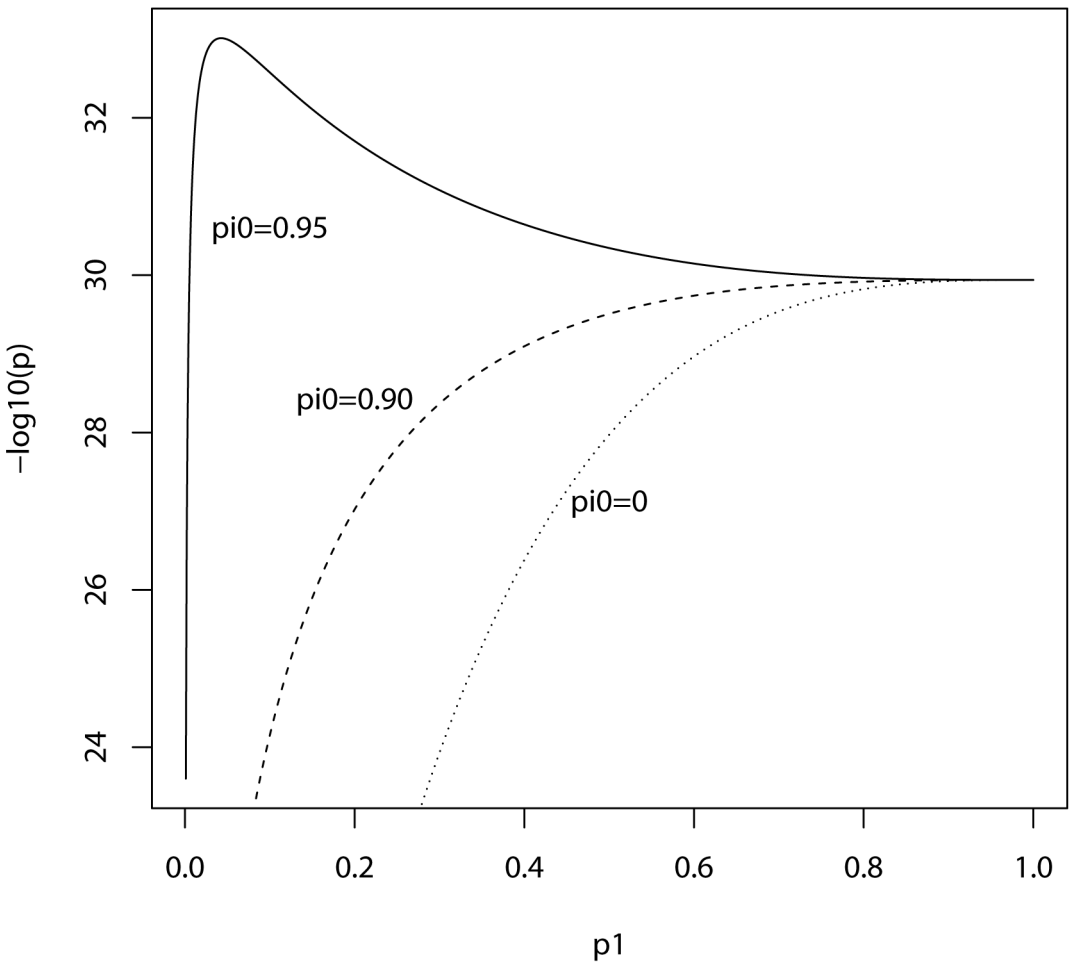

Figure 1. Expected $-\log_{10}(P)$ of linear regression estimate as a function of *P*-value threshold for selecting markers into the polygenic score. Training sample, 3322 cases and 3587 controls; replication sample, 2687 cases and 2656 controls. Marker panel of 74062 independent SNPs. Variance explained by markers, 28.7%. pi0, proportion of markers with no effect on disease.
doi:10.1371/journal.pgen.1003348.g001

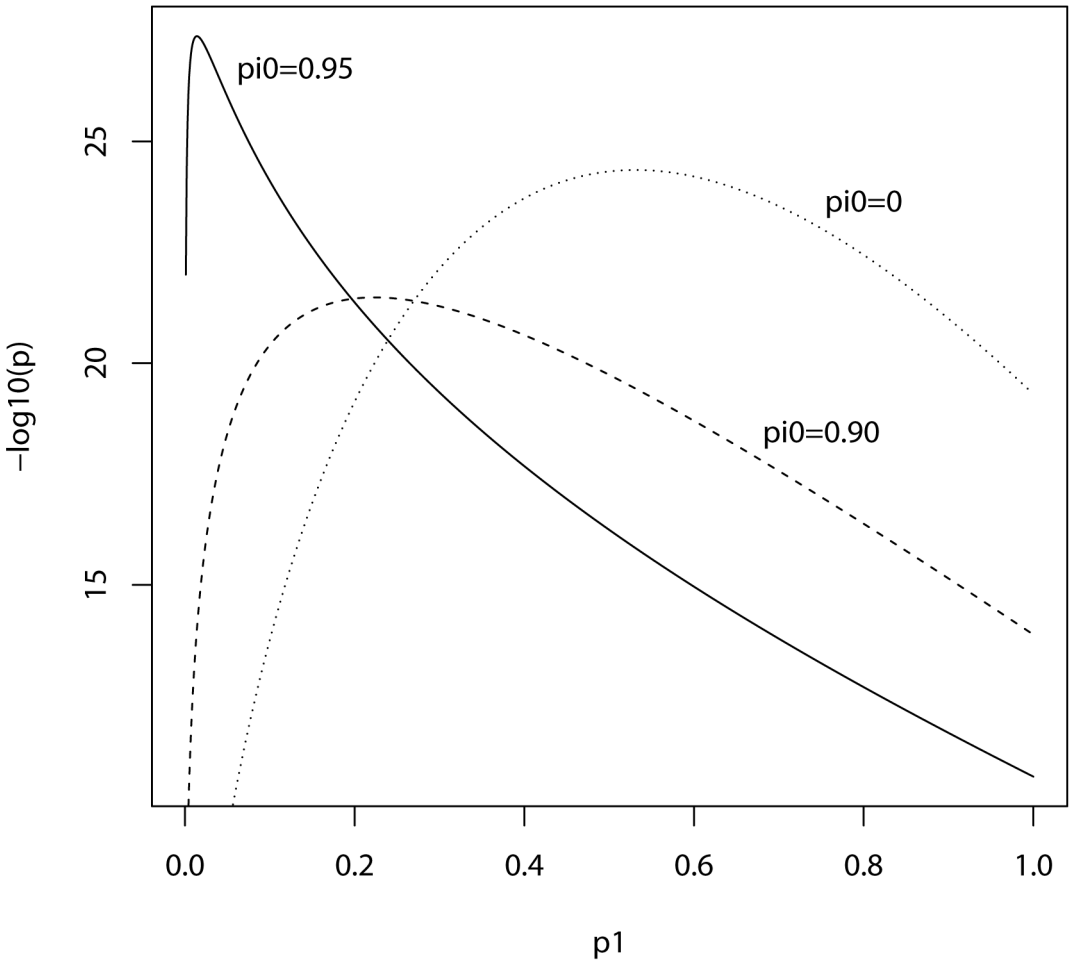

Figure 2. Expected $-\log_{10}(P)$ of allele score estimate as a function of *P*-value threshold for selecting markers into the polygenic score. Training sample, 3322 cases and 3587 controls; replication sample, 2687 cases and 2656 controls. Marker panel of 74062 independent SNPs. Variance explained by markers, 28.7%. pi0, proportion of markers with no effect on disease.
doi:10.1371/journal.pgen.1003348.g002

assuming that all SNPs have effects with explained variance 28.7% in both samples. If 99% of SNPs are assumed null, the estimated correlation reduces to 66.2% (95%CI: 48.1%–84.1%).

The International Multiple Sclerosis Consortium performed a similar exercise using a training sample of 931 cases and 2431 controls, a replication sample of 876 cases and 2077 controls, and a marker panel of 59470 nearly independent SNPs [6]. They also observed decreasing *P*-values for association as more SNPs were included in the score, obtaining $P = 6.12 \times 10^{-21}$ when all SNPs were included. Assuming prevalence of 0.1% this analysis has 80% power at nominal significance for explained genetic variance of 9.4%, and the observed result yields an estimate of 31.5% (95%CI: 24.9%–37.9%) assuming all SNPs have effects.

In applying these ideas to breast and prostate cancers, Machiela et al did not find significant associations of polygenic scores [14]. While this could be explained by the genetic architecture of the diseases, a possible explanation (noted by the authors) is the lower sample size together with the low heritability. Their breast cancer study used a total sample of 2287 subjects, approximately half of which were cases and half controls, which was split into training and testing subsets in a 9:1 ratio for 10-fold cross-validation. The marker panel consisted of 161,702 nearly independent SNPs. Assuming a prevalence of 3.6% and sibling relative risk of 2.5 [29], this design has only 17% power to detect an association of the polygenic score, even if the markers explain the full heritability. If the sample were split in a 1:1 ratio, the power would increase to 37%.

Their prostate cancer study had a total of 2277 subjects, approximately half of which were cases, again split in a 9:1 ratio and a marker panel of 165,508 nearly independent SNPs. Assuming a prevalence of 2.4% and sibling relative risk of 2.8 [29], this design has 19% power if the markers explain the full heritability. If the sample were split in a 1:1 ratio, the power would be 42%. It is clear that even with the optimistic assumption that the markers explain the full heritability, this study was unlikely to detect an association of the polygenic score for either cancer.

What sample size would have sufficient power to detect association of the polygenic score? For breast cancer the heritability of liability is estimated as 44% [21]. If the marker panel explains half of this heritability, roughly as in the ISC study, then two samples each of 1978 cases and 1978 controls would have 80% power at nominal significance. For prostate cancer the heritability of liability is also 44% and 1766 cases and controls would be required in each sample. For the ISC study, assuming explained genetic variance of 28.7%, 735 cases and controls in each sample are sufficient. Thus it appears that association testing is well powered at current sample sizes if two independent studies are used for training and testing, but less well powered if a single sample is split into two subsets.

As a final example of association testing, this time with a quantitative trait, Simonson et al studied the Framingham Risk Score for cardiovascular disease risk [8]. They also used 10-fold cross-validation of a single sample, giving training samples of 1575 subjects and testing samples of 175 subjects. They used a full set of 250,378 SNPs, which is here assumed to be similar to 100,000 independent SNPs. They first selected SNPs with P-values <0.1 into the score, then selected SNPs with $0.1<P<0.2$, then $0.2<P<0.3$ and so on, giving ten analyses. Even if the trait is fully heritable and explained by these markers, this analysis has 20% power for the SNPs with $P<0.1$, reducing with each P-value interval. For $0.4<P<0.5$, in which the authors found nominal significance of the score, the power is 6.7%. If all SNPs are included in the score, the power would be 38% if the trait is fully explained by the markers, but under a more conservative model in which the explained genetic variance is 30%, the power is just 8% and increases to 13% under an even split of training and testing samples. Again, splitting a single GWAS sample does not admit high power for testing a polygenic score.

Published data: Risk prediction

In their study of breast and prostate cancers Machiela et al also calculated the AUC for prediction of disease from the polygenic score. Here it is more important for the training sample to be large, ensuring accurate estimation of the score, justifying the 10-fold cross-validation design. Their AUC did not exceed 53% for breast cancer and 56.4% for prostate cancer. Under the same assumptions as above, the analytic AUC is 53.6% for breast cancer if the markers explain the full genetic variance, or 51.8% if they explain half. However if the sample were infinitely large, the AUCs would be 89% and 79% respectively. For prostate cancer, the analytic AUCs are 54.1% if the markers explain the full genetic variance, and 52% if they explain half; for a large sample they would be 90% and 80%. Thus, the low AUCs observed by Machiela et al are compatible with their study design, but they could be considerably higher if a larger training sample were available.

Evans et al considered prediction for the seven diseases of the WTCCC [19]. Approximately 2000 cases were available for each disease, with a common set of 1480 controls, and a marker panel of all SNPs on the Affymetrix 500K chip after quality control and exclusion of previously known loci. As this is the same chip used in the ISC study, it is assumed here that the panel is equivalent to 74062 independent SNPs. Logistic regression and allele score estimators were both used to construct scores, and a series of P-value thresholds from 10^{-5} to 0.8 were considered.

Table 2 compares the results of Evans et al to the analytic AUC for the diseases without strong MHC effects, using $P<0.8$ to select SNPs into the score, as that threshold generally gave the highest AUC. At that threshold, the choice of π_0 has little bearing on the results unless it is very close to 1, so it is set to 0. Also shown is the maximum AUC possible for each disease, obtained by letting the sample size grow to infinity. Bearing in mind that those authors noticed inflation in AUC for null SNPs, it is again clear that their modest results are compatible with the study design, and more encouraging results might be obtained from a larger sample. The calculations also confirm their observation that the allele count estimator is consistently less accurate than logistic regression; however while the two estimators give similar results at this sample size, more considerable differences emerge in the limit of large samples.

Several other studies have reported pseudo-R^2 from the regression of disease on the polygenic score [3,6,9]. Although prediction was not emphasised by those studies, they may still be evaluated for that purpose. Recently, Lee et al have argued that, for genetic predictors, R^2 on the liability scale is a more interpretable measure of accuracy for binary traits [25]. In Table 3, liability R^2 derived from those reports are compared to analytic values assuming different levels of heritability explained by the markers. The choice of π_0 has little bearing on these results so it is set to 0 throughout. The reported values are consistent with the markers explaining around half the heritability, with variation above and below. This is in line with the estimates of explained variance that were reported by those studies, and those estimates also agree well with those obtained using the method proposed here (equation 6). The low reported values of R^2 do not directly reflect the degree of missing heritability; rather they reflect the effect of sampling variation on the variance explained by an estimated score. Corresponding AUC values are also shown, and it is again clear that the currently modest utility of polygenic scores for discrimination is explained by limited training sample sizes, and much better results are possible through larger samples.

What sample size would permit estimation of a score with AUC at a clinical useful level, or otherwise close to its maximum value? The answer depends on π_0, the proportion of null markers in the panel, because if this is high then the individual marker effects will also be high and a low P-value threshold will eliminate much sampling error from the estimated score. Figure 3 shows AUC as a function of sample size for Crohn's disease, which has a high heritability of 76%, and breast cancer, which has low heritability of 44% [21], based on a panel of 100,000 independent markers. This is a similar number to current genotyping products, and results are given under a scenario in which the panel explains half the heritability [30]. For each sample size and π_0, the P-value threshold is applied that leads to the highest AUC. An AUC of 0.75 is generally regarded as the minimum useful level for screening subjects already considered at risk, whereas AUC of 0.99 is sufficient for screening the population at large [31]. For these two diseases the latter cannot be achieved from genetic data alone, so Table 4 gives minimum sample sizes for AUC of 0.75 and for 90%, 95% and 99% of the maximum possible AUC given the heritability.

The most favourable condition shown is $\pi_0 = 0.99$, that is there are 1000 markers with effects on disease. Figure 3 and Table 4 show that a few thousand cases and controls could yield a clinically useful AUC, but under most conditions several tens of thousands are needed. Under less favourable conditions – low heritability, low proportion of null markers – several hundred thousand cases and controls are needed to obtain an AUC within 10% of the achievable level, and even an AUC of 0.75 requires some tens of thousands of subjects. In the worst case the order of magnitude is of the millions.

Whole genome genotyping is now becoming feasible, under which the entire narrow-sense heritability would be represented. Assuming this is equivalent to about one million independent common SNPs [32], the required sample sizes are shown in Figure 4 and Table 5. Again, unless the heritability is explained by about 1000 markers, several tens to hundreds of thousands of subjects are needed to obtain a clinically useful AUC; for the genetic predictor to approach its potential, the order of magnitude is of the millions. The sample sizes to achieve AUC of 0.75 are larger than for 100,000 SNPs explaining half the heritability, but the latter scenario cannot achieve AUC of 0.99, so the clinical context can influence the choice of marker panel used to derive the predictor. It is clear that at current sample sizes, polygenic scores are only going to approach useful levels of discrimination if the marker panels include a high proportion of associated loci and the number of such loci is relatively small. Furthermore, for highly

Table 2. AUC calculated by Evans et al [19] compared to analytic values when ($\sigma^2_{g_1} = \frac{1}{2}h^2$) marker panel explains half the heritability, or ($\sigma^2_{g_1} = h^2$) marker panel explains the full heritability.

		Bipolar disorder	Coronary artery disease	Crohn's disease	Hypertension	Type-2 diabetes
	K	0.01	0.056	0.001	0.3	0.03
	h^2	0.69	0.72	0.76	1	0.6
Linear regression	Evans	0.668	0.595	0.614	0.61	0.601
	$\sigma^2_{g_1} = \frac{1}{2}h^2$	0.570 (0.890)	0.547 (0.843)	0.620 (0.948)	0.539 (0.841)	0.545 (0.832)
	$\sigma^2_{g_1} = h^2$	0.638 (0.974)	0.592 (0.948)	0.727 (0.995)	0.577 (0.971)	0.588 (0.934)
Allele count	Evans	0.653	0.599	0.617	0.602	0.589
	$\sigma^2_{g_1} = \frac{1}{2}h^2$	0.561 (0.827)	0.540 (0.780)	0.604 (0.894)	0.533 (0.770)	0.539 (0.772)
	$\sigma^2_{g_1} = h^2$	0.620 (0.922)	0.580 (0.880)	0.698 (0.970)	0.567 (0.885)	0.576 (0.868)

K, population prevalence and h^2, heritability of liability taken from Wray et al [21] except for hypertension which is assumed fully heritable for illustration. In parentheses, AUC achieved by an infinite sample.

doi:10.1371/journal.pgen.1003348.t002

polygenic conditions the sample sizes needed to approach this potential are an order of magnitude higher than are currently available.

Finally, Table 6 gives similar calculations for the correlation between predicted and observed quantitative traits with high ($h^2 = 0.8$) and moderate ($h^2 = 0.4$) heritability. The prospects here appear more challenging in terms of the sample sizes needed to approach the achievable correlation. For example, height has heritability of about 0.8, and the number of associated variants is known to be at least in the hundreds [7]. In the most optimistic scenario shown, 31,000 subjects would be required to derive a predictor with correlation 0.8 with the true height. In fact these sample sizes are now being approached by collaborative studies, and this result confirms that this is necessary for accurate prediction of quantitative traits in addition to the primary goal of identifying individually associated markers.

Discussion

To date polygenic score analyses have been performed opportunistically. The results provided here allow a more informed appraisal of these analyses, characterisation of the statistical properties of the methods, and insights into the future prospects of polygenic modelling. R code to compute the formulae in this paper is available from the author (sites.google.com/site/fdudbridge/software/).

Table 3. R^2 reported for complex diseases compared to analytic values when ($\sigma^2_{g_1} = \frac{1}{4}h^2, \frac{1}{2}h^2, h^2$) marker panel explains one quarter, one half or the full heritability.

		Schiz [3]	MS [6]	BrCa [14]	PrCa [14]	RA [9]	Celiac [9]	MI/CAD [9]	T2D [9]
	K	.01	.001	.036	.024	.0075	.0075	.056	.03
	h^2	.8	.5	.44	.44	.55	.55	.72	.6
	π_0	0	0	0	0	.97	.98	.98	.96
	Reported R^2	.013	.012	.001	.001	.003	.007	.007	.013
$\sigma^2_{g_1} = \frac{1}{4}h^2$	R^2	.006 (.2)	.002 (.125)	.0002 (.11)	.0002 (.11)	.001 (.1375)	.0008 (.1375)	.001 (.18)	.003 (.15)
	AUC	.56 (.81)	.54 (.81)	.51 (.71)	.51 (.72)	.52 (.52)	.52 (.77)	.52 (.75)	.53 (.75)
$\sigma^2_{g_1} = \frac{1}{2}h^2$	R^2	.024 (.4)	.008 (.25)	.0008 (.22)	.0009 (.22)	.006 (.275)	.003 (.275)	.004 (.36)	.010 (.3)
	AUC	.62 (.91)	.58 (.90)	.52 (.79)	.52 (.80)	.56 (.87)	.55 (.87)	.54 (.84)	.57 (.94)
$\sigma^2_{g_1} = h^2$	R^2	.089 (.8)	.03 (.5)	.003 (.88)	.003 (.44)	.025 (.55)	.013 (.55)	.017 (.72)	.013 (.6)
	AUC	.72 (.99)	.66 (.97)	.54 (.89)	.54 (.90)	.62 (.95)	.59 (.96)	.58 (.95)	.63 (.94)
	Reported $\sigma^2_{g_1}$	.3	na	na	na	.18	.44	.48	.49
	$\hat{\sigma}^2_{g_1}$	.29	.31	.30	.28	.21	.40	.47	.34

Schiz, schizophrenia. MS, multiple sclerosis. BrCa, breast cancer. PrCa, prostate cancer. RA, rheumatoid arthritis. Celiac, celiac disease. MI/CAD, early-onset myocardial infarction or coronary artery disease. T2D, type-2 diabetes. *K*, population prevalence and h^2, heritability of liability taken from Visscher et al [1] and Wray et al [21] except for celiac, assumed equal to RA. π_0, proportion of markers assumed to have no effects. Reported R^2, highest R^2 reported in cited publication, transformed to the liability scale. In parentheses, values achieved by an infinite training sample. Reported $\sigma^2_{g_1}$, variance explained by markers as estimated in cited publication. $\hat{\sigma}^2_{g_1}$, estimated variance explained using method proposed herein.

doi:10.1371/journal.pgen.1003348.t003

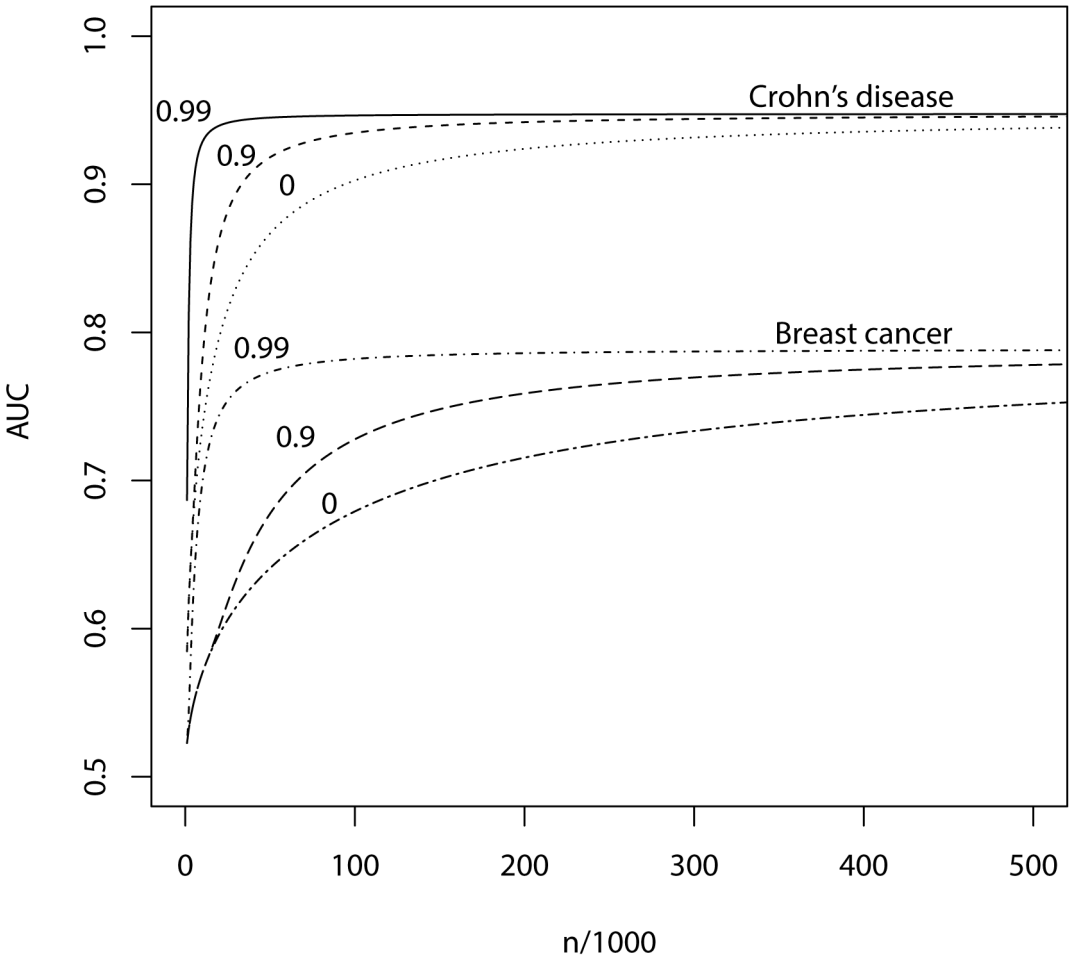

Figure 3. AUC as a function of sample size, using a panel of 100,000 markers that explains half the heritability of liability. *n*, number of cases and of controls in training sample. Heritability of liability, 76% for Crohn's disease. 44% for breast cancer. Line annotations are the proportion of markers with no effect on disease.
doi:10.1371/journal.pgen.1003348.g003

Current sample sizes are clearly adequate for testing association of a polygenic score in a replication sample, as long as full size samples are used for both training and testing. This is already apparent from the extraordinary significance levels reported in the seminal studies [3,6], but here it is shown that those results are compatible with realistic genetic models and are not necessarily explained by analytic biases that accumulate across SNPs. This had been previously shown by the ISC study, which simulated

Table 4. Numbers of cases and controls (in 1000s of each, rounded up) required to attain a specified AUC using a panel of 100,000 markers that explains half the heritability of liability.

	AUC	$\pi_0 = 0.99$	$\pi_0 = 0.90$	$\pi_0 = 0.75$	$\pi_0 = 0$
Crohn's disease ($h^2 = 0.76$, $K = 0.001$, Max = 0.95)	0.75	2 (0.0004)	9 (0.02)	12 (0.5)	12 (1)
	0.855 = 0.9*Max	3 (0.0004)	19 (0.01)	34 (0.06)	42 (1)
	0.9025 = 0.95*Max	6 (0.0004)	35 (0.008)	68 (0.04)	100 (1)
	0.9405 = 0.99*Max	23 (0.0003)	165 (0.004)	349 (0.02)	690 (1)
Breast cancer ($h^2 = 0.44$, $K = 0.036$, Max = 0.79)	0.75	23 (0.0004)	157 (0.008)	311 (0.03)	476 (1)
	0.711 = 0.9*Max	12 (0.0005)	77 (0.01)	144 (0.05)	183 (1)
	0.7125 = 0.95*Max	23 (0.0005)	159 (0.01)	315 (0.05)	484 (1)
	0.7821 = 0.99*Max	100 (0.00024)	755 (0.00389)	1610 (0.0147)	3281 (1)

π_0, proportion of SNPs having no effect on disease. Max, maximum AUC achievable given the genetic variance of the marker panel. In parentheses, *P*-value threshold that maximises the AUC.
doi:10.1371/journal.pgen.1003348.t004

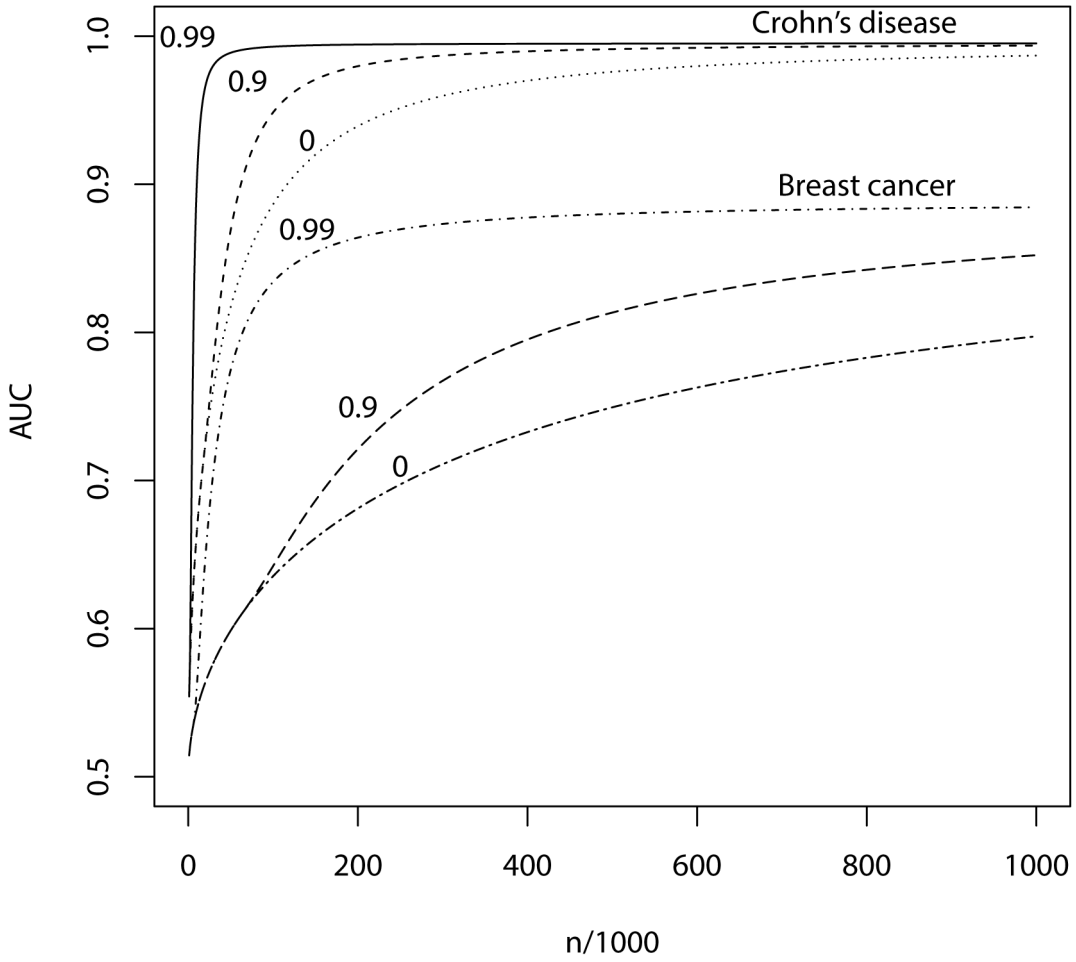

Figure 4. AUC as a function of sample size, using a panel of 1,000,000 markers that explains the full heritability. *n*, number of cases and of controls in training sample. Heritability of liability, 76% for Crohn's disease. 44% for breast cancer. Line annotations are the proportion of markers with no effect on disease.
doi:10.1371/journal.pgen.1003348.g004

plausible genetic models and showed that they led to similar results to those observed in the data [3]; here, the result is shown directly for the common quantitative model, without recourse to simulations. Studies that split a single sample into cross-validation subsets have been less successful [8,14], but here it is shown that this could be explained by their limited sample sizes, and more encouraging results for the same traits could be obtained with modestly increased samples.

Table 5. Numbers of cases and controls (in 1000s of each, rounded up) required to attain a specified AUC using a panel of 1,000,000 markers that explains the full heritability.

	AUC	$\pi_0 = 0.999$	$\pi_0 = 0.99$	$\pi_0 = 0.90$	$\pi_0 = 0.75$	$\pi_0 = 0$
Crohn's disease ($h^2 = 0.76$, $K = 0.001$, Max = 1.00)	0.75	1 (0.00007)	5 (0.0004)	25 (0.08)	27 (1)	27 (1)
	0.9 = 0.9*Max	2 (0.00007)	10 (0.0004)	62 (0.01)	107 (0.1)	117 (1)
	0.95 = 0.95*Max	3 (0.00007)	16 (0.0005)	103 (0.01)	190 (0.05)	243 (1)
	0.99 = 0.99*Max	8 (0.00007)	58 (0.0003)	413 (0.006)	847 (0.02)	1487 (1)
Breast cancer ($h^2 = 0.44$, $K = 0.036$, Max = 0.89)	0.75	6 (0.00007)	41 (0.0004)	256 (0.01)	448 (0.09)	505 (1)
	0.801 = 0.9*Max	9 (0.00007)	65 (0.0005)	428 (0.009)	806 (0.05)	1062 (1)
	0.8455 = 0.95*Max	17 (0.00007)	124 (0.0004)	857 (0.007)	1702 (0.03)	2656 (1)
	0.8811 = 0.99*Max	77 (0.00007)	566 (0.0002)	4305 (0.004)	9223 (0.01)	19191 (1)

π_0, proportion of SNPs having no effect on disease. Max, maximum AUC achievable given the genetic variance of the marker panel. In parentheses, *P*-value threshold that maximises the AUC.
doi:10.1371/journal.pgen.1003348.t005

Table 6. Numbers of subjects (in 1000s, rounded up) required to attain a specified correlation with a normal trait using a panel of 1,000,000 markers that explains the full heritability.

	Correlation	$\pi_0 = 0.999$	$\pi_0 = 0.99$	$\pi_0 = 0.90$	$\pi_0 = 0.75$	$\pi_0 = 0$
$h^2 = 0.8$ (Max = 0.894)	0.8046 = 0.9*Max	31 (0.00007)	227 (0.0004)	1601 (0.007)	3231 (0.03)	5329 (1)
	0.8493 = 0.95*Max	55 (0.00007)	411 (0.0003)	3004 (0.005)	6250 (0.02)	11571 (1)
	0.88506 = 0.99*Max	213 (0.00007)	1546 (0.0002)	12171 (0.003)	26724 (0.01)	61565 (1)
$h^2 = 0.4$ (Max = 0.632)	0.5688 = 0.9*Max	61 (0.00007)	453 (0.0004)	3201 (0.007)	6461 (0.03)	10658 (1)
	0.6004 = 0.95*Max	109 (0.00007)	821 (0.0003)	6007 (0.005)	12500 (0.02)	23141 (1)
	0.62568 = 0.99*Max	426 (0.00007)	3092 (0.0002)	24341 (0.003)	53448 (0.01)	123128 (1)

π_0, proportion of SNPs having no effect on the trait. Max, maximum correlation achievable given the genetic variance of the marker panel. In parentheses, *P*-value threshold that maximises the correlation.

doi:10.1371/journal.pgen.1003348.t006

When a sample is to be split into two subsets, a roughly even split yields the greatest power for testing association of the score. However, for predicting individual trait values it is more important for the training set to be large, and standard procedures such as 10-fold cross-validation remain preferable. If both testing and prediction are intended from a single sample, a pragmatic approach is to ensure adequate power by allocating about 2000 cases and 2000 controls to the replication sample, providing this is less than half the total, and then to ensure high predictive accuracy by allocating the remainder to the training sample.

The outlook for disease and trait prediction is more challenging. To date the severe shortfall in the accuracy of genetic predictors has generally been ascribed to incomplete coverage of marker panels or failure to identify sufficiently many associated markers. Here, however, no criteria for declaring individual significance are imposed, but neither does the calculation force the predictor to include markers that contribute no information. Under this pragmatic approach it results that tens of thousands of subjects, at least, are needed to derive predictors that are clinically useful. Furthermore, previous results on the potential accuracy of genetic prediction [17,20–22] only become relevant at very large sample sizes. Such numbers are now coming within reach of national biobank projects and international consortia, so the emergence of useful genetic predictors may not be too far off, although such large samples create issues of effect heterogeneity that are not addressed here. Recent estimates of the proportion of markers having effects also suggest that the more optimistic scenarios shown in Table 4, Table 5, Table 6 may apply [9,33]. Although the focus here is on AUC, various other measures of predictive accuracy are possible and can be computed within the same framework [24,25]. The expressions given here could be adapted to other measures without much difficulty.

For some diseases, fairly high AUC has already been observed [17,19]. This does not conflict with the present work but reflects the presence of major gene effects, usually in the MHC, which depart from the quantitative model treated here. Similarly, some diseases have non-genetic risk factors that already admit clinically useful predictors. There the more relevant issue is the extent to which genetics improves established models [34]. Again the focus has tended to lie on identifying specific markers to improve prediction, rather than the sample size needed to accurately estimate their combined effects. The approach taken here could easily be extended to accommodate additional fixed effects.

A fairly general construction of the polygenic score has been described, including weighted and unweighted methods from single marker analysis, and shrinkage methods used in multivariate analysis. There is little to choose between these estimators in terms of power, correlation or AUC, but the unweighted estimator will perform relatively worse as sample size increases since its sampling error does not reduce to zero. Shrinkage estimation leads to reduced mean square error for prediction and has some other advantages [35,36], but in the main applications for polygenic scores to date, namely association testing and AUC, it does not improve over the linear regression estimate.

However, some ideal conditions have been assumed including independence of markers and of study subjects. In reality markers will be in linkage disequilibrium and the approximation by an effective number of independent tests is heuristic. Similarly, subjects will be related, if distantly. Results from real data may depart from those presented here if proper account is taken of relationships between subjects. In particular, shrinkage estimation is likely to improve power and correlation, as well as mean square error, by analysing all markers simultaneously rather than each one marginally [37].

The assumption that effects are normally distributed is necessary when markers are selected by their *P*-values but not otherwise. Similarly, allowing a proportion of markers to have no effect only makes a difference when selecting markers by *P*-values. Thus the present results are relevant even if one does not entirely accept the polygenic model proposed. The normal distribution simplifies some calculations, but various heavy-tailed distributions have also been proposed for GWAS data [38,39] and would lead to improved prediction if such models held in truth. Furthermore the assumption of normality applies to effects on the standardised genotype scale, but there are plausible models for effect sizes as a function of allele frequency, leading to non-normal effects on the standardised scale. This may particularly affect the results for shrinkage estimation when the degree of shrinkage varies for markers with different allele frequencies. The numerical results presented are therefore not definitive but should be taken a guide to the likely magnitude of results in specific applications.

A novel approach to estimating parameters of the polygenic model has been proposed, showing promise for inferring the explained genetic variance and/or proportion of null markers. The method yields estimates that are similar to those obtained by existing approaches [1,37]. A similar approach to estimation has been developed by Stahl et al [9], based on simulating GWAS data from proposed models, and using rejection sampling to construct posterior distributions of their parameters. Apart from the accommodation of prior distributions (which were uninformative), this is essentially the same approach as used here except that whole genome simulation is used to obtain a sampling distribution. The analytic results provided here should allow this approach to be implemented more efficiently, and this will be attempted in future work.

The *P*-value thresholds that maximise the power and AUC are more permissive than the usual thresholds for individual markers. This means that polygenic analyses can be powerful while still including many non-significant markers, so they will continue to be useful as long as individually associated markers remain to be discovered. Although larger samples are needed for useful risk prediction, polygenic scores have an ongoing current role in assessing the variance explained by marker panels and the genetic correlation between related traits and populations.

Methods

Quantitative model

Recall equation 1 in which a pair of traits $\mathbf{Y}=(Y_1,Y_2)'$ is expressed as a linear combination of m genetic effects and an error term that includes environmental and unmodelled genetic effects:

$$\mathbf{Y}=\boldsymbol{\beta}'\mathbf{G}+\mathbf{E}=\left(\sum_{i=1}^{m}\beta_{i1}G_i+\mathrm{E}_1,\sum_{i=1}^{m}\beta_{i2}G_i+\mathrm{E}_2\right)'$$

where $\boldsymbol{\beta}$ is a $m\times 2$ matrix of coefficients, $\mathbf{G}$ is a m-vector of coded genetic markers, and $\mathbf{E}$ is a pair of random errors that are independent of $\mathbf{G}$. Assume that the markers G_i are independent and standardised. In the usual case of single nucleotide polymorphism (SNP) genotypes under Hardy-Weinberg Equilibrium, $G_i=(X_i-2f_i)/(2f_i(1-f_i))^{\frac{1}{2}}$ where X_i is the number of minor alleles and f_i is the minor allele frequency at SNP i. The genetic effects β_{ij} are regarded as fixed across samples but random over $i=1,\ldots,m$ with $E(\beta_{ij})=0$, $\mathrm{var}(\beta_{ij})=\sigma_j^2$ and $\mathrm{cov}(\beta_{i1},\beta_{i2})=\sigma_{12}$. Then the variance-covariance matrix of $\mathbf{Y}$ is written as

$$\Sigma_{\mathbf{Y}}=\begin{bmatrix} m\sigma_1^2+\sigma_{e_1}^2 & m\sigma_{12} \\ m\sigma_{12} & m\sigma_2^2+\sigma_{e_2}^2 \end{bmatrix}=\begin{bmatrix} \sigma_{g_1}^2+\sigma_{e_1}^2 & \rho\sigma_{g_1}\sigma_{g_2} \\ \rho\sigma_{g_1}\sigma_{g_2} & \sigma_{g_2}^2+\sigma_{e_2}^2 \end{bmatrix}$$

For continuous traits, assume without loss of generality that Y_1 and Y_2 are standardised so that $\sigma_{g_1}^2$ and $\sigma_{g_2}^2$ are the proportions of variation of each trait explained by $\mathbf{G}$. These quantities will be called the explained genetic variances, and are bounded above by the heritabilities.

The genetic effects on Y_1 are estimated from a sample of size n_1 and used to construct a polygenic score to be tested for association to Y_2 in an independent sample of size n_2. Define the polygenic score to be

$$\hat{S}=\sum_{i=1}^{m}\hat{\beta}_{i1}G_i$$

Clearly $E(\hat{S})=0$. Furthermore if $E(\hat{\beta}_{i1})=0$ then

$$\mathrm{var}(\hat{S})=\sum_{i=1}^{m}\mathrm{var}(\hat{\beta}_{i1}G_i)=\sum_{i=1}^{m}\hat{\beta}_{i1}^2\approx m\,\mathrm{var}(\hat{\beta}_{i1})$$

$$\begin{aligned}\mathrm{cov}(\hat{S},Y_2)&=E\left[\sum_{i=1}^{m}\hat{\beta}_{i1}G_i\sum_{i=1}^{m}\beta_{i2}G_i\right]=E\left[\sum_{i=1}^{m}\hat{\beta}_{i1}\beta_{i2}G_i^2\right]\\&=\sum_{i=1}^{m}\hat{\beta}_{i1}\beta_{i2}\approx m\,\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})\end{aligned}\qquad(7)$$

These expressions are equalities in the limit of large m but are approximations for a finite number of markers because the true effects β_{ij} are a sample from their random effects distribution.

Equations 2–6 in Results follow immediately, in which the key quantities are $\mathrm{cov}(\hat{\beta}_{i1},\beta_2)$ and $\mathrm{var}(\hat{\beta}_{i1})$. They in turn depend upon the form of the estimator $\hat{\beta}_{i1}$, for which three alternatives are now discussed.

Linear regression

A natural estimate of β_{i1} is the least squares estimate from the univariate linear regression of Y_1 on G_i. Then $\hat{\beta}_{i1}$ is asymptotically normally distributed with sampling mean β_{i1} and variance $\mathrm{var}(\hat{\beta}_{i1}-\beta_{i1})=(1-\sigma_1^2)n_1^{-1}$ since G_i is standardised by definition. Assuming that genetic effects are small, it is henceforth conservatively taken that $\mathrm{var}(\hat{\beta}_{i1}-\beta_{i1})\approx n_1^{-1}$ as previously suggested by Daetwyler et al [23]. The total variance of this estimator over markers and samples is $\mathrm{var}(\hat{\beta}_{i1})=\sigma_1^2+n_1^{-1}$, and its correlation with the effects on Y_2 is

$$\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})=\mathrm{cov}(\beta_{i1}+\varepsilon_{i1},\beta_{i2})=\mathrm{cov}(\beta_{i1},\beta_{i2})=\sigma_{12}$$

where ε_{i1} are the sampling errors. Immediate power and accuracy calculations are then available by substituting $\mathrm{var}(\hat{\beta}_{i1})=\sigma_1^2+n_1^{-1}$ and $\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})=\sigma_{12}$ into equations 2–5. When $\mathrm{cov}(\beta_{i1},\beta_{i2})=\sigma_1^2$, as when the same trait is considered in both samples, equation 2 gives the formula previously derived by Daetwyler et al [23], modified to allow for prediction of the phenotype rather than the genetic value. In the present notation,

$$R^2_{\hat{S},Y_2}=\frac{m^2\sigma_1^4}{(\sigma_1^2+n_1^{-1})\mathrm{var}(Y_2)}=\sigma_{g_1}^2\,\frac{\frac{n_1}{m}\sigma_{g_1}^2}{(\frac{n_1}{m}\sigma_{g_1}^2+1)\mathrm{var}(Y_2)}$$

corresponds to equation 1 of those authors, with the additional factor $\sigma_{g_1}^2$ being the genetic variance of the phenotype. This shows that the key determinants of the predictive accuracy are the variance explained by the markers and the ratio of the sample size to the number of markers.

Now suppose markers are only selected into the polygenic score if they have two-tailed *P*-values between thresholds p_0,p_1 where $0\le p_0\le p_1\le 1$. Asymptotically the equivalent constraint for $\hat{\beta}_{i1}$ is obtained from the Wald statistic as

$$n^{-\frac{1}{2}}q_1\le\left|\hat{\beta}_{1j}\right|\le n^{-\frac{1}{2}}q_0\qquad(8)$$

where $q_0=\Phi^{-1}(1-\frac{1}{2}p_0)$, $q_1=\Phi^{-1}(1-\frac{1}{2}p_1)$.

Suppose further that a proportion π_0 of the m markers have no effect on Y_1 (*ie.* $\beta_{i1}=0$), and the remaining markers have effects drawn from $N(0,(1-\pi_0)^{-1}\sigma_1^2)$. Then among the null markers the variance of $\hat{\beta}_{i1}$, conditional on selection into the polygenic score, is obtained from properties of the truncated normal distribution as [40]

$$n_1^{-1}\left[1+\frac{q_1\phi(q_1)-q_0\phi(q_0)}{\Phi(q_0)-\Phi(q_1)}\right]$$

Similarly, among the non-null markers the variance of $\hat{\beta}_{1j}$, conditional on selection into the polygenic score, is

$$((1-\pi_0)^{-1}\sigma_1^2+n_1^{-1})\left[1+\frac{r_1\phi(r_1)-r_0\phi(r_0)}{\Phi(r_0)-\Phi(r_1)}\right] \quad (9)$$

where $r_0=q_0(n_1(1-\pi_0)^{-1}\sigma_1^2+1)^{-\frac{1}{2}}$, $r_1=q_1(n_1(1-\pi_0)^{-1}\sigma_1^2+1)^{-\frac{1}{2}}$. The probability that a null marker is selected into the polygenic score is $2\pi_0(\Phi(q_0)-\Phi(q_1))$ and the corresponding probability for a non-null marker is $2(1-\pi_0)(\Phi(r_0)-\Phi(r_1))$. Therefore the total variance of $\hat{\beta}_{i1}$ is

$$\mathrm{var}(\hat{\beta}_{i1})=2\pi_0 n_1^{-1}[\Phi(q_0)-\Phi(q_1)+q_1\phi(q_1)-q_0\phi(q_0)]+$$
$$2(1-\pi_0)((1-\pi_0)^{-1}\sigma_1^2+n_1^{-1})[\Phi(r_0)-\Phi(r_1)+r_1\phi(r_1)-r_0\phi(r_0)] \quad (10)$$

Note that when $p_0=0$ and $p_1=1$, that is all markers are included in the score, then equation 10 reverts to $\mathrm{var}(\hat{\beta}_{i1})=\sigma_1^2+n_1^{-1}$, which is invariant to the proportion of null markers π_0 and does not assume a normal distribution for the non-null effects [23].

To obtain $\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})$ allowing for selection of markers, note that the regression of β_{i2} on $\hat{\beta}_{i1}$ has the same coefficient regardless of selection on $\hat{\beta}_{i1}$. For non-null markers this coefficient is

$$\frac{(1-\pi_0)^{-1}\sigma_{12}}{(1-\pi_0)^{-1}\sigma_1^2+n_1^{-1}}$$

and the covariance is this coefficient times the conditional variance of $\hat{\beta}_{i1}$ given in equation 9. For null markers the covariance is zero, so the total covariance is

$$\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})=\frac{(1-\pi_0)^{-1}\sigma_{12}}{(1-\pi_0)^{-1}\sigma_1^2+n_1^{-1}}\cdot$$
$$2(1-\pi_0)((1-\pi_0)^{-1}\sigma_1^2+n_1^{-1})[\Phi(r_0)-\Phi(r_1)+r_1\phi(r_1)-r_0\phi(r_0)]$$
$$=2\sigma_{12}[\Phi(r_0)-\Phi(r_1)+r_1\phi(r_1)-r_0\phi(r_0)]$$

This expression is substituted into equations 2–5 together with the variance in equation 10 to obtain the power and accuracy of the polygenic score when markers are selected into the score based on their *P*-values.

Shrinkage estimation

It is well known that estimation and prediction for multivariate models can be improved, in terms of mean squared error, by assuming that their effects come from a common underlying distribution. A common approach in quantitative genetics is to fit a mixed model in which genetic markers have random effects for which best linear unbiased predictors (BLUPs) are obtained [35]. This is one of several closely related formulations of multilevel models [36]. As these approaches tend to give similar results when the number of markers is large, a basic Bayesian estimation scheme is outlined here and will be assumed to give typical results for a shrinkage estimator.

Suppose β_{i1} has the prior distribution $\beta_{i1}\sim N(0,\sigma_1^2)$, and let the "data" consist of the univariate linear regression estimates, $\hat{\beta}_{i1}|\beta_{i1}\sim N(\beta_{i1},n_1^{-1})$. Then the posterior for β_{i1} given $\hat{\beta}_{i1}$ is also normal, $\beta_{i1}|\hat{\beta}_{i1}\sim N(A\hat{\beta}_{i1},An_1^{-1})$ where $A=\frac{\sigma_1^2}{\sigma_1^2+n_1^{-1}}$ [41]. A natural estimator for β_{i1} is therefore the posterior mean $\tilde{\beta}_{i1}=A\hat{\beta}_{i1}$ for which $\mathrm{var}(\tilde{\beta}_{i1})=A^2\,\mathrm{var}(\hat{\beta}_{i1})$ and $\mathrm{cov}(\tilde{\beta}_{i1},\beta_{i2})=A\,\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})$. Since all effects are shrunk by the same factor A it follows that this approach leads to the same power and correlation as the linear regression estimator $\hat{\beta}_{i1}$, but the mean square error is reduced to $mA^2\,\mathrm{var}(\hat{\beta}_{i1})-2mA\,\mathrm{cov}(\hat{\beta}_{i1},\beta_{i2})+1$.

Allele count

A currently common approach is to construct the polygenic score by summing the number of trait-increasing alleles across selected markers, without considering their effect sizes other than to identify the direction of association at each marker. This may be called an unweighted score, in contrast to the above approaches that estimate weights for each marker. The unweighted score may be more robust against errors in estimating the effect sizes arising from limited sample size, population heterogeneity, "winner's curse" bias, and confounding by population structure. Here a related approach is considered in which all markers are given the same absolute effect size on the standardised genotype scale. This is equivalent to the allele counting approach when all markers have the same allele frequency. When allele frequencies are heterogeneous, allele counting assumes that all markers have the same effect on the trait, whereas the present approach assumes that all markers contribute the same proportion of variance to the trait. Both models can be criticised but the present approach will allow the comparison of weighted to unweighted scores without considering the distribution of allele frequencies or their relation to the effect sizes.

The polygenic score is now calculated as

$$\hat{S}=\sum_{i=1}^{m}\mathrm{sgn}(\hat{\beta}_{i1})G_i$$

where $\mathrm{sgn}(x)=\frac{|x|}{x}$ and $\hat{\beta}_{i1}$ is the linear regression estimate as before. Clearly $\mathrm{var}(\mathrm{sgn}(\hat{\beta}_{i1}))=1$. The covariance $\mathrm{cov}(\mathrm{sgn}(\hat{\beta}_{i1}),\beta_{i2})$ is obtained by integrating over the distribution of β_{i1}. Allow again for selection of markers by their *P*-values as in equation 8 and denote the selection event by $\zeta: n^{-\frac{1}{2}}q_1\le\left|\hat{\beta}_{1j}\right|\le n^{-\frac{1}{2}}q_0$. Then using the symmetry of the distribution of β_{i1} the required covariance is

$$\mathrm{cov}(\mathrm{sgn}(\hat{\beta}_{i1}),\beta_{i2})=E(\mathrm{sgn}(\hat{\beta}_{i1})\beta_{i2})$$
$$=2\int_{x=0}^{\infty}\Pr(\beta_{i1}=x|\zeta)$$
$$\left[\Pr(\hat{\beta}_{i1}>0|\beta_{i1}=x,\zeta)-\Pr(\hat{\beta}_{i1}<0|\beta_{i1}=x,\zeta)\right]E(\beta_{i2}|\beta_{i1}=x)dx$$
$$=2\int_{x=0}^{\infty}\frac{\Pr(\beta_{i1}=x)\Pr(\zeta|\beta_{i1}=x)}{\Pr(\zeta)}\frac{(\Pr(\hat{\beta}_{i1}>0,\zeta|\beta_{i1}=x)-\Pr(\hat{\beta}_{i1}<0,\zeta|\beta_{i1}=x))}{\Pr(\zeta|\beta_{i1}=x)}$$
$$E(\beta_{i2}|\beta_{i1}=x)dx$$
$$=\frac{2}{\Pr(\zeta)}\int_{x=0}^{\infty}\Pr(\beta_{i1}=x)\left[\Pr(\hat{\beta}_{i1}>0,\zeta|\beta_{i1}=x)-\Pr(\hat{\beta}_{i1}<0,\zeta|\beta_{i1}=x)\right]$$
$$E(\beta_{i2}|\beta_{i1}=x)dx$$

The probabilities in this expression are as follows. The selection probability is again

$$\Pr(\zeta) = 2(1-\pi_0)(\Phi(r_0)-\Phi(r_1)) + 2\pi_0(\Phi(q_0)-\Phi(q_1)).$$

The probability density for nonzero β_{i1} is

$$\begin{aligned}\Pr(\beta_{i1}=x, \beta_{i1}\neq 0) &= \Pr(\beta_{i1}\neq 0)\Pr(\beta_{i1}=x|\beta_{i1}\neq 0)\\ &= (1-\pi_0)^{\frac{3}{2}}\sigma_1^{-1}\phi(x(1-\pi_0)^{\frac{1}{2}}\sigma_1^{-1}).\end{aligned}$$

Given some value of $\beta_{i1}>0$ the probability that its estimator is also positive, and the marker is selected into the score, is

$$\Pr(\hat{\beta}_{i1}>0, \zeta|\beta_{i1}) = \Phi(q_0 - n_1^{\frac{1}{2}}\beta_{i1}) - \Phi(q_1 - n_1^{\frac{1}{2}}\beta_{i1})$$

Similarly given $\beta_{i1}>0$

the probability that its estimator is negative, and the marker is selected, is

$$\Pr(\hat{\beta}_{i1}<0, \zeta|\beta_{i1}) = \Phi(-q_1 - n_1^{\frac{1}{2}}\beta_{i1}) - \Phi(-q_0 - n_1^{\frac{1}{2}}\beta_{i1})$$

Finally the conditional mean of β_{i2} given β_{i1} is given by properties of the bivariate normal distribution as $E(\beta_{i2}|\beta_{i1}) = \frac{\sigma_{12}}{\sigma_1^2}\beta_{i1}$. The integral can be evaluated numerically, yielding values for power and accuracy from equations 2–5.

Binary traits

The forgoing is based on linear regression, which is the usual approach for quantitative traits. For binary traits the standard analysis is logistic regression, used both for estimating the coefficients β_{i1} in the polygenic score and for testing the association of the score in a replication sample. For small effects the log-odds are approximately linear in the predictors, so we may continue to work in a linear regression framework for estimating power and accuracy. That is, the binary trait is coded as 0/1 and treated as the response in ordinary linear regression. The variance of Y_2 in equation 2 is now the binomial variance $P_2(1-P_2)$ where P_j is the proportion of study subjects with $Y_j=1$. In a prospective sample, P_j is the population proportion of the trait, whereas in a case/control sample (to be discussed further below), it is the sampling proportion of cases.

The binary traits are now assumed to arise from a liability threshold model, under which all individuals have an underlying normally distributed trait, called the liability, and all those whose liability exceeds a fixed threshold will exhibit the trait. Although the liability is not directly observed, this model has several advantages for modelling polygenic effects, including independence of the genetic effects from the trait prevalence, and an elegant linear transformation between effects on liability to corresponding effects on the observed (0/1) trait. This model has recently been elucidated by several authors for studying the quantitative genetics of binary traits in humans, and the reader is referred to their papers for more detailed discussion [21,24,30].

Assuming the marginal liabilities L_j are distributed as a standard normal, the threshold for exhibiting trait $j\in\{1,2\}$ is $\tau_j = \Phi^{-1}(1-K_j)$ where K_j is the population prevalence. The genetic effects β_{ij} are now taken to act on liability, and for small effects a linear transformation to the corresponding effect on the observed trait may be obtained as [30]

$$\beta_{ij}\frac{\mathrm{cov}(L_j, Y_j)}{\mathrm{var}(L_j)} = \beta_{ij}E(L_j Y_j) = \beta_{ij}K\frac{\phi(\tau_j)}{K} = \phi(\tau_j)\beta_{ij} \qquad (11)$$

Given the genetic variance-covariance matrix $\mathbf{\Sigma_Y}$ on the liability scale, the statistical properties of the polygenic score may now be calculated as before, but substituting $\phi(\tau_j)^2\sigma_j^2$ for σ_j^2 and $\phi(\tau_1)\phi(\tau_2)\sigma_{12}$ for σ_{12} throughout, and using $P_1(1-P_1)n_1^{-1}$ as the sampling variance of $\hat{\boldsymbol{\beta}}_{ij}$.

Sensitivity and specificity are often of interest in the prediction of binary traits. In particular, the accuracy of a predictor can be assessed by the AUC constructed as follows. Subjects are classified such that those with a polygenic score above a fixed threshold are predicted to have the trait, those below the threshold to not have it. Sensitivity is the proportion of subjects with the trait who are correctly predicted as such, and specificity the proportion of subjects without the trait correctly predicted as such. Each possible threshold leads to a value of sensitivity and specificity, defining the receiver operator characteristic curve by plotting sensitivity against 1-specificity. The AUC can be defined as the probability that a pair of subjects, one with the trait and one without, is correctly classified by the predictor. Because the central limit theorem implies that the polygenic score is normally distributed, the expected AUC can be calculated as [21]

$$AUC = \Phi\left(\frac{E(\hat{S}|Y_2=1) - E(\hat{S}|Y_2=0)}{\sqrt{\mathrm{var}(\hat{S}|Y_2=1) + \mathrm{var}(\hat{S}|Y_2=0)}}\right) \qquad (12)$$

In this expression, $\hat{S}$ is formed from effects on Y_1 estimated on the observed scale whereas the conditional means and variances are conveniently calculated on the liability scale for Y_2. There is a linear transformation between effects on Y_1 and those on Y_2, defined by their bivariate normal distribution, and equation 11 gives another linear transformation between effects on Y_2 and those on L_2. Equation 12 may therefore be equivalently written in terms of effects on L_2 with the corresponding score denoted $\hat{S}_{L_2}$. The conditional means and variances are functions of the variance in L_2 explained by $\hat{S}_{L_2}$ [21,40], which is $R^2_{\hat{S},Y_2}\,\mathrm{var}(Y_2/\phi(\tau_2))$, giving

$$E(\hat{S}_{L_2}|Y_2=1) = \frac{\phi(\tau_2)}{K_2}\frac{R^2_{\hat{S},Y_2}K_2(1-K_2)}{\phi^2(\tau_2)} = \frac{R^2_{\hat{S},Y_2}(1-K_2)}{\phi(\tau_2)}$$

and

$$\mathrm{var}(\hat{S}_{L_2}|Y_2=1) = \frac{R^2_{\hat{S},Y_2}K_2(1-K_2)}{\phi(\tau_2)^2}\left[1 - \frac{R^2_{\hat{S},Y_2}(1-K_2)}{\phi(\tau_2)}\left(\frac{\phi(\tau_2)}{K_2} - \tau_2\right)\right].$$

Similarly

$$E(\hat{S}_{L_2}|Y_2=0)=\frac{-R^2_{\hat{S},Y_2}K_2}{\vartheta(\tau_2)}$$

and

$$\mathrm{var}(\hat{S}_{L_2}|Y_2=0)=\frac{R^2_{\hat{S},Y_2}K_2(1-K_2)}{\vartheta(\tau_2)^2}\left[1-\frac{R^2_{\hat{S},Y_2}K_2}{\vartheta(\tau_2)}\left(\frac{\vartheta(\tau_2)}{1-K_2}+\tau_2\right)\right].$$

Case/control studies

In case/control studies the increased ascertainment of cases leads to departure from the normal distribution of liability assumed in the previous subsection. To overcome this problem, it is again assumed that there is a linear transformation from an effect β_{i1} on liability to one on the observed trait in which the 0/1 response denotes ascertained case status.

When there is no selection on Y_j or G_i the regression of Y_j on G_i has coefficient $\frac{\mathrm{cov}(Y_j,G_i)}{\mathrm{var}(G_j)}=\mathrm{cov}(Y_j,G_i)=\vartheta(\tau_j)\beta_{ij}$ from equation 11. The converse regression of G_i on Y_j has coefficient $\frac{\mathrm{cov}(Y_j,G_i)}{\mathrm{var}(Y_j)}=\frac{\vartheta(\tau_j)\beta_{ij}}{K_j(1-K_j)}$. The latter will also apply when there is ascertainment on Y_j, but the regression of G_i on ascertained Y_j can also be written as $\frac{\mathrm{cov}(Y_j,G_i|A)}{\mathrm{var}(Y_j|A)}$, where A denotes ascertainment, so that

$$\mathrm{cov}(Y_j,G_i|A)=\frac{\vartheta(\tau_j)\beta_{ij}\,\mathrm{var}(Y_j|A)}{K_j(1-K_j)}$$

The desired quantity is the coefficient for the regression of ascertained Y_j on G_i which is thus

$$\frac{\vartheta(\tau_j)\beta_{ij}\,\mathrm{var}(Y_j|A)}{K_j(1-K_j)\mathrm{var}(G_i|A)}$$

In general the variance of genetic markers G_i will differ from 1 under ascertainment but it will henceforth be assumed that its expectation over markers is approximately 1. A heuristic justification for this assumption is given in the Text S1. Based on this assumption, an effect β_{ij} on liability is transformed by the factor

$$\vartheta(\tau_j)\frac{P_j(1-P_j)}{K_j(1-K_j)} \tag{13}$$

to the observed case/control scale. Similarly to before, given the genetic variance-covariance matrix $\mathbf{\Sigma_Y}$ on the liability scale, the properties of the polygenic score can be calculated on the observed scale, substituting $\vartheta(\tau_j)^2\frac{P_j^2(1-P_j)^2}{K_j^2(1-K_j)^2}\sigma_j^2$ for σ_j^2 and $\vartheta(\tau_1)\vartheta(\tau_2)\frac{P_1(1-P_1)P_2(1-P_2)}{K_1(1-K_1)K_2(1-K_2)}\sigma_{12}$ for σ_{12} throughout, and using $P_1(1-P_1)n_1^{-1}$ as the sampling variance of $\hat{\beta}_{ij}$.

To obtain the AUC, the same approach as before is used, but now using

$$R^2_{\hat{S},Y_2}\,\mathrm{var}\left(\frac{K_2(1-K_2)}{\vartheta(\tau_2)P_2(1-P_2)}Y_2\right)$$

as the variance in L_2 explained by $\hat{S}_{L_2}$. Therefore,

$$E(\hat{S}_{L_2}|Y_2=1)=\frac{\phi(\tau_2)}{K_2}R^2_{\hat{S},Y_2}\left(\frac{K_2(1-K_2)}{\phi(\tau_2)P_2(1-P_2)}\right)^2P_2(1-P_2)$$
$$=\frac{R^2_{\hat{S},Y_2}K_2(1-K_2)^2}{\phi(\tau_2)P_2(1-P_2)}$$

and

$$\mathrm{var}(\hat{S}_{L_2}|Y_2=1)=R^2_{\hat{S},Y_2}\left(\frac{K_2(1-K_2)}{\phi(\tau_2)P_2(1-P_2)}\right)^2P_2(1-P_2)$$
$$\left[1-\frac{R^2_{\hat{S},Y_2}K_2(1-K_2)^2}{\phi(\tau_2)P_2(1-P_2)}\left(\frac{\phi(\tau_2)}{K_2}-\tau_2\right)\right]$$

Similarly

$$E(\hat{S}_{L_2}|Y_2=0)=\;=\frac{-R^2_{\hat{S},Y_2}K_2^2(1-K_2)}{\vartheta(\tau_2)P_2(1-P_2)}$$

and

$$\mathrm{var}(\hat{S}_{L_2}|Y_2=0)=R^2_{\hat{S},Y_2}\left(\frac{K_2(1-K_2)}{\phi(\tau_2)P_2(1-P_2)}\right)^2P_2(1-P_2)$$
$$\left[1-\frac{R^2_{\hat{S},Y_2}K_2^2(1-K_2)}{\phi(\tau_2)P_2(1-P_2)}\left(\frac{\phi(\tau_2)}{K_2}+\tau_2\right)\right]$$

The transformation from liability to observed scales differs from that of Lee et al [30], which is for the total genetic liability (their equation 19). Here the interest is in the individual marker effects on the observed scale, because they are what are estimated when constructing the polygenic score.

Liability R^2

The derived expressions involve $R^2_{\hat{S},Y_2}$ which is the coefficient of determination on the observed scale. Lee et al have argued that, for a genetic predictor, R^2 on the liability scale is more interpretable for binary traits as it is invariant to the population prevalence and sampling ratio [25]. An approximate transformation to the liability scale is obtained by transforming the genetic effects using equation 13 and rescaling the trait variance from the binomial variance on the observed scale to the unit variance on the liability scale. Therefore,

$$R^2_{liab}\approx R^2_{\hat{S},Y_2}\frac{K_2^2(1-K_2)^2}{\vartheta(\tau_2)^2P_2(1-P_2)}$$

Log-risk model

An alternative to the liability threshold model is the log-risk model for binary traits, which is equivalent to the logistic model in the limit of low prevalence. Here the polygenic score estimates the

log risk of disease, which is assumed to be normally distributed in the population with mean $\log K - \log \lambda_S$ and variance $2 \log \lambda_S$, where λ_S is the sibling relative recurrence risk [17,20]. Under this model the log risk has the same variance in cases and controls, but the mean log risk among cases is increased by that same variance, becoming $\log K + \log \lambda_S$. This model allows a simpler calculation of AUC for rare disease, which is given here but not pursued further.

Given K_j and λ_{S_j} and denoting log-risk of trait j by R_j, the transformation from log-risk to observed scales is

$$\frac{\mathrm{cov}(R_j, Y_j)}{\mathrm{var}(R_j)} = \frac{E(R_j Y_j) - E(R_j)E(Y_j)}{2 \log \lambda_{S_j}} = \frac{K_j(\log K_j + \log \lambda_{S_j}) - (\log K_j - \log \lambda_{S_j})K_j}{2 \log \lambda_{S_j}} = K_j$$

with the same adjustment for case/control ascertainment (equation 13). The difference in polygenic scores between cases and controls is the variance of the score,

$$E(\hat{S}_{L_2} | Y_2 = 1) - E(\hat{S}_{L_2} | Y_2 = 0) = R^2_{\hat{S},Y_2} \left(\frac{K_2(1-K_2)}{K_2 P_2 (1-P_2)} \right)^2 P_2(1-P_2) = \frac{R^2_{\hat{S},Y_2}(1-K_2)^2}{P_2(1-P_2)}$$

Since the polygenic score has the same variance in cases and controls, equation 12 gives the AUC as

$$AUC = \Phi \left(\sqrt{\frac{R^2_{\hat{S},Y_2}(1-K_2)^2}{2P_2(1-P_2)}} \right)$$

Simulations

The derived expressions were compared to simulations in which the major assumptions were examined under realistic scenarios. These assumptions include a large number of markers with effects, for equality in equation 7, and small genetic effects, so that effects on the liability scale are approximately linear. In case/control designs the disease prevalence is assumed to be not too small, so that the variance of the ascertained genotypes remains near 1 as assumed in equation 13 and Text S1. Effects are assumed to be normally distributed on the standardised genotype scale. Sample sizes are assumed large so that estimates of genetic effects are normally distributed.

A baseline scenario was defined to reflect that seen in recent studies, as follows. Two normally distributed traits were simulated with explained genetic variances 0.4 and 0.3 and correlation of genetic effects of 0.65. Genotypes from 100,000 independent SNPs were simulated, with minor allele frequencies uniformly distributed on (0.01, 0.5). This reflects current marker panels that directly explain about half the heritability [1]. The proportion of null SNPs was 0.95 or 0.99 [9], with the same SNPs having effects for both traits. Their effect sizes were drawn from the bivariate normal distribution such that the desired variances and covariance were attained. The traits were then generated from the quantitative model in equation 1.

The polygenic score was estimated using the first trait in a sample of 4000 unrelated subjects. The score was constructed using P-value thresholds of 0.1 and 0.001 for $\pi_0 = 0.95$ and 0.99 respectively; these thresholds yielded the highest R^2 and AUC values. The score was then tested for association with the second trait in an independent sample of 4000 subjects. The correlation and mean square error between the score and the second trait were also estimated in the second sample. The association tests were used in equation 6 to estimate the explained genetic variances in the first and second samples in turn, and then the covariance between effects in the two samples, each time keeping other parameters fixed to their simulation values.

Table S1 shows estimates from 1000 simulations compared to the analytic values, for the three estimators discussed. Mean square error for the allele count estimator is not meaningful without further scaling of the polygenic score, which is a further problem not of present interest. All simulations agree well with the analytic results. Because the variances and covariance are bounded in (0,1), their median estimates are shown with the coverage, rather than their means. The proposed estimating equations are seen to be accurate, but the confidence intervals are anti-conservative when the number of markers with effects is low, here 1000. This is because the realised variance and covariance in equation 7 depart from their large m expectation, with resulting over-dispersion in the estimating equation (left hand side of equation 6). However when the number of markers with effects is 5000, the correct coverage is attained.

The traits were then treated as liabilities for binary diseases with prevalence 0.2. Disease status was simulated prospectively, as in a cohort study. The polygenic score was estimated and tested using both linear and logistic regression. Table S2 shows estimates of power and AUC compared to the analytic values. Results for the shrinkage estimator are identical to the regression estimator and are not shown. All simulations agree well with the analytic results, and the proposed estimating equations are accurate. The results for logistic regression agree well with those for linear regression, justifying the use of the latter to derive the analytic results.

Then, a case/control design was simulated in which the disease prevalence was now 0.001. The same total sample sizes were used but included equal numbers of cases and controls. A computationally efficient approach to this simulation is described in Text S2. The results are given in Table S3. Again all simulations are seen to agree with the analytic values, but when the number of markers with effects is low, there is a downward bias in the parameter estimates and the confidence intervals of the parameter estimates are anti-conservative. Again the logistic regression results agree well with those for linear regression. Taking Tables S1, S2, S3 together, the analytic methods are accurate for the strongest effects likely to be seen in current studies, but when the number of SNPs with effects is about 1000, there is downward bias in the effect estimates and under-coverage of the confidence intervals, the degree of which appears to vary with the strength of the association.

To assess robustness to normality of the marker effects, the simulations were repeated with the effects drawn from Laplace distributions and then rescaled to give the same explained variance and correlation as before. Instead of $\pi_0 = 0.95$ and $P<0.1$, simulations with $\pi_0 = 0$ and $P<1$ were performed to verify that this situation does not assume normality. The results in Table S4, Table S5 and Table S6 confirm this to be the case, whereas when $\pi_0 = 0.99$ and $P<0.001$ the analytic expressions tend to underestimate the power and accuracy. This is due to the heavier tails of the Laplace distribution compared to the normal, and quantitatively different results would be seen for different generating models. Again, bias and under-coverage is seen when there are 1000 markers with effects.

Supporting Information

Table S1 Simulations of quantitative traits compared to analytic results. Analytic values are in parentheses. Genotypes for 100,000 SNPs were simulated in 4000 subjects in each of two samples. Minor allele frequencies were drawn from Unif(0.01,0.5). Effect sizes in the two samples were drawn from the bivariate normal distribution with marginal variances 0.4, 0.3 and correlation 0.65. π_0, proportion of SNPs having no effect on traits. *P*, *P*-value for including SNP in the polygenic score. NCP, non-centrality parameter. Power computed at $\alpha = 0.05$. MSE, mean square error. σ_1^2, σ_2^2, ρ, median estimates of model parameters, with coverage of 95%CI in brackets.
(DOCX)

Table S2 Simulations of binary traits in prospective samples compared to analytic results. Analytic values are in parentheses. Genotypes for 100,000 SNPs were simulated in 4000 subjects in each of two samples. Minor allele frequencies were drawn from Unif(0.01,0.5). Effect sizes on liability were drawn from the bivariate normal distribution with marginal variances 0.4, 0.3 and correlation 0.65. Trait prevalence was 0.2 in both samples. π_0, proportion of SNPs having no effect on traits. *P*, *P*-value for including SNP in the polygenic score. NCP, non-centrality parameter. Power computed at $\alpha = 0.05$. AUC, area under receiver-operator characteristic curve. σ_1^2, σ_2^2, ρ, median estimates of model parameters, with coverage of 95%CI in brackets.
(DOCX)

Table S3 Simulations of binary traits in case/control samples compared to analytic results. Analytic values are in parentheses. Genotypes for 100,000 SNPs were simulated in 4000 subjects in each of two samples. Minor allele frequencies were drawn from Unif(0.01,0.5). Effect sizes on liability were drawn from a bivariate normal distribution with marginal variances 0.4, 0.3 and correlation 0.65. Trait prevalence was 0.001 in both samples, cases and controls sampled in equal proportion. π_0, proportion of SNPs having no effect on traits. *P*, *P*-value for including SNP in the polygenic score. NCP, non-centrality parameter. Power computed at $\alpha = 0.05$. AUC, area under receiver-operator characteristic curve. σ_1^2, σ_2^2, ρ, median estimates of model parameters, with coverage of 95%CI in brackets.
(DOCX)

Table S4 Simulations of quantitative traits compared to analytic results. Analytic values are in parentheses. Genotypes for 100,000 SNPs were simulated in 4000 subjects in each of two samples. Minor allele frequencies were drawn from Unif(0.01,0.5). Effect sizes in the two samples were drawn from Laplace distributions such that their marginal variances were 0.4, 0.3 and their correlation was 0.65. π_0, proportion of SNPs having no effect on traits. *P*, *P*-value for including SNP in the polygenic score. NCP, non-centrality parameter. Power computed at $\alpha = 0.05$. MSE, mean square error. σ_1^2, σ_2^2, ρ, median estimates of model parameters, with coverage of 95%CI in brackets.
(DOCX)

Table S5 Simulations of binary traits in prospective samples compared to analytic results. Analytic values are in parentheses. Genotypes for 100,000 SNPs were simulated in 4000 subjects in each of two samples. Minor allele frequencies were drawn from Unif(0.01,0.5). Effect sizes on liability were drawn from Lapace distributions such that their marginal variances were 0.4, 0.3 and their correlation was 0.65. Trait prevalence was 0.2 in both samples. π_0, proportion of SNPs having no effect on traits. *P*, *P*-value for including SNP in the polygenic score. NCP, non-centrality parameter. Power computed at $\alpha = 0.05$. AUC, area under receiver-operator characteristic curve. σ_1^2, σ_2^2, ρ, median estimates of model parameters, with coverage of 95%CI in brackets.
(DOCX)

Table S6 Simulations of binary traits in case/control samples compared to analytic results. Analytic values are in parentheses. Genotypes for 100,000 SNPs were simulated in 4000 subjects in each of two samples. Minor allele frequencies were drawn from Unif(0.01,0.5). Effect sizes on liability were drawn from Laplace distributions such that their marginal variances were 0.4, 0.3 and their correlation was 0.65. Trait prevalence was 0.001 in both samples, cases and controls sampled in equal proportion. π_0, proportion of SNPs having no effect on traits. *P*, *P*-value for including SNP in the polygenic score. NCP, non-centrality parameter. Power computed at $\alpha = 0.05$. AUC, area under receiver-operator characteristic curve. σ_1^2, σ_2^2, ρ, median estimates of model parameters, with coverage of 95%CI in brackets.
(DOCX)

Text S1 Variance of genetic marker in a case/control sample.
(DOCX)

Text S2 Simulation of genotypes in a case/control study.
(DOCX)

Author Contributions

Conceived and designed the experiments: FD. Performed the experiments: FD. Analyzed the data: FD. Wrote the paper: FD.

References

1. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90: 7–24.
2. Wray NR, Goddard ME, Visscher PM (2007) Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17: 1520–1528.
3. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752.
4. Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, et al. (2011) Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43: 969–976.
5. Hamshere ML, O'Donovan MC, Jones IR, Jones L, Kirov G, et al. (2011) Polygenic dissection of the bipolar phenotype. Br J Psychiatry 198: 284–288.
6. Bush WS, Sawcer SJ, de Jager PL, Oksenberg JR, McCauley JL, et al. (2010) Evidence for polygenic susceptibility to multiple sclerosis–the shape of things to come. Am J Hum Genet 86: 621–625.
7. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.
8. Simonson MA, Wills AG, Keller MC, McQueen MB (2011) Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. BMC Med Genet 12: 146.
9. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, et al. (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet 44: 483–489.
10. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42: 937–948.
11. Peterson RE, Maes HH, Holmans P, Sanders AR, Levinson DF, et al. (2011) Genetic risk sum score comprised of common polygenic variation is associated with body mass index. Hum Genet 129: 221–230.
12. Carayol J, Schellenberg GD, Tores F, Hager J, Ziegler A, et al. (2010) Assessing the impact of a combined analysis of four common low-risk genetic variants on autism risk. Mol Autism 1: 4.
13. Kang J, Kugathasan S, Georges M, Zhao H, Cho JH (2011) Improved risk prediction for Crohn's disease with a multi-locus approach. Hum Mol Genet 20: 2435–2442.

14. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, et al. (2011) Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol 35: 506–514.
15. Witte JS, Hoffmann TJ (2011) Polygenic modeling of genome-wide association studies: an application to prostate and breast cancer. OMICS 15: 393–398.
16. Pharoah PD, Antoniou AC, Easton DF, Ponder BA (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 358: 2796–2803.
17. Clayton DG (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5: e1000540. doi:10.1371/journal.pgen.1000540.
18. Sawcer S, Ban M, Wason J, Dudbridge F (2010) What role for genetics in the prediction of multiple sclerosis? Ann Neurol 67: 3–10.
19. 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.
20. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, et al. (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31: 33–36.
21. Wray NR, Yang J, Goddard ME, Visscher PM (2010) The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6: e1000864. doi:10.1371/journal.pgen.1000864.
22. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, et al. (2006) Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med 8: 395–400.
23. Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3: e3395. doi:10.1371/journal.pone.0003395.
24. So HC, Sham PC (2010) A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. PLoS Genet 6: e1001230. doi:10.1371/journal.pgen.1001230.
25. Lee SH, Goddard ME, Wray NR, Visscher PM (2012) A better coefficient of determination for genetic profile analysis. Genet Epidemiol 36: 214–224.
26. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, et al. (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460: 753–757.
27. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, et al. (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet 44: 247–250.
28. Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, et al. (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet 43: 977–983.
29. Risch N (2001) The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol Biomarkers Prev 10: 733–741.
30. 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.
31. Janssens AC, Moonesinghe R, Yang Q, Steyerberg EW, van Duijn CM, et al. (2007) The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. Genet Med 9: 528–535.
32. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 32: 227–234.
33. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, et al. (2011) Genomic inflation factors under polygenic inheritance. Eur J Hum Genet 19: 807–812.
34. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, et al. (2010) Performance of common genetic variants in breast-cancer risk models. N Engl J Med 362: 986–993.
35. Goddard ME, Wray NR, Verbyla K, Visscher PM (2009) Estimating Effects and Making Predictions from Genome-Wide Marker Data. Statistical Science 24: 517–529.
36. Greenland S (2000) Principles of multilevel modelling. Int J Epidemiol 29: 158–167.
37. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.
38. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.
39. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4: e1000130. doi:10.1371/journal.pgen.1000130.
40. Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics: Longman.
41. Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression: Cambridge University Press.