Summary statistic analyses do not correct confounding bias

John B. Holmes*[1], Doug Speed[2,3], and David J. Balding[1,2]

[1]Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Australia.
[2]Aarhus Institute of Advanced Studies (AIAS), Aarhus University, Denmark.
[3]UCL Genetics Institute, University College London, United Kingdom.

January 25, 2019

Abstract

LD SCore regression (LDSC) has become a popular approach to estimate confounding bias, heritability and genetic correlation using only genome wide association study (GWAS) test statistics. SumHer is a newly-introduced alternative with similar aims. We show using theory and simulations that both approaches fail to adequately account for confounding bias, even when the assumed heritability model is correct. Consequently, these methods may estimate heritability poorly if there was inadequate adjustment for confounding in the original GWAS analysis. We also show that choice of summary statistic for use in LDSC or SumHer can have a large impact on resulting inferences. Further, covariate adjustments in the original GWAS can alter the target of heritability estimation, which can be problematic when LDSC or SumHer is applied to test statistics from a meta-analysis of GWAS with different covariate adjustments.

LD SCore regression (LDSC) uses genome-wide association test statistics to estimate confounding bias, the heritability tagged by SNPs (h^2_{SNP}), how h^2_{SNP} is distributed across the genome and the genetic correlation of pairs of traits [1, 2, 3, 4]. Its use of test statistics rather than individual genotype data means that it is effectively unlimited in sample size, and can make use of published studies that do not release the genotypes of participants. Moreover the test statistics can be obtained from a single GWAS or from a meta-analysis of multiple GWAS. These advantages have led to LDSC being very widely used.

LDSC regresses the test statistic at each SNP on an "LD score", defined as a sum of linkage disequilibrium (LD) coefficients over neighbouring SNPs. The regression slope and intercept are interpreted as, respectively, h^2_{SNP} and confounding bias not corrected in the GWAS analysis. SumHer

*corresponding author, email: john.holmes@unimelb.edu.au

[5] and S-LDSC [3, 4] generalise LDSC by introducing weights into the LD score. The weights correspond to a heritability model that relates the expected heritability of a SNP to its properties known *a priori*. SumHer uses fixed, SNP-specific weights reflecting LD and minor allele fraction (MAF). In the most recent version of S-LDSC [4], weights based on LD and MAF as well as functional annotations are estimated in the summary statistic analysis. HESS [6] and RSS [7] are other summary statistic methods that require more information than association test statistics.

Researchers using LDSC or SumHer rely on assumptions about the test statistics. Usually, these researchers have not performed the underlying GWAS analyses, but use test statistics obtained from public data repositories [8] that may lack information required to check these assumptions. Here, we examine the linear regression models underpinning these methods and assess their validity under a range of scenarios. We do not revisit the topic of the underlying heritability model [5], rather we will highlight problems that arise even when the simulation and analysis heritability models are the same.

We derive expected values of association statistics and show that confounding effects are SNP dependent, and correlated with LD score (Appendix, sections 1-2), contravening a fundamental assumption of LDSC and its SumHer analogue. Thus a global adjustment term can fail to remove confounding effects, although a multiplicative adjustment can correct an over-conservative use of genomic control [5].

We illustrate the magnitude of the problem through simulations. Our investigation covers two possible summary statistics and we show that inferences from LDSC or SumHer can be greatly impacted by this choice. Further, we show that the definition of h^2_{SNP} targeted by LDSC and SumHer varies with the covariates fitted in the GWAS analyses. This can be important in meta-analysis: if the component studies use different covariate adjustments, any subsequent summary statistic heritability analysis will merge estimates of different quantities.

Choice of test statistics

LDSC and SumHer both fit a linear regression to summary statistics $S_j, j = 1, \dots, m$, obtained from a GWAS on n individuals. Associated with SNP j is $h^2_j \geq 0$, interpreted as the expected heritability attributable uniquely to that SNP, with $h^2_{\mathrm{SNP}} = \sum_{j=1}^m h^2_j$. SumHer is the case $A = 1$ in

$$\mathrm{E}[S_j] = C\big(A + n\sum_i h^2_i r^2_{ij}\big), \tag{1}$$

while LDSC assumes both $C = 1$ and $h^2_j = h^2_{\mathrm{SNP}}/m$ for all j. In (1), r^2_{ij} is an estimated LD coefficient with $r_{ii} = 1$ (see Methods), while A and C are alternative adjustments for confounding effects not accounted for in the GWAS analysis; as they cannot both be estimated, we fix either $A = 1$ or $C = 1$ or both. Estimates of C using SumHer were reported to be much lower than the corresponding estimates of A from LDSC [5], but this was due to the difference in heritability model rather than whether the confounding term was additive or multiplicative. The SumHer results indicated that many GWAS had over-corrected for confounding ($C < 1$), whereas LDSC analyses of the same data typically found $A > 1$, indicating a need for further confounding adjustment [5].

In practice, S_j is often the Wald statistic T_j^2 from a classical simple linear regression [9, 1, 3], which can be inferred from p-values. Its null distribution is $F_{1,n-2}$, which converges to χ_1^2 as n increases. However, LDSC was proposed assuming that $S_j = n\hat{\alpha}_j^2$, where α_j is the effect of SNP j when both $\mathbf{Z}$ and $\mathbf{y}$ are standardised [1]. Assuming that no covariates were fitted, S_j is $n/(n{-}1)$ times the standardised regression sum of squares $S\tilde{S}R_j = (n{-}1)SSR_j/SST$. When covariates are included this equality no longer holds, but we check using simulation that $\mathrm{E}[S\tilde{S}R_j] \approx \mathrm{E}[n\hat{\alpha}_j^2]$, and where convenient we compute $\mathrm{E}[S\tilde{S}R_j]$ rather than $\mathrm{E}[n\hat{\alpha}_j^2]$.

Recently, GWAS test statistics have often been derived from a mixed regression model [10, 11] in which SNP j is tested while other SNPs are used to compute the variance structure of a random effect modelling the role of cryptic kinship and/or population structure. We report the expectation of a general mixed model test statistic (Appendix, section 2.7), but we do not examine it further here because the expectation requires quantities not usually available from GWAS data repositories.

A general model

To assess the validity of (1), we derive approximations for $\mathrm{E}[S_j]$ and perform simulation studies using the phenotype model

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I}), \tag{2}$$

where μ is an intercept, $\mathbf{Z}$ an $n \times m$ matrix of standardised SNP genotypes, $\boldsymbol{\alpha}$ a vector of SNP effect sizes, and $\boldsymbol{\Sigma}$ a diagonal matrix with jth entry σ_j^2. The $n \times p$ matrix $\mathbf{X}$ contains column-standardised covariate values, while $\boldsymbol{\beta}$ is a vector of covariate effects. If $\mathrm{Cor}(\mathbf{X}, \mathbf{Z}) \neq \mathbf{0}$, the $\mathbf{X}$ are confounders for genetic association analysis. The most important example is population structure when both $\mathbf{X}$ and $\mathbf{Z}$ vary with, for example, geography or social strata.

When (2) is used as a simulation model, usually only a subset of the available SNPs are assigned non-zero effects. When used as an analysis model, because the causal SNPs are unknown all available SNPs should be included in (2). This mismatch between simulation and analysis models arises as it is impossible in practice to limit analyses to causal SNPs.

Two definitions of h^2_{SNP}

We define $\sigma^2_y = \sum_{j=1}^m \sigma^2_j + \sigma^2_e$, the phenotypic variance after conditioning on covariates/confounders $\mathbf{X}$. However $\mathbf{X}$ may not be recorded, or may be omitted from the analysis, in which case σ^2_y cannot be estimated, and only the total phenotypic variance $\sigma^2_y + \sigma^2_c$ is available, where $\sigma^2_c = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/(n-1)$ with $'$ denoting transpose. This leads to two definitions of the heritability of SNP j [12]:

$$\begin{aligned} h^2_{j,\text{a}} &= \sigma^2_j/\sigma^2_y, & (3) \\ h^2_{j,\text{b}} &= \sigma^2_j/(\sigma^2_y + \sigma^2_c). & (4) \end{aligned}$$

The conditional heritability, $h^2_{j,\text{a}}$, is standard when the $\mathbf{X}$ are modelled as fixed effects [13, 14], while the marginal heritability, $h^2_{j,\text{b}}$, is usually preferred for random-effect covariates [15, 16]. We use h^2_j when there are no covariates or it is unimportant to distinguish $h^2_{j,\text{a}}$ from $h^2_{j,\text{b}}$. Henceforth we assume that the phenotype vector $\mathbf{y}$ is sample standardised, in which case $\sigma^2_j = h^2_j$.

Results

For derivations of the expectations given below, see Supporting Information, section B. Simulation results reported here are for SumHer analyses of LDAK phenotypes (see Methods); corresponding results using LDSC analyses of GCTA phenotypes are broadly similar (Appendix, Figures S1-4).

No confounding ($\text{Cor}(\mathbf{X}, \mathbf{Z}) = 0$)

For a single GWAS with no covariate/confounder effects

$$\text{E}[T^2_j] \approx c_j\left(1 + n\sum_i r^2_{ij}h^2_i\right) \tag{5}$$

which is (1) with $A = 1$ and $C = c_j = 1/(1 - \sum_i r^2_{ij}h^2_i)$. (5) resembles (1) with $A = 1$ and $C \neq 1$, but the deviation from unity does not indicate confounding. For complex traits h^2_i is typically small, so that c_j slightly exceeds 1 for many j. Further,

$$\text{E}[(n-1)\hat{\alpha}^2_j] = \text{E}[S\tilde{S}R_j] = \frac{\text{E}[SSR_j]}{\text{E}[SST]} \approx n\frac{\sigma^2_y + n\sum_i r^2_{ij}\sigma^2_i}{(n-1)\sigma^2_y} \approx 1 + n\sum_i r^2_{ij}h^2_i, \tag{6}$$

which is (1) with $A = C = 1$.

SumHer estimates of h^2_{SNP} based on GWAS summary statistics in the absence of covariate/confounder effects are centred close to the true value of 0.5 for both statistics (Figure 1(a)), so that for our simulations the deviation of the c_j from 1 appears to be negligible. The mean estimate of h^2_{SNP} does not noticeably change when A or C is estimated rather than fixed at the true values ($A = C = 1$), but the variance increases due to uncertainty arising from the additional parameter estimation.

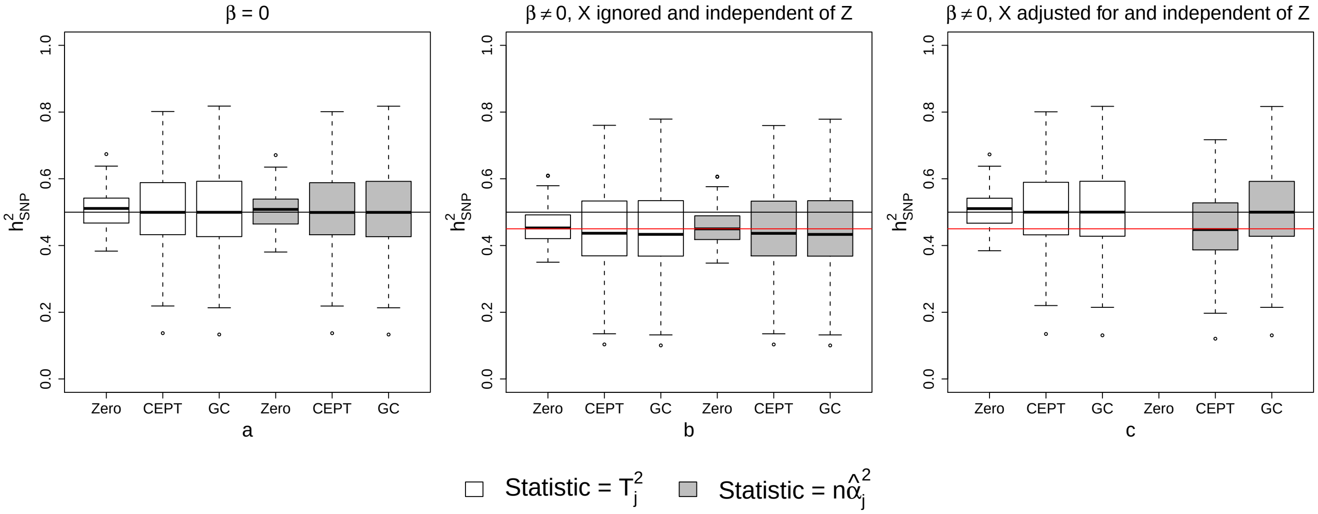

Figure 1: Estimates of h^2_{SNP} obtained from SumHer analysis of summary statistics calculated from a GWAS of LDAK simulated phenotypes. The black and red horizontal lines indicate the values of h^2_{SNPa} and h^2_{SNPb}, the SNP heritability without and with conditioning on covariates. Zero, CEPT and GC refer to no, A and C confounding terms in the analysis model. (a) Phenotypes with no covariate effects. (b) Phenotypes with covariate effects but $\mathbf{X}$ ignored in the analysis. (c) Phenotypes with covariate effects and $\mathbf{X}$ adjusted for in the analysis. The "Zero" estimates when $S_j = n\hat{\alpha}_j^2$ are all negative and are not shown.

When covariates affect $\mathbf{y}$ but $\mathbf{X}$ is ignored in the GWAS analysis, $h^2_{j,\text{b}}$ is estimated rather than $h^2_{j,\text{a}}$ because now $\text{E}[SST] \approx \sigma_y^2 + \sigma_c^2$ rather than σ_y^2. Again, the average estimate of h^2_{SNP} changes little when A or C is estimated rather than fixed at 1 (Figure 1(b)). When $\mathbf{X}$ is included in the GWAS analysis,

$$\text{E}[T_j^2] = \text{E}[\text{E}[T_j^2|\mathbf{X}]] \approx \frac{\sigma_y^2 + (n-p)\sum_i r_{ij}^2\sigma_i^2}{\sigma_y^2 - \sum_i r_{ij}^2\sigma_i^2} = c_j\left(1 + (n-p)\sum_i r_{ij}^2 h_{i,\text{a}}^2\right), \tag{7}$$

which is the same as (5) but with $n-p$ in place of n and $h_i^2 = h_{i,\text{a}}^2$. Further,

$$\text{E}[S\tilde{S}R_j] \approx C\left(1 + (n-p)\sum_i r_{ij}^2 h_{i,\text{a}}^2\right) = A + (n-p)\sum_i r_{ij}^2 h_{i,\text{b}}^2, \tag{8}$$

where $A = C = \sigma_y^2/(\sigma_y^2 + \sigma_c^2) = h^2_{\text{SNPb}}/h^2_{\text{SNPa}}$. We see from (7) that only h^2_{SNPa} can be estimated from T_j^2 (see also Figure 1(c)) whereas in (8) either h^2_{SNPa} or h^2_{SNPb} can be estimated, according to whether A or C is fitted.

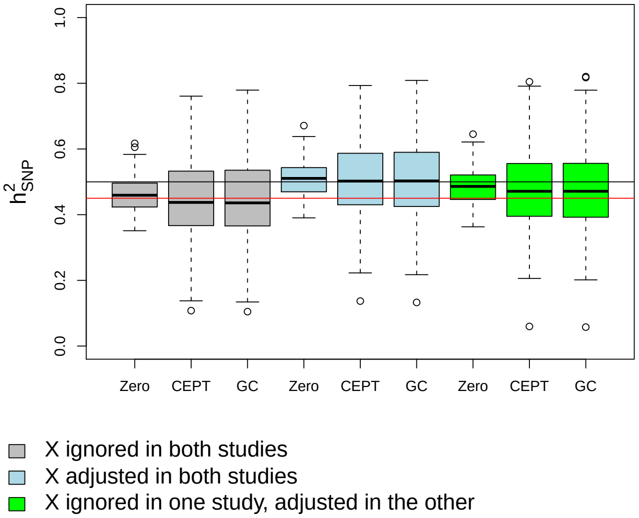

Figure 2: Estimates of h^2_{SNP} obtained from SumHer analysis of summary statistics calculated from a meta-analysis of two GWAS. The black and red horizontal lines indicate the values of h^2_{SNPa} and h^2_{SNPb}. Zero, CEPT and GC refer to no, A and C confounding terms in the analysis model.

Figure 2 shows that h^2_{SNPa} and h^2_{SNPb} are estimated with no apparent bias if, respectively, both studies did and did not adjust for covariates in a two-GWAS meta-analysis. Again, the inclusion of A or C terms has little effect on the mean estimates, as there is no confounding. When there is a mismatch in covariate adjustments between the two GWAS, the estimate of h^2_{SNP} is intermediate between h^2_{SNPa} and h^2_{SNPb} (Figure 2, green bars). In practice many meta-analyses do combine studies with different covariate adjustments, which may not adversely affect association tests but does affect heritability analyses. Examples include the meta-analyses of height [17] and blood pressure [18] re-analysed using LDSC [1], and those of psychiatric traits [19] and type 2 diabetes [20].

Confounding $(\mathbf{Cor}(\mathbf{X}, \mathbf{Z}) \neq \mathbf{0})$

When confounder $\mathbf{X}$ is ignored in the GWAS analysis:

$$\mathrm{E}[T_j^2] \approx c_j \left(1 + \frac{n a_j}{\sigma_y^2 + \sigma_c^2} + n \sum_i r_{ij}^2 h_{i,\mathrm{b}}^2\right) = c_j \mathrm{E}[S\tilde{S}R_j], \tag{9}$$

where $1/c_j = 1 - a_j/(\sigma_y^2+\sigma_c^2) - \sum_{i \neq j} r_{ij}^2 h_{i,\mathrm{b}}^2$ and $a_j = \left(\sum_{k=1}^p \mathrm{Cor}(\mathbf{Z}_j, \mathbf{X}_k)\boldsymbol{\beta}_k\right)^2$ with $\mathbf{X}_k$ denoting column k of $\mathbf{X}$. Assuming $c_j \approx 1$, only $h_{j,\mathrm{b}}^2$ is estimable, and (9) includes an additive constant resembling A in (1). However, this term is SNP-dependent, and for it to correspond to A in (1) we require a_j to be independent of LD score, which typically does not hold (see Appendix, section 1). Instead, we expect the estimate of h^2_{SNP} to be inflated by an amount that depends on both Cov($\mathbf{X}, \mathbf{Z}$) and $\sigma_c^2/(\sigma_y^2+\sigma_c^2)$. Replacing (9) with the SumHer regression model leads to similar difficulties.

When $\mathbf{X}$ is included in the GWAS analysis, the estimated SNP effect, $\hat{\alpha}_j$, can be obtained from the linear regression of the residuals of $\mathbf{y}|\mathbf{X}$ on the residuals of $\mathbf{Z}|\mathbf{X}$ [21], and

$$\mathrm{E}[T_j^2|\mathbf{Z},\mathbf{X}] \approx \frac{b\sigma_e^2 + n\sum_i (\hat{r}_{ij} - \hat{\boldsymbol{\gamma}}_i' \hat{\boldsymbol{\Sigma}}_{\mathbf{X}}^2 \hat{\boldsymbol{\gamma}}_j)^2 \sigma_i^2 / \bar{R}_j^2}{b\sigma_e^2 + \sum_{i\neq j} \bar{R}_i^2 \sigma_i^2 - \sum_{i\neq j} (\hat{r}_{ij} - \hat{\boldsymbol{\gamma}}_i' \hat{\boldsymbol{\Sigma}}_{\mathbf{X}}^2 \hat{\boldsymbol{\gamma}}_j)^2 \sigma_i^2 / \bar{R}_j^2}, \tag{10}$$

where $\boldsymbol{\Sigma}_{\mathbf{X}}^2 = \mathrm{Var}[\mathbf{X}]$ and, from the regression of $\mathbf{Z}_j$ on $\mathbf{X}$, $\bar{R}_j^2$ is one minus the coefficient of determination and $\hat{\boldsymbol{\gamma}}_j$ is the vector of estimated coefficients, while $b = (n{-}p{-}2)/(n{-}2)$. Further,

$$\mathrm{E}[S\tilde{S}R_j|\mathbf{Z}] \quad \approx \quad \frac{\sigma_y^2}{\sigma_y^2 + \sigma_c^2} (b(1{-}h^2_{\mathrm{SNPa}}) + n\sum_i (\hat{r}_{ij} - \hat{\boldsymbol{\gamma}}_i' \hat{\boldsymbol{\Sigma}}_{\mathbf{X}}^2 \hat{\boldsymbol{\gamma}}_j)^2 h^2_{i,\mathrm{a}} / \bar{R}_j^2). \tag{11}$$

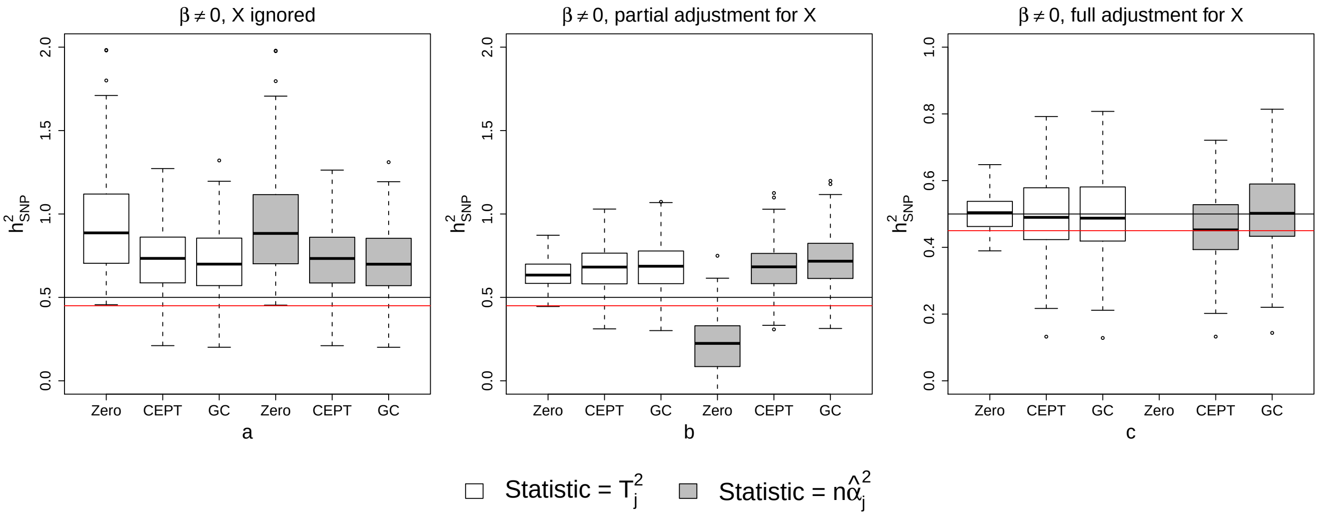

Figure 3: Similar to Figure 1, but here GWAS phenotypes are subject to confounding: phenotype means differ among three subpopulations that each consist of three sub-subpopulations. Subpopulations were constructed by applying k-means clustering to principal components of the SNPs with non-zero LDAK weight. Estimates of h^2_{SNP} from a GWAS with (a) no covariate adjustment, (b) adjustment for the three subpopulations but not the sub-subpopulations, (c) full covariate adjustment. Note that the y-axis differs among (a), (b) and (c).

Now, $S\tilde{S}R_j = (n{-}1{-}\mathbf{Z}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_j)\hat{\alpha}_j^2$ and, unlike when Cor($\mathbf{X}, \mathbf{Z}$) $= \mathbf{0}$, the term multiplying $\hat{\alpha}_j^2$ varies over SNPs.

As expected, ignoring confounders results in inflated estimates of h^2_{SNP} (Figures 3(a) and 4(a, b)). The estimable heritability parameter is h^2_{SNPb}, which is 0.45 for $C1$, 0.475 for $C2$, and 0.49 for

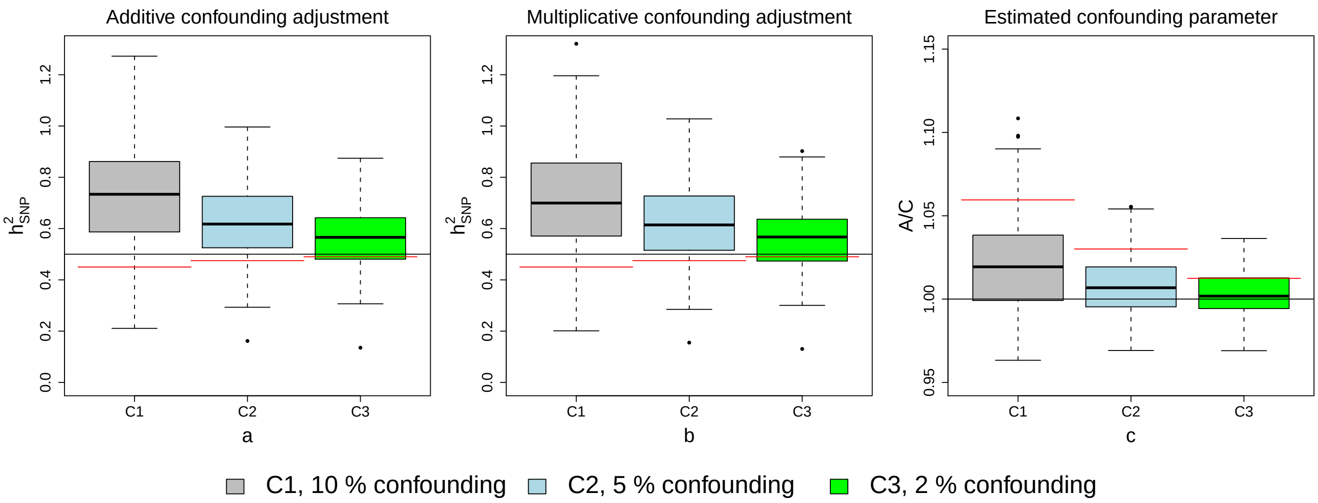

Figure 4: Estimating h^2_{SNP} and confounding parameters from phenotypes with differing proportions of phenotypic variance due to confounding when $h^2_{\text{SNPa}} = 0.5$. The confounding corresponds to ignoring subpopulations, which were constructed by applying k-means clustering to principal components of the SNPs with non-zero LDAK weight. The black lines in (a,b) indicate the simulated value of h^2_{SNPa} and the red lines the simulated value of h^2_{SNPb}, while the box-plot shows the distribution of h^2_{SNP} estimates when applying the confounding adjustment indicated in the plot heading. In (c), the black line at $A/C = 1$ corresponds to an estimate of zero confounding bias. Note that the y-axis differs between (a,b) and (c).

$C3$ phenotypes, yet the average estimates of h^2_{SNP} are in the reverse order ($C1 > C2 > C3$) because of the inadequately-corrected confounding (Figure 4(a,b)).

The A/C estimates are consistently too low (Figure 4(c)) because the positive association between a_j and LD score leads to some of the confounding being misinterpreted as heritability. Comparing $C1$, $C2$ and $C3$ phenotypes, we find that the bias in h^2_{SNP} is, like a_j, a function of $\sigma^2_c/(\sigma^2_y+\sigma^2_c)$, the proportion of phenotypic variance due to confounding. Figure 4(c) shows average A/C estimates > 1 in the presence of confounding, but this does not always hold, and estimates $A < 1$ have been reported, such as in GWAS of rheumatoid arthritis [1], age at first birth and number of children born [22], body mass index and hip-waist ratio [5] among others, which could be due to confounding of the type considered here. The level of bias in h^2_{SNP} was higher in the LDAK simulations than for the GCTA simulations (Appendix, section 3.1).

Our finding of inadequate adjustment for confounding is concordant with the results of two recent papers [23, 24] that analysed stratified populations, but not with Lee et al. [25] which considered confounding by parental genotype. This is because parental average genotype generates an additive genetic effect, which inflates the slope but not the intercept of the summary statistic regression. Their examples of non-trivial intercepts are based on twins, which can be viewed as combining two dependent samples each of unrelated individuals. This deviation from model assumptions, which is not confounding as defined at (2) because the inflation was not generated by an unaccounted effect $\mathbf{X}\boldsymbol{\beta}$, is corrected using an additive adjustment. A meta-analysis with overlapping samples also generates an intercept different from 1 that can be corrected using an additive adjustment (see Appendix, section 2.6 and Figure S9). In contrast, the population structure in our simulations implies some relatedness among all individuals, leading to a non-ignorable relationship between confounding and LD.

Partial covariate adjustment in the GWAS analysis reduced but did not eliminate bias in h^2_{SNP} estimates (Figure 3(b)) and led to divergence of the estimates based on T_j^2 and $n\hat{\alpha}_j^2$. Full covariate adjustment did lead to unbiased estimates of h^2_{SNPa} when $S_j = T_j^2$ whether or not we allowed $A \neq 1$ or $C \neq 1$ (Figure 3(c)). When $S_j = n\hat{\alpha}^2$, allowing $A \neq 1$ or $C \neq 1$ led to estimates of h^2_{SNPb} and h^2_{SNPa}, respectively. These results indicate that although confounding adjustment can mask causal signal, which is intuitively why (10) and (11) differ from (7) and (8), the differences appear negligible in this case. However, for populations with much stronger stratification adjusted for in the GWAS stage, the distinction between (10), (11) and (7), (8) was reported to be important for estimating h^2_{SNP} [26].

Discussion

We have shown theoretically, and illustrated using simulation, that GWAS confounding bias is in general SNP dependent and correlated with LD, so that the adjustment terms in the summary-statistic regression models of LDSC [1] and SumHer [5] can fail to adequately account for confounding bias, and hence also h^2_{SNP}, if the original GWAS analysis did not avoid confounding effects. This finding accords with findings by others [23, 24] and some statements and results in the original LDSC paper [1]. Firstly, in Bulik-Sullivan et al. [1] a small amount of polygenicity was inferred in simulations of confounding-only phenotypes, which was attributed to linked selection generating the correlation between confounding effect and LD score. Secondly, statements on interpreting the intercept were based on average results from distinct populations, not replicate samples from the same population. This ignores the structure, and hence confounding, that is specific to a population. For example, Supplementary Table 4C in Bulik-Sullivan et al. [1] shows that in the presence of confounding only ($h^2_{\text{SNP}} = 0$), one population (cou3) has higher LDSC h^2_{SNP} estimates for all three traits (0.144, 0.254, 0.229) than for any of the other 18 trait/population combinations (average: 0.030). Most importantly, the claim that LD score is not associated with confounding [1] was based on marginalising over the confounding component, defined as a function of allele frequency change, which is inappropriate as the test statistic and LD score are both SNP specific.

One source of error in published GWAS test statistics is genomic control, which applies a common multiplicative adjustment to all statistics, derived under an assumption of sparse causal effects. It tends to over-adjust for highly-polygenic traits, and can be corrected by use of C, the SumHer multiplicative adjustment term [5].

When covariates were fitted in the GWAS analysis, we found very different results according to whether S_j was chosen to be the Wald statistic T_j^2 or the statistic $n\hat{\alpha}_j^2$ used to justify LDSC [1]. Further, the estimable definition of h^2_{SNP} varies with the covariate adjustment performed in the original association analysis. The statistic $S\tilde{S}R$, closely related to $n\hat{\alpha}_j^2$, can be used to estimate h^2_{SNPb} regardless of the (non-confounder) covariates fitted, and hence a valid meta-analysis of h^2_{SNP} estimates is possible. However, $S\tilde{S}R$ is often not available in published GWAS results, and like T_j^2 it is subject to SNP-dependent confounding that can bias estimates of h^2_{SNP}.

We have only considered quantitative phenotypes, and we have not examined in detail the question of the validity of h^2_{SNP} analyses based on mixed model association statistics. We have shown that the expected values of such statistics contain terms not usually obtainable from public databases, but this may not preclude the development of h^2_{SNP} estimation based on linear regression of mixed model test statistics. As the expectation contains both intercept and slope terms, the relationship between the two can involve either only a shift (suggesting fitting A) or only a change in scale (suggesting fitting C) for h^2_{SNP} to remain estimable. We believe that change in scale is a more plausible relationship, both because of published simulations [5] and because theoretical properties suggest that mean estimates of $\hat{\alpha}_j$ (but not their variances) should be similar whether obtained using linear regression or a mixed model.

Methods

Data processing

We used genotypes from the eMERGE network [27], following the same quality control steps as [5]. From the 25,875 individuals, we randomly selected 8000 to form the study population, simulated their phenotypes and computed GWAS summary statistics. The remaining 17,875 individuals were used as a reference panel to compute r^2 values for the summary statistic analyses.

We also generated three meta-analyses by dividing the study population randomly into two studies of size 4000, and calculating summary statistics for each study, both without and with covariate adjustment. Each meta-analysis used within-study phenotype standardisation, and computed T_j^2 using inverse-variance [28] weighting.

Of the SNPs remaining after quality control, 558,431 had non-zero LDAK weights [29] and only these SNPs contribute causal effects under the LDAK model and to SumHer analyses. We also restricted LDSC analyses, and simulations under the GCTA model [30], to a set of 558,431 randomly-chosen SNPs.

Simulation of phenotypes and summary statistics

The GCTA model [30] is the special case of (2) in which, like the LDSC model, $\sigma_j^2 = h^2_{\mathrm{SNP}}/m$ for all j. LDAK [29, 31] is another special case of (2), and we adopt in the SumHer model its SNP weights based on LD and MAF [5]. For 150 iterations, we randomly sampled 35,000 causal SNPs and, under each of the GCTA and LDAK models, we generated five phenotypes with different covariate and confounding effects, such that $h^2_{\mathrm{SNPa}} = 0.5$ in all cases.

The five phenotypes are $\mathbf{y}_A$ (no covariates or confounding), $\mathbf{y}_B$ (covariate effect, no confounders), and $\mathbf{y}_{Ci}, i = 1, 2, 3$ (confounding, $\mathrm{Cor}(\mathbf{X}, \mathbf{Z}) \neq \mathbf{0}$). For $\mathbf{y}_B$, $\mathbf{X}$ has two columns, and the simulated effects were such that $\sigma_c^2 = \sigma_y^2/9$, so that $h^2_{\mathrm{SNPb}} = 0.45$. To explore incomplete control of confounding, for all $\mathbf{y}_C$ phenotypes confounders correspond to a two-level hierarchical population structure. First, three subpopulations were identified using k-means clustering on the leading 2 principal components (PCs) of the SNP correlation matrix $\mathbf{Z}^T\mathbf{Z}/m$, restricted to SNPs with non-zero LDAK weight in order to minimise any effect of correlated SNPS.

Within each of these subpopulations, three sub-subpopulations were defined by k-means clustering on the two leading PCs computed only from subpopulation members. We assigned different phenotype means to the nine sub-subpopulations, while SNP effect sizes remained the same. For $\mathbf{y}_C$ phenotypes we consider both $\mathbf{X}$ corresponding to the three subpopulations (two columns), and $\mathbf{X}$ corresponding to all nine sub-subpopulations (eight columns). The h^2_{SNPb} values were 0.45 (C1), 0.475 (C2) and 0.49 (C3).

$\mathbf{y}_A$ phenotypes and principal components were calculated using LDAK software, while k-means clustering and the simulation of $\mathbf{y}_B$ and $\mathbf{y}_C$ phenotypes was undertaken in R [32].

For all phenotypes we compute T_j^2 and $n\hat{\alpha}_j^2$, both with and without adjusting for covariates $\mathbf{X}$. Based on these statistics we estimate h^2_{SNP} using SumHer for LDAK phenotypes (results in main text) while for GCTA phenotypes (Appendix, section 3.1) we used LDSC as implemented in the LDAK software [29]. The two sets of results are broadly similar; we comment in the text on notable differences.

Large-effect SNPs

In part because of the problem of the unknown number of causal SNPs, but also due to model misspecification such as incomplete control of confounding, in many GWAS values of S_j arise that are extreme outliers under the GWAS assumed analysis model. Ideally, the solution would be to improve the analysis model, for example using a distribution with thicker tails than the Gaussian, or assigning an atom of prior probability at each SNP to a zero effect. However, because of computational advantages associated with model (2), in practice an ad-hoc data filtering approach is often adopted in which a SNP is removed if its estimated effect size is too large to be well-supported under the model. As we have control over confounding in our simulations, our main results do not use filtering. In Appendix, section 3.2, we consider the impact of filtering, where we follow [8] and exclude from analysis any SNP with $S_j > 80$. In the analysis of $\mathbf{y}_A$ and $\mathbf{y}_B$ simulations, no SNP was excluded, while for 32 C1, 2 C2, and 0 C3 LDAK and 44 C1, 14 C2, and 0 C3 GCTA simulations, at least one SNP was excluded for both T_j^2 and $n\hat{\alpha}_j^2$ when $\mathbf{X}$ was ignored in the GWAS analysis.

References

[1] Brendan K. Bulik-Sullivan, Po-Ru Loh, Hilary Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47:291–295, 2015.

[2] B. K. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium, Laramie Duncan, J. R. B. Perry, Nick Patterson, Elise B. Robinson, M. J. Daly, A. L. Price, and B. M. Neale. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, 47(11):1236–1241, 2015.

[3] Hilary K. Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh Farh, Stephan Ripke, Felix R. Day, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R. B. Perry, Yukinori Okada, Soumya Raychaudhuri, Mark Daly, Nick Patterson, Benjamin M. Neale, and Alkes L. Price. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, 47(11):1228–1235, 2015.

[4] Steven Gazal, Hilary K. Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoeh, Brendan K. Bulik-Sullivan, Benjamin M. Neale, Alexander Gusev, and Alkes L. Price. Linkage disequilibrium dependent architecture of human complex traits reveals action of negative selection. *Nat. Genet.*, 49:1421–1427, 2017.

[5] Doug Speed and David J. Balding. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.*, 2018.

[6] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.*, 99:139–153, 2016.

[7] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.*, 11:1561–1592, 2017.

[8] Jie Zheng, A. Mezut Erzurumluoglu, Benjamin L. Elsworth, John P. Kemp, Laurence Howe, Philip C. Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, Early Genetics and Lifecourse Epidemiology (EAGLE) Consortium, Beate St Pourcain, Nicole M. Warrington, Hilary K.

Finucane, Alkes L. Price, B. K. Bulik-Sullivan, Verneri Anttila, Lavinia Paternoster, Tom R. Gaunt, David M. Evans, and Benjamin M. Neale. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, 33(2):272–279, 2017.

[9] Jian Yang, Michael N. Weedon, Shaun Purcell, Guilaume Lettre, Karol Estrada, Cristen J. Willer, Albert V. Smith, Erik Ingelsson, Jeffrey R. O'Connell, Massimo Mangino, Reedik Mägi, Pamela A. Madden, Andrew C. Heath, Dale R. Nyholt, Nicholas G. Martin, Grant W. Montgomery, Timothy M. Frayling, Joel N. Hirschhorn, Mark I. McCarthy, Michael E. Goddard, Peter M. Visscher, and the GIANT Consortium. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.*, 19:807–812, 2011.

[10] Christoph Lippert, Jennifer Listgarten, Ting Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nat. Methods*, 10: 833–837, 2011.

[11] Po-Ru Loh, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjálmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, Paul M. Ridker, Benjamin M. Neale, Bonnie Berger, Nick Patterson, and Alkes L. Price. Efficient bayesian mixed model analysis increases association power in large cohorts. *Nat. Genet.*, 47:284–290, 2015.

[12] Omer Weissbrod, Jonathan Flint, and Saharon Rosset. Estimating snp-based heritability and genetic correlation in case-control studies directly and with summary statistics. *Am. J. Hum. Genet.*, 103:89–99, 2018.

[13] R.A. Mrode. *Linear models for the prediction of animal breeding values.* CABI publishing, 3 edition, 2014.

[14] Matti Pirinen, Peter Donnelly, and Chris C.A. Spencer. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.*, 7 (1):369–390, 2013.

[15] D. Heckerman, Deepti Gurdasani, Carl Kadie, Cristina Pomilla, Tommy Carstensen, Hilary Martin, Kenneth Ekoru, Rebecca N. Nsubuga, Gerald Ssenyomo, Anatoli Kamali, Pontiano Kaleebu, Christian Widmer, and Manjinder S. Sandhu. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl. Acad. Sci. U.S.A.*, 113 (27):7377–7382, 2016.

[16] Pierre De Villemereuil, Michael B. Morrissey, Shinichi Nagakawa, and Holger Schielzeth. Fixed effect variance and the estimation of repeatabilities and heritabilities: Issues and solutions. *J. Evol. Biol.*, 31(4):621–632, 2018.

[17] Allen H. Lango, K. Estrada, G. Lettre, S.I. Berndt, M.N. Weedon, F. Rivadeneira, C.J. Willer, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[18] The International Consortium for Blood Pressure Genome-Wide Association Studies et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, 2011.

[19] Aysu Okbay, Bart M.L Baselmans, Jan-Emmanuel De Neve, Patrick Turley, Michel G. Nivard, Mark Alan Fontana, S. Fleur W. Meddens, et al. Genetic variants associated with subjective well-being, depressive symptoms and neuroticism identified through genome-wide analyses. *Nat. Genet.*, 48:624–633, 2016.

[20] Robert A Scott, Laura J Scott, Reedik Mägi, Letizia Marullo, Kyle J Gaulton, Marika Kaakinen, et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes*, 66(11):2888–2902, 2017.

[21] Ragnar Frisch and Fredrick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933.

[22] Nicola Barban, Rick Jansen, Ronald de Vlaming, Ahmad Vaez, Jornt J. Mandemakers, et al. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat. Genet.*, 48(12):1462–1472, 2016.

[23] Ronald DeVlaming, Magnus Johannesson, Patrik K.E. Magnusson, M. Arfan Ikram, and Peter M. Visscher. Equivalence of LD-Score Regression and Individual-Level-Data Methods. 2017.

[24] Jermey J. Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Jørgensen, Hakhamanesh Mostafavi, Yair Field, Evan A. Boyle, Xinjun Zhang, Fernando Racimo, Jonathan K. Pritchard, and Graham Coop. Reduced signal for polygenic adaptation of height in UK Biobank. 2018.

[25] James J. Lee, Matt McGue, and Carson C. Iacomo, William G.and Chow. The accuracy of LD score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genet. Epidemiol.*, 42:783–795, 2018.

[26] Yang Luo, Xinyi Li, Xin Wang, Steven Gazal, Josep Maria Mercader, 23 and Me Research Team, SIGMA Type 2 Diabetes Consortium, Benjamin M. Neale, Jose C. Florez, Adam Auton, Alkes L. Price, Hilary K. Finucane, and Soumya Raychaudhuri. Estimating heritability of complex traits in admixed populations with summary statistics. 2018.

[27] Shefali S. Verma, Mariza de Andrade, Gerard Tromp, Helena Kuivaniemi, Elizabeth Pugh, Bahram Namjou-Khales, Shubhabrata Mukherjee, Gail P. Jarvik, Leah C. Kottyan, Amber Burt, Yuki Bradford, Gretta D. Armstrong, Kimberly Derr, Dana C. Crawford, Jonathan L. Haines, Rongling Li, David Crosslin, and Marylyn D. Ritchie. Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.*, 5:1–15, 2014.

[28] Cristen J. Willer, Yun Li, and Goncalo R. Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[29] D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.*, 91:1011–1021, 2012.

[30] J. Yang, S. Hong Lee, M.E. Goddard, and P. M. Visscher. Gcta: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, 88:76–82, 2011.

[31] Doug Speed, Na Cai, the UCLEB Consortium, Michael R. Johnson, Sergey Nejentsev, and David J. Balding. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.*, 49 (7):986–992, 2017.

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `http://www.R-project.org/`.