Use of allele scores as instrumental variables for Mendelian randomization

Stephen Burgess Simon G. Thompson

October 5, 2012

Summary

Background: An allele score is a single variable summarizing multiple genetic variants associated with a risk factor. It is calculated as the total number of risk factor-increasing alleles for an individual (unweighted score), or the sum of weights for each allele corresponding to estimated genetic effect sizes (weighted score). An allele score can be used in a Mendelian randomization analysis to estimate the causal effect of the risk factor on an outcome.

Methods: Data were simulated to investigate the use of allele scores in Mendelian randomization where conventional instrumental variable techniques using multiple genetic variants demonstrate “weak instrument” bias. The robustness of estimates using the allele score to misspecification (for example non-linearity, effect modification) and to violations of the instrumental variable assumptions was assessed.

Results: Causal estimates using an allele score were unbiased with appropriate coverage levels. The estimates were generally robust to misspecification of the allele score, but not to instrumental variable violations, even if the majority of variants in the allele score were valid instruments. Using a weighted rather than an unweighted allele score increased power, but the increase was small when genetic variants had similar effect sizes. Using the data under analysis to choose which variants to include in an allele score, or for deriving weights, resulted in substantial biases.

Conclusions: Allele scores enable valid causal estimates with large numbers of genetic variants. The stringency of criteria for genetic variants in Mendelian randomization should be maintained for all variants in an allele score.

Keywords: Mendelian randomization, allele scores, genetic risk scores, instrumental variables, weak instruments.

Word count: Abstract = 244; Paper = 4257.
Introduction

Allele scores (also called genetic risk scores, gene scores, or genotype scores) are a convenient way of summarizing a large number of genetic variants associated with a risk factor. An unweighted allele score is constructed as the total number of risk factor-increasing alleles present in the genotype of an individual. A weighted allele score can also be considered, where each allele contributes a weight reflecting an estimate of the effect of the corresponding genetic variant on the risk factor. These weights can be internally derived from the data under analysis, or externally derived from prior knowledge or an independent data source. In this way, multidimensional genetic data on variants associated with a risk factor can be collapsed into a single variable. Allele scores have been constructed for many traits, including fasting glucose [1], blood pressure [2] and high-density lipoprotein cholesterol [3].

Allele scores are important for the modelling of multifactorial polygenic traits, particularly when the allele score consists either of many common variants with small effects, or of rare variants. When several such variants are combined into an allele score, the score may explain a considerable proportion of variation in the risk factor, even if none of the variants individually does.

Mendelian randomization

In this paper, we consider the use of allele scores in Mendelian randomization: that is the application of instrumental variable methods with genetic instruments to estimate the causal effect of a risk factor on an outcome from observational data [4, 5]. Under the assumption that the genetic instruments used are specifically associated with the risk factor of interest, and not directly associated with either the outcome, nor associated with any potential confounding variable, a genetic instrumental variable divides the population into subgroups which systematically differ in the risk factor, but not in any competing risk factor [6]. The genetically-defined subgroups are analogous to treatment arms in a randomized controlled trial [7]. Any difference in the outcome between the subgroups is inferred to be causally due to the risk factor of interest, subject to the validity of the instrumental variable assumptions [8].

Violation of instrumental variable assumptions

Violation of these assumptions can occur for a number of biologically plausible reasons, including pleiotropic association of the genetic variant with a confounding variable, linkage disequilibrium with another functional variant associated with a confounding variable, and population stratification meaning that genetic associations reflect latent strata in the population [9, 10]. However, where there is substantial biological knowledge about a genetic variant (or variants) and the plausibility of a specific association with the risk factor, and where this association has been robustly demonstrated, the instrumental variable estimate can be reasonably assumed to represent a causal effect. Examples of genetic variants which have been used in this way for coronary heart disease include variants in the CRP gene for the causal effect of C-reactive protein [11],
and variants in the *IL6R* gene for the causal effect of interleukin-6 receptor [12].

**Using allele scores in Mendelian randomization**

Allele scores are used in Mendelian randomization for reasons of simplicity, increased power [13] and avoidance of weak instrument bias [14]. Their use requires the assumption that the allele score is an instrumental variable [15], and so is specifically associated with the risk factor and not with the outcome or confounders as above. This means that each variant which contributes to the allele score must be an instrumental variable [13]. As the biological effects of the variants in an allele score are typically not well-known, the instrumental variable assumptions may not be satisfied for all the variants. We demonstrate the problems resulting from departures from these assumptions, as well as from assumptions which are commonly made for mathematical convenience, such as the use of additive genetic models with no interactions between genetic variants. The aim of this paper is to show how use of an allele score resolves some of these problems; first in an idealized simulation study, and then in a range of more realistic scenarios.

**Review of current practice**

The term “genetic risk score” returns 2620 hits in the search engine Google Scholar and 72 hits in conjunction with the term “Mendelian randomization” (04/10/2012), so a comprehensive review of the available literature on allele scores is impractical. We here provide some references as examples of how allele scores have been used in practice. Lin et al. [16] used an unweighted and a weighted allele score based on 15 genetic variants in the context of risk prediction, deriving weights from the data under analysis. They found that a weighted allele score provided greater discrimination than an unweighted score when used in conjunction with conventional risk factors. Rasmussen et al. [1] and Ehret et al. [2] used a weighted allele score in the context of Mendelian randomization, deriving weights from the data under analysis. Rasmussen et al. chose 5 variants from genetic regions which showed significant *p*-values in the dataset, although the precise choice of variants was from a separate meta-analysis (which included the study under analysis). In Ehret et al., several of the 29 variants used in the allele score were novel, and were chosen on the basis of *p*-values in the dataset. Voight et al. [3] used a weighted allele score with 14 variants to perform Mendelian randomization, deriving weights from a published meta-analysis, although some studies were in common between the two analyses. It is not clear how the specific variants were chosen, although they are reported as having significant *p*-values in the dataset.
Initial analysis: valid instruments with equal-sized effects

We initially generate simulated data for the comparison of instrumental variable methods assuming that each of the genetic variants is a valid instrumental variable for the risk factor, and that each variant has the same effect on the risk factor. In our simulation model, we have a risk factor \( X \) which is a linear sum of a confounder \( U \), assumed unmeasured, a set of \( J \) independently distributed genetic variants \( G_j \) for \( j = 1, \ldots, J \) and a normally distributed error term. The outcome \( Y \) is a continuous variable calculated as the linear sum of the risk factor, the confounder and an independent error term. The data-generating model for individual \( i \) is:

\[
X_i = \sum_{j=1}^{J} \alpha G_{ij} + \alpha U_i + \epsilon_{Xi} \quad (1)
\]

\[
Y_i = \beta X_i + \beta U_i + \epsilon_{Yi}
\]

\( U_i, \epsilon_{Xi}, \epsilon_{Yi} \sim N(0, 1) \) independently.

To investigate realistic settings, data were simulated with 9, 25 and 100 genetic variants for 3000 individuals. We set \( \alpha_U, \beta_U = 1 \) so that the risk factor and outcome are positively correlated even without a causal effect of the risk factor on the outcome. Three values are taken for the causal effect \( \beta_X \) of 0, 0.2 and 0.4. We chose \( \alpha_G = 0.1 \) for 9 variants, \( \alpha_G = 0.06 \) for 25 variants, and \( \alpha_G = 0.03 \) for 100 variants with a minor allele frequency of 0.3 for each variant, meaning that the proportion of variation in the risk factor explained by the genetic variants above that expected by chance (the adjusted \( R^2 \)) was approximately 1.9% throughout, similar to the \( R^2 \) for the allele score \( = \sum_j G_{ij} \). This is a fairly typical proportion for many biomarkers (for example [11]). Although many traits have a heritability which is much greater than 1.9% [17], it is unlikely that this heritability can be attributed to genetic variants which are specifically associated with the trait of interest rather than those associated with potential confounders.

For each of 1000 simulated datasets, we calculate estimates of the causal effect using the allele score as an instrumental variable using the two-stage least squares (2SLS) method to give a point estimate and standard error [18]. In comparison, we also present results taking each genetic variant as a separate variable using the 2SLS and limited information maximum likelihood (LIML) methods [19] implemented using the \textit{ivreg2} command in Stata [20]. LIML and 2SLS give identical estimates in the case of a single instrumental variable such as an allele score. We focus on bias, coverage (proportion of 95% confidence intervals containing the true causal effect), and power (chance of detecting a non-null causal effect at a significance level of 5%).

When the strength of the instruments is low, estimates using multiple instruments are known to be biased in the direction of the observational confounded association and have poor coverage properties [21, 22]. The strength of the instruments is measured by the F statistic from the regression of the risk factor on the instruments [23]. Conventionally, instruments with an F statistic less than 10 are labelled as “weak”.
[24], although so-called ‘weak instrument bias’ is a continuous rather than a binary phenomenon. Instruments with lower expected F statistics correspond to estimates which are more biased.

In our simulations, the mean F statistic from the regression of the risk factor on the allele score is almost 60, meaning that causal estimates using the allele score should be unaffected by weak instrument bias. The F statistics using each of the genetic variants as a separate variable are much lower. Estimates from the LIML method are less affected by weak instrument bias than those from the 2SLS method [25].

Results

Table 1 displays results from each method: the median estimate across simulations, interquartile range (IQR) of estimates, coverage and power. The median estimate is given rather than the mean as the distribution of estimates has several extreme values. With the allele score and LIML methods, the theoretical mean estimate is undefined. The Monte Carlo standard error (the expected variation from the true value based on our 1000 simulations) of the median estimate is 0.004, and of the coverage is 0.7%.

We see that the estimates using 2SLS are biased throughout and coverage is less than the nominal 95% level. Bias acts in the direction of confounding, and is especially serious with large numbers of genetic variants of smaller effect size. Although the power reaches 100% in some cases, this is meaningless when the coverage is below the nominal level and is due to the large bias. The LIML estimates show good performance with bias compatible with zero, but coverage levels decrease as the number of variants increases. The median estimates using the allele score are compatible with being unbiased with correct coverage levels throughout. The precision of the allele score method is greater (i.e. has a lower IQR) than LIML, but the power is similar (where the coverage of each method are close to the correct 95%). This is expected for variants with equal-sized effects as, in this case, no information is lost by converting the multivariate data on genetic variants to a univariate unweighted allele score.

In conclusion, when correctly specified, allele scores allow valid estimation of causal effects using large numbers of genetic variants where conventional methods suffer from problems of bias and reduced coverage (overly narrow confidence intervals).

Investigating departures from the model

Strong assumptions are necessary for the correct specification of an unweighted allele score. In this section, we consider various departures from the data-generating model (1) considered above, which reflect more realistic situations. We examine how estimates from instrumental variable methods using the allele score are affected by these changes. We first introduce the various scenarios to be considered, before presenting and discussing the results. Unless otherwise stated, all parameters will take the same values as in the simulation considered above. In the main paper, we consider situations with 25 variants; results from models with 9 and 100 variants obtained by scaling the genetic parameters accordingly are given in the Web Appendix. Parameters have
Table 1: Instrumental variable estimates for genetic variants with equal-sized effects from allele score analysis and multivariable analyses using two-stage least squares (2SLS) and limited information maximum likelihood (LIML) methods: median estimate across simulations, interquartile range (IQR) of estimates, coverage and power (%)

<table>
<thead>
<tr>
<th></th>
<th>Null effect ($\beta_X = 0$)</th>
<th>Small effect ($\beta_X = 0.2$)</th>
<th>Moderate effect ($\beta_X = 0.4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median IQR Coverage</td>
<td>Median IQR Coverage Power</td>
<td>Median IQR Coverage</td>
</tr>
<tr>
<td>Data-generating model with 9 genetic variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00 0.19 95.0</td>
<td>0.20 0.19 94.5 35.6</td>
<td>0.40 0.17 96.7 79.7</td>
</tr>
<tr>
<td>2SLS</td>
<td>0.06 0.17 90.8</td>
<td>0.26 0.17 89.1 55.8</td>
<td>0.47 0.16 89.7 91.8</td>
</tr>
<tr>
<td>LIML</td>
<td>0.00 0.20 95.0</td>
<td>0.20 0.20 94.0 39.5</td>
<td>0.41 0.19 95.7 77.8</td>
</tr>
<tr>
<td>Data-generating model with 25 genetic variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00 0.18 96.9</td>
<td>0.20 0.19 95.2 36.3</td>
<td>0.40 0.18 96.6 77.5</td>
</tr>
<tr>
<td>2SLS</td>
<td>0.15 0.14 69.2</td>
<td>0.35 0.14 68.8 86.9</td>
<td>0.55 0.14 67.9 99.1</td>
</tr>
<tr>
<td>LIML</td>
<td>0.01 0.21 92.6</td>
<td>0.20 0.20 92.4 36.2</td>
<td>0.40 0.22 93.5 72.8</td>
</tr>
<tr>
<td>Data-generating model with 100 genetic variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>-0.01 0.18 95.4</td>
<td>0.20 0.18 95.7 35.7</td>
<td>0.40 0.17 95.2 77.0</td>
</tr>
<tr>
<td>2SLS</td>
<td>0.32 0.10 1.3</td>
<td>0.52 0.09 1.4 100.0</td>
<td>0.72 0.09 0.9 100.0</td>
</tr>
<tr>
<td>LIML</td>
<td>-0.01 0.30 79.2</td>
<td>0.21 0.30 80.5 42.6</td>
<td>0.41 0.27 82.1 70.2</td>
</tr>
</tbody>
</table>

been chosen to take plausible values with reference to real examples and principles, for example that main effects are generally larger than interaction terms.

1. Unequal variants: valid instruments with different-sized effects

In practice, it may be that some genetic variants have stronger associations with the risk factor than others. To model this, we draw the genetic effect sizes $\alpha_{Gj}$ for each genetic variant $j$ from independent normal distributions with mean 0.06 and standard deviation 0.018; so nearly all of the genetic effects sizes are between 0.02 and 0.12. In addition to a standard allele score where each risk-increasing allele contributed the same value to the allele score, we constructed a weighted allele score ($= \sum_j w_j G_{ij}$). The weights ($w_j$) were determined in three ways: internally from the same data used in the analysis, externally from an independent source, and from the coefficients in the generating model. In the first case, the weighted allele score method is identical to a 2SLS method, as the weights are the same as the coefficients from the first-stage regression in the 2SLS analysis. In the second case, weights are generated by sampling from a normal distribution around the true weight with a standard deviation of 0.04 and of 0.01. This represents uncertainty in the estimation of weights taken from the regression of the risk factor on the variants in an external data source of approximately the same size as the original dataset (3000 participants, imprecise weights) and of 16 times the size of the original dataset (48000 participants, precise weights). In the third case, the coefficients from the generating model are the true weights.
2. Main and secondary variants: valid instruments with a few large and many small effects

In some practical examples, a small number of main variants have large effects (here, two) and other secondary variants may have smaller effects, a model called a “major-gene/polygene model” by Pierce et al. [13]. We additionally consider a composite approach using 2SLS, estimating separate coefficients for the main variants, and including others in an unweighted allele score. This is compared with the weighted and unweighted allele score and 2SLS/LIML methods. In the generating model, the effect size for the two main variants is set at five times the size of the effect of the secondary variants. We set $\alpha_G = 0.046$ for the secondary variants and $\alpha_G = 0.23$ for the main variants so that the proportion of variation in the risk factor explained by the allele score ($R^2$) is 1.9%.

3. Selected variants: instruments chosen due to strength of association in the data under analysis

In practice, it may be that the investigator is uncertain if each of the alleles is truly associated with the risk factor in the population of interest and decides to only include in an allele score the variants which show the strongest association with the risk factor. To illustrate this approach, variants were ranked according to their strength of association with the risk factor and selected using two criteria: a fixed number of variants (5, 10), and a threshold $p$-value (0.05, 0.01).

4. Non-linear genetic effects: valid instruments with non-linear effects

In practice, it may be that some genetic variants do not have linear (that is additive or per allele) effects on the risk factor. We modify the data-generating model (1) by replacing the first line with:

$$X_i = \sum_{j=1}^{f} (\alpha_{G1} G_{ij} + \alpha_{G1j} \mathbb{1}_{(G_{ij}=1)}) + \alpha_U U_i + \epsilon_{Xi}$$

where $\mathbb{1}_{(\cdot)}$ is an indicator function, taking the value one when the subscripted condition is satisfied and zero otherwise. We set $\alpha_{G1} = 0.06$ and draw the effects $\alpha_{G1j}$ from a normal distribution with mean 0 and standard deviation 0.036. For heterozygotes ($G_{ij} = 1$), nearly all values of $\alpha_{G1} + \alpha_{G1j}$ are in the range $-0.02$ to $0.14$; 0 corresponds to a recessive model, and 0.12 to a dominant model.

5. Interactions between genetic variants: valid instruments with genetic interactions

In practice, it may be that there are statistical interactions between the genetic variants. These are often called gene–gene interactions, though are more properly thought
of as variant–variant interactions [26].

We modify the data-generating model (1) by replacing the first line with:

\[ X_i = \sum_{j=1}^{J} \alpha_{G1} G_{ij} + \sum_{j=1}^{J} \sum_{k > j}^{J} \alpha_{Gjk2} G_{ij} G_{ik} + \alpha_U U_i + \epsilon_{Xi}. \]  

We set \( \alpha_{G1} = 0.06 \) and draw the effects \( \alpha_{Gjk2} \) from a mixture distribution taking the value zero with probability 0.95 and a random value from a normal distribution with mean 0 and standard deviation 0.036 with probability 0.05. With 25 genetic variants, in each simulated dataset there will be an average of 15 interactions between genetic variants out of the 300 pairs of variants; these include a range of interactions from strongly negative (e.g. \( \alpha_{Gjk2} = -0.06 \)) to strongly positive (e.g. \( \alpha_{Gjk2} = +0.06 \)).

6. Interactions between a genetic variant and a covariate: valid common instruments with environmental interactions

In practice, it may be that there are statistical interactions between a genetic variant and a covariate which is not a confounder. These are often called gene–environment interactions, though are more properly thought of as examples of effect modification.

We modify the data-generating model (1) by replacing it with:

\[ X_i = \sum_{j=1}^{J} (\alpha_{G1} G_{ij} + \alpha_{Gj3} G_{ij} V_i) + \alpha_U U_i + \alpha_V V_i + \epsilon_{Xi}. \]  

\[ Y_i = \beta_X X_i + \beta_U U_i + \epsilon_{Yi} \]

\( U_i, V_i, \epsilon_{Xi}, \epsilon_{Yi} \sim \mathcal{N}(0, 1) \) independently.

The variable \( V \) is introduced as a covariate affecting the risk factor but not the outcome, so that \( V \) is not a confounder but an effect-modifier. To ensure that the model is similar to those considered previously in terms of instrument strength, we let \( \alpha_U = \alpha_V = \frac{1}{\sqrt{2}} \). We set \( \alpha_{G1} = 0.06 \) and draw the modifying effects \( \alpha_{Gj3} \) from a mixture distribution taking the value zero with probability 0.5 and a random value from a normal distribution with mean 0 and standard deviation 0.018 with probability 0.5. With 25 genetic variants, in each simulated dataset there will be an average of 12.5 interactions between a genetic variant and the covariate.

7. Association between a genetic variant and a confounder: invalid instruments

In practice, it may be that some of the genetic variants are not specifically associated with the risk factor of interest, but instead with another variable which is a confounder in the association between the risk factor and outcome. Although they will be correlated with the factor of interest, this will be due to the effect of the confounder rather than a direct effect of the variant on the risk factor.
Unlike the previous departures from the data-generating model, which represent misspecification of the analysis model, in this case the departure is a violation of the instrumental variable assumptions. If the confounder is unmeasured, it will be impossible empirically to distinguish between this scenario and the initial scenario.

We modify the data-generating model (1) by replacing the first line with:

\[
X_i = \sum_{j=1}^{J} \alpha_G Z_j G_{ij} + \alpha_U U_i + \epsilon_i
\]

\[
U_i \sim N\left(\sum_{j=1}^{J} \alpha_G (1 - Z_j) G_{ij}, 1\right).
\]

The \(Z_j\) are dummy variables taking the value one if the genetic variant \(j\) is directly associated with the risk factor \(X\) (a valid instrument) and zero if the variant is associated with the confounder \(U\) (an invalid instrument). The strength of association between the variant and either \(X\) or \(U\) is constrained to be the same. We draw the \(Z_j\) randomly, taking the probability of the instrument being valid as 0.9, 0.7 and 0.5.

**Results**

Table 2 gives results for each of the seven scenarios described above for three values of the causal effect (\(\beta_X = 0, 0.2, 0.4\)) for data-generating models with 25 genetic variants. Results for 9 and 100 genetic variants are given in the Web Appendix. In each case, we present the median estimates across simulations, and the coverage; the IQR of estimates for a null effect and power for a non-null effect are also shown.

We see that:

1–2. For variants with different sizes of effect, the use of true weights rather than an unweighted allele score gave some improvement in power. When the alleles have similar sizes of effect (scenario 1), the gain in power was generally only 3–4%, whereas when the alleles had considerably different sizes of effect (scenario 2), the gain was 12–15%. Results using an unweighted allele score were unbiased even though the model was misspecified. Use of weights derived from the same data under analysis resulted in severe bias. Use of precisely estimated externally derived weights was as efficient as use of the true weights, although power dropped off when the weights were less precisely estimated even, in some cases, to below that of the unweighted score. The composite method results indicate a small amount of bias consistent with weak instrument bias. Nominal coverage levels are maintained, while the apparent power is slightly greater than when the true weights are used, possibly due to the slight upward bias in estimates.

3. The use of variants chosen based on their strength of association with the risk factor in the data under analysis gave seriously biased estimates, with bias in the direction of the confounded association. This bias echoes a similar bias in the multiple variant case where variants are chosen according using data-derived criteria [21].
4-6. None of the ways of misspecifying the analysis model considered (non-linear genetic effects, variant–variant and variant–environment interactions) affected the bias, coverage or power of estimates using the unweighted allele score.

7. The use of invalid genetic variants in an allele score severely biased estimates of causal effects, even when 90% of the variants in the score were valid instruments.

To summarize, the use of an allele score does not seem to be sensitive to implicit parametric assumptions made by the procedure, such as the linearity of effects. However, estimates are sensitive to how the score is constructed, both how the variants included in the score are chosen, and how the weights in a weighted score are determined. In practice, when an allele score is proposed to be used in a Mendelian randomization analysis, researchers should make clear precisely how the decisions leading to the construction of the score were made.

**Discussion**

The overall conclusion from this simulation study is that unweighted allele scores can be used as instruments in Mendelian randomization if each of the variants used in constructing the allele score satisfies the assumptions of an instrumental variable. The validity of the unweighted allele score is not biased by misspecifications to the analysis model, such as the assumption of equal effect sizes for variants, non-linear genetic effects, or effect modification by variant–variant or variant–environment interactions. This is important because, in practice, what assumptions are true is unknown. If variants have different sizes of effect on the risk factor, then precision can be gained by using a weighted allele score, although the use of an unweighted score gave reasonable estimates in the examples considered. If variants have considerably different sizes of effect, then a weighted allele score would be thought to be advisable, although the weights should not be generated from the data under analysis [13]. If the weights are imprecisely measured, then estimates remain unbiased, although gains in power are somewhat reduced.

The use of an allele score enables robust instrumental variable analysis with much larger numbers of genetic variants than conventional methods (2SLS, LIML) can handle. Although LIML performed reasonably well with a small number (9) of variants, the coverage of the LIML estimate was low with large numbers of variants. For variants with different sizes of effect on the risk factor, LIML did not dominate a weighted allele score method in terms of precision. Estimates from the 2SLS method showed bias and poor coverage throughout, a manifestation of the problems of weak instrument bias.

In order to make comparisons across simulations with different numbers of variants, we have assumed that the effect size is smaller when there are more variants in the allele score. In practice, there is no tradeoff that the effect size decreases as the number of instruments increases. The choice as to how many (and which) variants to include in an allele score should be a question addressed using scientific knowledge, rather than statistical testing. Unless variants are very highly correlated, all variants which
<table>
<thead>
<tr>
<th></th>
<th>Null effect ($\beta_X = 0$)</th>
<th>Small effect ($\beta_X = 0.2$)</th>
<th>Moderate effect ($\beta_X = 0.4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median</td>
<td>IQR</td>
<td>Coverage</td>
</tr>
<tr>
<td>1. Unequal effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unweighted score</td>
<td>0.00</td>
<td>0.18</td>
<td>96.7</td>
</tr>
<tr>
<td>Internal weights (2SLS)</td>
<td>0.14</td>
<td>0.13</td>
<td>71.7</td>
</tr>
<tr>
<td>External weights (imprecise)</td>
<td>0.00</td>
<td>0.17</td>
<td>95.7</td>
</tr>
<tr>
<td>True weights</td>
<td>0.00</td>
<td>0.17</td>
<td>96.3</td>
</tr>
<tr>
<td>LIML</td>
<td>0.00</td>
<td>0.20</td>
<td>92.4</td>
</tr>
<tr>
<td>2. Main and secondary effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unweighted score</td>
<td>0.00</td>
<td>0.18</td>
<td>96.9</td>
</tr>
<tr>
<td>Internal weights (2SLS)</td>
<td>0.10</td>
<td>0.12</td>
<td>76.6</td>
</tr>
<tr>
<td>External weights (imprecise)</td>
<td>0.00</td>
<td>0.15</td>
<td>95.3</td>
</tr>
<tr>
<td>External weights (precise)</td>
<td>0.00</td>
<td>0.14</td>
<td>95.9</td>
</tr>
<tr>
<td>True weights</td>
<td>0.00</td>
<td>0.14</td>
<td>95.7</td>
</tr>
<tr>
<td>Composite approach</td>
<td>0.01</td>
<td>0.14</td>
<td>95.2</td>
</tr>
<tr>
<td>LIML</td>
<td>0.00</td>
<td>0.16</td>
<td>93.3</td>
</tr>
<tr>
<td>3. Selected variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 5 variants</td>
<td>0.23</td>
<td>0.12</td>
<td>62.4</td>
</tr>
<tr>
<td>Top 10 variants</td>
<td>0.19</td>
<td>0.10</td>
<td>64.5</td>
</tr>
<tr>
<td>Variants with $p &lt; 0.05$</td>
<td>0.21</td>
<td>0.10</td>
<td>61.8</td>
</tr>
<tr>
<td>Variants with $p &lt; 0.01$</td>
<td>0.26</td>
<td>0.10</td>
<td>59.1</td>
</tr>
<tr>
<td>4. Non-linear effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00</td>
<td>0.18</td>
<td>96.8</td>
</tr>
<tr>
<td>5. Interactions between variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00</td>
<td>0.18</td>
<td>97.0</td>
</tr>
<tr>
<td>6. Interactions between a variant and covariate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00</td>
<td>0.18</td>
<td>96.9</td>
</tr>
<tr>
<td>7. Invalid variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% valid variants</td>
<td>0.10</td>
<td>0.19</td>
<td>83.3</td>
</tr>
<tr>
<td>70% valid variants</td>
<td>0.30</td>
<td>0.21</td>
<td>35.8</td>
</tr>
<tr>
<td>50% valid variants</td>
<td>0.49</td>
<td>0.20</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Table 2: Instrumental variable estimates in a range of scenarios from allele score analysis and multivariable analyses using two-stage least squares (2SLS) and limited information maximum likelihood (LIML) methods in data-generating model with 25 genetic variants: median estimate across simulations, interquartile range (IQR) of estimates, coverage and power (%)  

1Use of a weighted allele score with internal weights (weights derived from the data under analysis) is equivalent to use of the 2SLS method.
can be reasonably assumed to be valid instruments should be included in a Mendelian randomization analysis to improve the precision of the causal estimate [14].

The use of large numbers of genetic variants associated with a risk factor has been proposed in Mendelian randomization, on the premise that pleiotropic effects may be expected to “balance out” [27]. This is similar to the expecting the effects of confounding on observational estimates of association to cancel out. The results of this paper demonstrate that the criteria for the inclusion of a genetic variant in an allele score should be just as stringent as those for any other Mendelian randomization analysis.

Comparison with previous work

Previous work on the use of multiple genetic variants in the context of Mendelian randomization has demonstrated that using an allele score results in increases in power compared to using single genetic variants, with slight reductions in power compared to using multiple variants, but better bias properties [13, 14]. This paper confirms these findings and further emphasizes the problem of poor coverage with large numbers of variants. The major contributions of this paper are the comparison of weighted and unweighted scores, and addressing the robustness of estimates using allele scores to misspecification of the score. A key novel finding of this paper affecting the use of allele scores in practice is that the procedure used for constructing an allele score, or for deriving weights for a weighted score, has a considerable impact on the bias of estimates.

Limitations of this paper

Although the simulations have covered a range of different scenarios, the conclusions are limited by the reliance on simulated rather than theoretical results. Other departures could be investigated in further simulations. For example, we have here considered genetic interactions on a linear scale; interactions could be considered on a multiplicative scale. We have limited this paper to the case of a continuous outcome. Although the outcome in Mendelian randomization is often binary, binary outcomes result in other difficulties in effect estimation [28, 29]. However, we have no reason to doubt that the general findings of this paper would be applicable to the binary case.

We have here assumed that the external weights used in calculating a weighted allele score are relevant estimates of the true weights. If the external source is from a different population, then these weights may be biased for the true weights. As the use of an unweighted score, which is known to be misspecified with variants of different strengths, did not result in bias, it is unlikely that the use of mis-estimated weights would lead to serious bias. However, when choosing a source to derive external weights, it is best to choose a source from a similar population with enough participants to ensure precisely estimated relevant weights.

One assumption which we have not varied is the independence of genetic variants. If several variants are included in an allele score which are in high linkage disequilibrium (highly correlated), then it would be unnecessary to include all the variants in an
allele score, especially if they all happened to be correlated with the same functional variant. This would lead to difficulties in estimating and interpreting weights in a weighted allele score.

A disadvantage of using multiple genetic variants, and allele scores in particular, is sporadic missing data leading to reduced sample sizes for analysis [14]. In the multiple variant setting, imputation methods have been shown to be effective in mitigating against any reduction in power due to missing data [30].

Key messages:

- The use of an allele score rather than multiple genetic variants resolves the problem of weak instrument bias in Mendelian randomization.

- If genetic variants have approximately equal-sized effects, then an unweighted allele score gives unbiased estimates with coverage close to the nominal level, which are robust to misspecifications of the assumptions of linearity and additivity made by the allele score.

- If genetic variants have considerably different-sized effects, then a weighted allele score gives more efficient estimates. These weights should be derived from an independent dataset; use of the dataset under analysis to derive weights leads to severe bias.

- Choosing genetic variants to include in an allele score based on observed strength in the dataset under consideration also leads to severe bias. More generally, the procedure for constructing an allele score to be used in an analysis should be made clear, as this has a considerable impact on bias.

- Inclusion of variants in an allele score which are invalid instruments results in severe bias. (165 words)

Acknowledgements

The authors would like to thank colleagues at the Cardiovascular Epidemiology Unit, University of Cambridge for helpful discussions in the formation of this paper: Jo Howson, Adam Butterworth, Stephen Kaptoge, Pei Gao, and Edwin Grappin.

References


Web Appendix

Additional parameters and results

We briefly list the parameters used in the simulations in each of the ten scenarios considered with 9, 25 and 100 genetic variants.

1. **Unequal variants:** Genetic effect sizes $\alpha_{Gj}$ for each genetic variant $j$ from a normal distribution with mean (0.1, 0.06, 0.03) and standard deviation (0.03, 0.018, 0.009). Independent weights are generated from a normal distribution with standard deviation of 0.04 and 0.01.

2. **Main and secondary variants:** Genetic effect sizes $\alpha_G = (0.054, 0.046, 0.0285)$ for secondary variants and $5\alpha_G = (0.27, 0.23, 0.1425)$ for two main variants.

3. **Selected variants:** The fixed numbers of variants were (2, 5, 10) and (4, 10, 20).

4. **Non-linear genetic effects:** Genetic effect sizes $\alpha_G = (0.1, 0.06, 0.03)$, effects $\alpha_{Gj1}$ drawn from a normal distribution with mean 0 and standard deviation (0.06, 0.036, 0.018).

5. **Interactions between genetic variants:** Genetic effect sizes $\alpha_G = (0.1, 0.06, 0.03)$, effects $\alpha_{Gjk2}$ drawn from a mixture distribution taking the value zero with probability (0.8, 0.95, 0.98) and a random value from a normal distribution with mean 0 and standard deviation (0.06, 0.036, 0.018) with probability (0.2, 0.05, 0.02). With (9, 25, 100) genetic variants, in each simulated dataset there will be an average of (7.2, 15, 90) interactions between genetic variants out of the (45, 300, 4500) pairs of variants.

6. **Interactions between a genetic variant and a covariate:** Genetic effect sizes $\alpha_G = (0.1, 0.06, 0.03)$, effects $\alpha_{Gj3}$ drawn from a mixture distribution taking the value zero with probability 0.5 and a random value from a normal distribution with mean 0 and standard deviation (0.03, 0.018, 0.009) with probability 0.5. With (9, 25, 100) genetic variants, in each simulated dataset there will be an average of (4.5, 12.5, 50) interactions between a genetic variant and the covariate.

7. **Invalid variants:** Parameters as in initial analysis.

The results with 9 and 100 genetic variants are given in Web Tables A1 and A2 respectively.
<table>
<thead>
<tr>
<th></th>
<th>Null effect ($\beta_X = 0$)</th>
<th>Small effect ($\beta_X = 0.2$)</th>
<th>Moderate effect ($\beta_X = 0.4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median</td>
<td>IQR</td>
<td>Coverage</td>
</tr>
<tr>
<td>1. Unequal effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unweighted score</td>
<td>0.00</td>
<td>0.19</td>
<td>95.2</td>
</tr>
<tr>
<td>Internal weights (2SLS)</td>
<td>0.06</td>
<td>0.16</td>
<td>90.7</td>
</tr>
<tr>
<td>External weights (imprecise)</td>
<td>0.00</td>
<td>0.19</td>
<td>96.1</td>
</tr>
<tr>
<td>External weights (precise)</td>
<td>0.00</td>
<td>0.18</td>
<td>95.9</td>
</tr>
<tr>
<td>True weights</td>
<td>0.00</td>
<td>0.18</td>
<td>95.7</td>
</tr>
<tr>
<td>LIML</td>
<td>0.00</td>
<td>0.18</td>
<td>95.3</td>
</tr>
<tr>
<td>2. Main and secondary effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unweighted score</td>
<td>0.00</td>
<td>0.18</td>
<td>95.1</td>
</tr>
<tr>
<td>Internal weights (2SLS)</td>
<td>0.03</td>
<td>0.12</td>
<td>92.8</td>
</tr>
<tr>
<td>External weights (imprecise)</td>
<td>0.00</td>
<td>0.14</td>
<td>96.5</td>
</tr>
<tr>
<td>External weights (precise)</td>
<td>0.00</td>
<td>0.13</td>
<td>95.9</td>
</tr>
<tr>
<td>True weights</td>
<td>0.00</td>
<td>0.14</td>
<td>95.9</td>
</tr>
<tr>
<td>Composite approach</td>
<td>0.01</td>
<td>0.13</td>
<td>95.5</td>
</tr>
<tr>
<td>LIML</td>
<td>0.00</td>
<td>0.14</td>
<td>95.4</td>
</tr>
<tr>
<td>3. Selected variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 2 variants</td>
<td>0.17</td>
<td>0.15</td>
<td>80.5</td>
</tr>
<tr>
<td>Top 4 variants</td>
<td>0.13</td>
<td>0.12</td>
<td>84.0</td>
</tr>
<tr>
<td>Variants with $p &lt; 0.05$</td>
<td>0.08</td>
<td>0.10</td>
<td>88.5</td>
</tr>
<tr>
<td>Variants with $p &lt; 0.01$</td>
<td>0.13</td>
<td>0.11</td>
<td>83.3</td>
</tr>
<tr>
<td>4. Non-linear effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00</td>
<td>0.19</td>
<td>95.1</td>
</tr>
<tr>
<td>5. Interactions between variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00</td>
<td>0.19</td>
<td>95.4</td>
</tr>
<tr>
<td>6. Interactions between a variant and covariate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allele score</td>
<td>0.00</td>
<td>0.19</td>
<td>95.3</td>
</tr>
<tr>
<td>7. Invalid variants</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90% valid variants</td>
<td>0.09</td>
<td>0.20</td>
<td>80.2</td>
</tr>
<tr>
<td>70% valid variants</td>
<td>0.30</td>
<td>0.26</td>
<td>36.9</td>
</tr>
<tr>
<td>50% valid variants</td>
<td>0.49</td>
<td>0.27</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Web Table A1: Instrumental variable estimates in a range of scenarios from allele score analysis and multivariable analyses using two-stage least squares (2SLS) and limited information maximum likelihood (LIML) methods in data-generating model with 9 genetic variants: median estimate across simulations, interquartile range (IQR) of estimates, coverage and power (%)

1Use of a weighted allele score with internal weights (i.e. weights derived from the data under analysis) is equivalent to use of the 2SLS method.
Web Table A2: Instrumental variable estimates in a range of scenarios from allele score analysis and multivariable analyses using two-stage least squares (2SLS) and limited information maximum likelihood (LIML) methods in data-generating model with 100 genetic variants: median estimate across simulations, interquartile range (IQR) of estimates, coverage and power (%).

1Use of a weighted allele score with internal weights (i.e. weights derived from the data under analysis) is equivalent to use of the 2SLS method.