Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases

Marie Verbanck1,2,7, Chia-Yen Chen4,5,6,7, Benjamin Neale4,5,6,8* and Ron Do1,2,3,8*

Horizontal pleiotropy occurs when the variant has an effect on disease outside of its effect on the exposure in Mendelian randomization (MR). Violation of the 'no horizontal pleiotropy' assumption can cause severe bias in MR. We developed the Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO) test to identify horizontal pleiotropic outliers in multi-instrument summary-level MR testing. We showed using simulations that the MR-PRESSO test is best suited when horizontal pleiotropy occurs in <50% of instruments. Next we applied the MR-PRESSO test, along with several other MR tests, to complex traits and diseases and found that horizontal pleiotropy (i) was detectable in over 48% of significant causal relationships in MR; (ii) introduced distortions in the causal estimates in MR that ranged on average from ~131% to 201%; (iii) induced false-positive causal relationships in up to 10% of relationships; and (iv) could be corrected in some but not all instances.

Epidemiological studies have established correlations between numerous exposures and complex diseases1. Drawing causal inferences from such studies can be challenging due to reverse causation and confounding and/o other biases2.

Mendelian randomization (MR) is a commonly used human genetics approach that can be used to infer causality of an exposure for a complex disease outcome3-4. MR presents a number of advantages over observational epidemiology, including the ability to control for non-heritable environmental confounders in such analyses and the use of genetic instruments to evaluate the impact of an exposure without necessitating the measurement of that exposure in the outcome group. MR uses genetic variants as instrumental variables (IVs) that are robustly associated with the exposure of interest and tests whether the effects of the variants on the exposure result in proportional effects on the outcome.

In response to the advent of the genome-wide association (GWA) study and subsequent identification of thousands of trait-associated loci, multiple MR methods that leverage GWA summary statistics have been developed. These multi-instrument MR methods aggregate estimates from multiple IVs, testing for a causal relationship between a given exposure and outcome in a linear regression framework in which the variants’ effects on the outcome are regressed on the same variants’ effects on the exposure2.

A fundamental assumption of MR is the ‘no horizontal pleiotropy’ assumption (also called the ‘exclusion-restriction criterion’), which requires that the IV used for MR analysis acts on the target outcome exclusively through the exposure of interest5-6. Horizontal pleiotropy occurs when the variant has an effect on other traits outside of the pathway of the exposure of interest and has an impact on the target outcome, or when the variant has a direct effect on the target outcome7. As a violation of the ‘no horizontal pleiotropy’ assumption, horizontal pleiotropy can distort MR tests, leading to inaccurate causal estimates, loss of statistical power and potential false-positive causal relationships.

Emerging evidence has supported the possibility of a pervasive role of pleiotropy among loci identified from GWA studies. Studies have shown that many traits are genetically correlated with each other8. Furthermore, studies have shown that hundreds of individual variants identified from GWA studies are associated with multiple traits9-14. Studies in causal relationships inferred by MR is currently unknown.

As a result, there has recently been discussion regarding the potentially serious consequences horizontal pleiotropy may have on the validity of previous and current MR studies. Some have raised skepticism about the MR approach due to the pervasiveness of pleiotropy among trait-associated variants15-16, while others have defended MR by noting that horizontal pleiotropy has long been known to impose limits on MR16. Regardless, the fact remains that the extent to which horizontal pleiotropy affects causal relationships inferred by MR is currently unknown.

Here we conduct a systematic evaluation of the role of horizontal pleiotropy in MR. We developed the Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO) approach to detect and correct for horizontal pleiotropic outliers in multi-instrument summary-level MR testing. In extensive simulations, we then evaluated the performance of the MR-PRESSO test and compared it with other complementary methods, including methods that measure and correct for an average horizontal pleiotropic effect across all variants, as well as outlier-robust methods. Finally, we applied these methods to 4,250 MR tests of complex traits and diseases derived from 82 summary-level GWA datasets.

Results
The MR-PRESSO test and its components. We developed the MR-PRESSO test to evaluate horizontal pleiotropy in multi-instrument

1The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 2The Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 3Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. 4Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. 5Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA. 6Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA. 7These authors contributed equally: Marie Verbanck and Chia-Yen Chen. 8These authors jointly supervised this work: Benjamin Neale and Ron Do.
*e-mail: bneale@broadinstitute.org; ron.do@mssm.edu
summary-level MR. In brief, MR-PRESSO has the following three components (Methods and Fig. 1): (a) detection of horizontal pleiotropy (the MR-PRESSO global test); (b) correction for horizontal pleiotropy via outlier removal (the MR-PRESSO outlier test); and (c) testing of significant differences in the causal estimates before and after correction for outliers (the MR-PRESSO distortion test). The MR-PRESSO test relies on a regression framework in which the variants’ effects on the outcome are regressed on the same variants’ effects on exposure, with the slope of the regression line providing an estimate of the causal effect of the exposure on the outcome. The MR-PRESSO global test evaluates overall horizontal pleiotropy among all IVs in a single MR test by comparing the observed distance of all the variants to the regression line (residual sum of squares) with the expected distance under the null hypothesis of no horizontal pleiotropy. The MR-PRESSO outlier test evaluates the presence of specific horizontal pleiotropic outlier variants by using the observed and expected distributions of the tested variant. Finally, the MR-PRESSO distortion test evaluates the significance of the distortion between the causal estimate before removal of the horizontal pleiotropic outlier variants and the causal estimate after such removal (detected from the outlier test of MR-PRESSO).

Comparison of statistical properties of MR methods. We assessed the statistical performances of several methods designed either to detect horizontal pleiotropy (the MR-PRESSO global test, Q test\(^{17,18}\), Q (modified) test\(^{19}\), Q’ test\(^{17,18,20}\) and Q’ (modified) test\(^{19}\)) or to correct for horizontal pleiotropy (correction of the global average horizontal pleiotropic effect using MR-Egger regression\(^{20,21}\) or multi-variate MR (MMR)\(^{22,23}\); outlier detection and removal approaches using the MR-PRESSO outlier test, Cook’s distance\(^{24,25}\), Studentized residuals\(^{24}\), the Q (modified) outlier test or Q’ (modified) outlier test in multi-instrument summary-level MR. For these comparisons, we performed 10,000 simulations under different scenarios using the model described in Supplementary Fig. 1b. We varied several parameters in the simulations, including the main causal effect (\(\beta_{\text{main}} = 0, 0.1, 0.2, 0.5\) with \(\beta_{\text{pleiotropy}} = 0.1\)), the percentage of horizontal pleiotropic outlier variants (0%, 2%, 4%, 10%, 50% or 90%), the type of horizontal pleiotropy (positive or balanced) and whether or not we assumed the ‘instrument strength independent of direct effect’ (InSIDE) condition\(^{20}\) (Methods and Supplementary Note 1).

False-positive rate and/or power to detect horizontal pleiotropy. Using simulations, we first assessed the false-positive rate (type 1 error) and power (1 – type 2 error) of the MR methods noted above that detect horizontal pleiotropy, including the MR-PRESSO global test, Q test\(^{17,18}\), Q (modified) test\(^{19}\), Q’ test\(^{17,18,20}\) and Q’ (modified) test\(^{19}\) (Methods and Supplementary Note 1).

Under the null hypothesis of no horizontal pleiotropy (Supplementary Table 1), the MR-PRESSO global test, Q (modified) test and Q’ (modified) test had controlled false-positive rates (~5%), whereas the rates of the original Q and Q’ tests were inflated (between 5% and 25%). Since the Q and Q’ tests were found to be inflated, we proceeded with the MR-PRESSO global test, Q (modified) test and Q’ (modified) test in power analyses (Table 1). We observed acceptable power (> 70%) to detect horizontal pleiotropy across all three tests in simulations in which the percentage of horizontal pleiotropic variants was ≥10% (which corresponds to 5 horizontal pleiotropic variants out of 50). Similar results were observed across all magnitudes of causal estimates, including instances in which there was no causal relationship (\(\beta_{\text{main}} = 0\) ) (Table 1). In addition, we conducted several sensitivity analyses (violation of the InSIDE condition, frequency of horizontal pleiotropic variants ≥50% and perfectly overlapping samples to estimate the effect sizes on the outcome and exposure) to further evaluate the robustness of these methods. All three tests had high power in the detection of horizontal pleiotropy under a wide range of these...
horizontal pleiotropy). The false-positive rate (family-wise error rate) was lower than did the MR-PRESSO outlier test (inflated false-positive rate (family-wise error rate), whereas Studentized residuals had low power (Supplementary Table 2). The Q (modified) outlier test require, as a baseline assumption, that at least 50% of the genetic variants be valid instruments (no horizontal pleiotropy) and have balanced pleiotropy and that the InSIDE assumption be valid. Conversely, methods that correct for a global average horizontal pleiotropic effect among all variants (either by including the regression intercept in MR-Egger or by covariate adjustment in MMR) are best suited when there is a large percentage of horizontal pleiotropic variants (>50%).

Detection of horizontal pleiotropy in MR for complex traits. We applied the MR-PRESSO global test, along with two methods (the Q (modified) test and Q’ (modified) test) that can detect horizontal pleiotropy in MR, to all possible pairs of 82 complex traits and diseases retrieved from publicly available GWA datasets (Table 2 and Methods). In total, we conducted 4,250 tests for each of the three MR approaches. We accounted for multiple testing of the 4,250 tests using the Bonferroni correction. We note that this correction is overly stringent, since many of the traits and diseases are correlated. Using a Bonferroni-corrected threshold of $P < 1.17 \times 10^{-5}$, the MR-PRESSO global test was statistically significant in 21.69% ($n = 922$) of the 4,250 tests. When we restricted this to statistically significant causal estimates in the IVW meta-analysis, we detected significance of the MR-PRESSO global test at a higher rate of 48.69% ($n = 93$ of 191 tests). Both the Q (modified) test and the Q’ (modified) test provided similar estimates. As a sensitivity analysis, we restricted this to a subset of traits and diseases that were less correlated (for example, Pearson $r < 0.30$). We observed that 24.17% of MR tests were statistically significant for the MR-PRESSO global test among significant causal relationships (Supplementary Table 10).

Horizontal pleiotropy correction in MR for complex traits. We evaluated five methods to correct for horizontal pleiotropy in MR (Table 3). This included the following: (1) removing outliers detected by the MR-PRESSO outlier test, Q (modified) outlier test

---

**Table 1 | Power to detect horizontal pleiotropy in MR, for different methods**

| Causal effect | Horizontal pleiotropic variants (%) | Type of pleiotropy | MR-PRESSO global test | Q (modified) test | Q’ (modified) test |
|---------------|-------------------------------------|--------------------|-----------------------|-----------------|-------------------|
| 0 2           | Balanced                            | 25.34              | 25.40                 | 22.04           |
| 0 2           | Positive                            | 25.01              | 25.01                 | 22.00           |
| 0 4           | Balanced                            | 51.79              | 51.96                 | 47.80           |
| 0 4           | Positive                            | 50.88              | 51.32                 | 47.34           |
| 0 10          | Balanced                            | 95.53              | 95.58                 | 94.47           |
| 0 10          | Positive                            | 94.27              | 94.26                 | 92.85           |
| 0.1 2         | Balanced                            | 24.10              | 24.41                 | 21.58           |
| 0.1 2         | Positive                            | 23.60              | 23.89                 | 20.96           |
| 0.1 4         | Balanced                            | 51.17              | 51.49                 | 47.51           |
| 0.1 4         | Positive                            | 50.67              | 50.59                 | 46.69           |
| 0.1 10        | Balanced                            | 95.56              | 95.56                 | 94.31           |
| 0.1 10        | Positive                            | 93.73              | 93.77                 | 92.08           |
| 0.2 2         | Balanced                            | 22.42              | 22.70                 | 19.74           |
| 0.2 2         | Positive                            | 22.95              | 22.92                 | 19.97           |
| 0.2 4         | Balanced                            | 48.37              | 48.29                 | 44.45           |
| 0.2 4         | Positive                            | 46.89              | 46.82                 | 43.18           |
| 0.2 10        | Balanced                            | 94.08              | 94.11                 | 92.78           |
| 0.2 10        | Positive                            | 91.85              | 91.91                 | 90.24           |
| 0.5 2         | Balanced                            | 16.72              | 16.70                 | 14.90           |
| 0.5 2         | Positive                            | 16.76              | 16.55                 | 15.15           |
| 0.5 4         | Balanced                            | 33.71              | 33.79                 | 31.00           |
| 0.5 4         | Positive                            | 32.52              | 32.66                 | 29.93           |
| 0.5 10        | Balanced                            | 81.15              | 81.22                 | 79.07           |
| 0.5 10        | Positive                            | 76.99              | 77.09                 | 74.75           |

The simulation scenarios included variations of the causal effect of exposure 1 on the outcome, the percentage of horizontal pleiotropic variants among the total of 50 variants and the type of pleiotropy. The InSIDE condition was satisfied in all reported scenarios.

parameters; however, simulations showed a reduction in power in those with perfectly overlapping samples (Supplementary Note 1 and Supplementary Tables 2–4).

**Evaluation of bias in the causal estimates.** We investigated how the causal effect estimate (bias) and corresponding s.d. (precision) of the MR-PRESSO outlier test were affected by horizontal pleiotropy. We then compared the MR-PRESSO test with other established MR methods that can correct for horizontal pleiotropy, including those that correct for a global average horizontal pleiotropic effect across all variants (for example, inclusion of the intercept in MR-Egger regression [10] and adjustment for multiple exposures in MMR [22]), horizontal pleiotropic outlier-detection methods (Cook’s distance [13] and Studentized residuals [14], the Q (modified) outlier test and the Q’ (modified) outlier test) and outlier-robust methods (weighted median [26] and mode-based estimate [27]) (Methods).

The MR-PRESSO outlier test and Cook’s distance had similar power in identifying the correct horizontal pleiotropic outliers, whereas Studentized residuals had low power (Supplementary Table 5 and Supplementary Note 1). However, Cook’s distance had lower specificity in identifying the correct horizontal pleiotropic outliers than did the MR-PRESSO outlier test (inflated false-positive rate (family-wise error rate), $>97\%$ when there was no horizontal pleiotropy). The false-positive rate (family-wise error rate) and power of the Q (modified) outlier test and the Q’ (modified) outlier test were very similar to those of the MR-PRESSO outlier test (Supplementary Table 5).

MR-Egger regression generally had lower precision than the MR-PRESSO outlier test and MMR. The $I^2$ index, which informs on when MR-Egger regression should be employed (i.e., $I^2 > 90\%$), was lower than 90% on average in most settings (Supplementary Table 6 and Supplementary Note 1).

Cook’s distance and Studentized residuals had bias and precision in the causal estimate similar to that of the MR-PRESSO outlier test (Supplementary Table 7). The weighted median had less bias but also less precision in the causal estimate than did the MR-PRESSO outlier test, particularly when the frequency of horizontal pleiotropic variants was $<50\%$. The mode-based estimate generally had very low precision compared with that of the other methods.

All four methods (mode-based estimate, weighted median, Cook’s distance and Studentized residuals) had limitations similar to those of the MR-PRESSO outlier test (for example, inflated causal estimates and low precision) when there was a very high percentage of horizontal pleiotropic variants (≥50%). The weighted median had less bias but lower specificity in identifying the correct horizontal pleiotropic outliers, whereas Studentized residuals had low power (Supplementary Table 6 and Supplementary Note 1). Our simulations showed that when the InSIDE assumption was invalid, all methods exhibited bias in the causal estimate (Supplementary Tables 8 and 9 and Supplementary Note 1).

The standard inverse-variance weighted (IVW) approach showed the expected bias in the causal estimate due to horizontal pleiotropy. When the InSIDE assumption was valid and the percentage of horizontal pleiotropic variants was small ($\leq 10\%$), the causal estimate of the MR-PRESSO outlier adjustment was less biased and had better precision (smaller s.d.) than IVW, MR-Egger or MMR. However, when the percentage of horizontal pleiotropic variants was high ($\geq 50\%$), the opposite was found. These trends were expected, since outlier-detection methods such as the MR-PRESSO outlier test require, as a baseline assumption, that at least 50% of the genetic variants be valid instruments (no horizontal pleiotropy) and have balanced pleiotropy and that the InSIDE assumption be valid. Conversely, methods that correct for a global average horizontal pleiotropic effect among all variants (either by including the regression intercept in MR-Egger or by covariate adjustment in MMR) are best suited when there is a large percentage of horizontal pleiotropic variants ($>50\%$).
and Q' (modified) outlier test (at the Bonferroni-corrected threshold); and (2) adjusting for significant covariates individually or all together from the main MR test (for example, significant causal effect in IVW meta-analysis at the Bonferroni-corrected threshold).

As shown in the simulations, the MR-PRESSO global test could be used to determine if there were any remaining horizontal pleiotropy after the application of correction strategies to minimize initial horizontal pleiotropy in the MR test (Supplementary Note 1 and Supplementary Table 11). In Table 3, we observed that the outlier-removal approach using the MR-PRESSO outlier test was effective in eliminating statistical significance in the MR-PRESSO global test in 46% of the 922 tests. The Q (modified) outlier test and Q' (modified) outlier test provided similar estimates. Furthermore, the covariate-adjustment approach, defined by accounting for traits that were shown to have a significant causal effect on the same outcome, eliminated significance in the MR-PRESSO global test in 22% (n=20) of the 93 tests when adjusting for a single covariate. When adjusting for all significant covariates in the same model, the covariate adjustment approach eliminated significance in 34% (n=22) of the 42 tests. Taken together, these two correction strategies (MR-PRESSO outlier removal and MMR) were successful in 47% (n=438) of the 922 total tests. Furthermore, we note that the covariate-adjustment approach is limited in that it requires a priori knowledge of the trait responsible for the horizontal pleiotropic effect.

Outliers effect on the distortion of MR causal estimates. We evaluated the extent to which outliers cause distortion in the causal estimates resulting from MR. Using the MR-PRESSO distortion test, we compared the causal estimates from the IVW meta-analysis before removal of outlier variants detected by the MR-PRESSO outlier test with those after such removal (Methods). Using a Bonferroni-corrected threshold, we observed a significant distortion (of ~93% and 35%) in 2.5% (n=2) of significant causal estimates (n=81 total). Since the Bonferroni correction is overly stringent, we considered the commonly used nominal threshold of P<0.05 that the majority of MR studies to date have used for statistical significance. A significant distortion was observed in almost 10% (n=22) of the causal relationships (n=229 total), with a distortion between −131% and 201%, on average (Fig. 2).

Below, we provide one example to highlight the role of pleiotropy in MR (Supplementary Fig. 2). We observed that the causal effect of body mass index (BMI) on C-reactive protein was estimated to be 0.39 (P=7.02×10−4) by IVW. The MR-PRESSO global test showed statistical significance (P<10−4), and the MR-PRESSO outlier test identified one significant outlier variant (P<10−4; rs2075650 in the APOE locus, which encodes apolipoprotein E). Examining this further, we observed that this variant was highly pleiotropic, having associations with several traits and diseases, including Alzheimer’s disease, BMI, C-reactive protein, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, plasma triglycerides, waist circumference, hip circumference and waist/hip ratio at P=6×10−4 (Bonferroni-corrected P=0.05 / 82 (where the denominator is the number of traits); Supplementary Table 12). Furthermore, this variant was associated with several other traits and diseases in the public NHGRI-EBI GWAS catalog5 (P<5×10−8; Supplementary Table 13). We believe some of these genetic associations with other traits and diseases will be found to be due to horizontal pleiotropy, while others will be found to be in the same causal pathway (for example, vertical pleiotropy). After removing this outlier variant, we observed a lower estimation of the causal estimate for BMI on C-reactive protein (βBMI=0.35, P=3.45×10−5), with this single variant alone causing a 12% distortion in the causal estimate (Supplementary Fig. 2).

Outliers effect on false-positive causal relations in MR. We evaluated the extent to which outliers (as detected by the MR-PRESSO outlier test) can induce false-positive causal relationships. A false-positive causal relationship was defined as an exposure–outcome pair in which the causal estimate was no longer statistically significant in the outlier-corrected IVW model but was previously significant in the naïve IVW model. According to this definition, we identified false-positive relationships in 10% of putatively causal relationships (n=24 of 229 total tests) using the common nominal threshold of P<0.05 and 1.2% (n=1 out of 81 total tests) using the stringent Bonferroni-corrected threshold (P<1.17×10−4). We note that the outlier-removal approach decreases the power of the outlier-corrected IVW model test because a smaller number of variants is included after removal of horizontal pleiotropic outlier variants. Therefore, we expect that a small number of false-positive causal relationships might be due to reduced statistical power.

Causal relationships inferred from MR testing. We identified 191 significant causal relationships out of a total of 4,250 MR tests using the Bonferroni-corrected threshold of P<0.05 / 4,250 < 1 × 10−5. We note that many of the traits examined are closely related to each other (for example, BMI–waist circumference–hip circumference, and low-density-lipoprotein cholesterol–total cholesterol) and therefore the tests are not completely independent of each other. After
correction for horizontal pleiotropy via outlier removal using the MR-PRESSO outlier test, we validated known causal relationships, including the effect of low-density-lipoprotein cholesterol on coronary artery disease ($\beta_{\text{causal}} = 0.52, P = 5.15 \times 10^{-10}$), systolic blood pressure on coronary artery disease ($\beta_{\text{causal}} = 0.05, P = 1.78 \times 10^{-6}$), and BMI on C-reactive protein ($\beta_{\text{causal}} = 0.35, P = 3.45 \times 10^{-19}$), among the strongest findings. Furthermore, we observed an effect of BMI on uric acid ($\beta_{\text{causal}} = 0.31, P = 3.29 \times 10^{-10}$) and plasma triglycerides ($\beta_{\text{causal}} = 0.20, P = 6.9 \times 10^{-19}$), although these were significant for the MR-PRESSO global test even after correction via outlier removal.

**Discussion**

In summary, we have evaluated horizontal pleiotropy in the context of MR testing across pairwise comparisons of a large number of complex traits and diseases. We have (i) developed the MR-PRESSO method to detect and correct for horizontal pleiotropic outliers in MR and compared it with several established methods; (ii) applied several of these MR methods to complex traits and diseases and showed that horizontal pleiotropy occurs in over 48% of causal relationships between complex traits and diseases inferred by MR; (iii) observed average distortion between −131% and 201% in the causal estimates of MR due to horizontal pleiotropy; and (iv) showed that horizontal pleiotropy can be minimized and corrected in some cases through outlier detection and/or secondary phenotype adjustment when the mediating trait is known.

By applying the MR-PRESSO global test to detect horizontal pleiotropy in a wide array of complex traits and diseases, we observed horizontal pleiotropy in approximately 48% of inferred causal relationships. This is consistent with emerging evidence that many disease-associated variants identified from GWA studies have effects on multiple traits\(^a\). Since these variants are used as IVs in multi-instrument MR, it is likely that a non-negligible number of these variants do not meet the ‘no horizontal pleiotropy’ assumption in MR. Furthermore, Hemani et al.\(^b\) have evaluated horizontal pleiotropy in a study concomitant with our study. Their study proposed a mixture-of-experts machine-learning framework to select the most appropriate MR method among a variety of standard and horizontal pleiotropy-robust MR methods associated with an IV-selection procedure. The framework selected a method that involved horizontal pleiotropy (pleiotropy-robust method or IV filtering) in 90% of the MR tests. These results indicate that horizontal pleiotropy is commonplace and highlight the need for the evaluation of horizontal pleiotropy for variants acting as IVs as a necessary and standard test for the performance of MR.

Horizontal pleiotropy in MR has direct implications for genetics-guided drug discovery and validation. Accurate estimates of causal effects between biomarkers and diseases can inform dose–response curves for drug efficacy and safety\(^\d\). In the present study, we show that horizontal pleiotropy can induce distortion in the causal estimates in MR and that this distortion is pervasive among many causal relationships. Second, there is increasing interest in using surrogate endpoints for drugs in clinical trials. Identifying true causal relationships using MR can pinpoint biomarkers that are causal and hence identify those surrogate endpoints that are most relevant to disease\(^\d\).

Our current study has several strengths. Outlier-detection methods are useful because they work within the framework of IVW. Furthermore, the MR-PRESSO global test (as well as the Q* (modified) test and Q* (modified) test) is adequately powered to detect horizontal pleiotropy among even a small subset of loci. Finally, outlier-detection methods can be used in several different MR tests, including IVW and MMR and even within the framework of MR-Egger regression. The MR-PRESSO method also has limitations. There were instances in which correction strategies (outlier removal or covariate adjustment) could not completely remove horizontal pleiotropy as detected by the MR-PRESSO global test. Possible reasons for this include violation of the InSIDE condition (as shown in the simulations), which is untestable; a percentage of horizontal pleiotropic outlier variants that is not suitable for the particular correction strategy (outlier-detection method with >50% of horizontal pleiotropic variants); no applicable covariate adjustment; or other sources of heterogeneity in the effect sizes other than horizontal pleiotropy, such as gene–gene and gene–environment interactions. Furthermore, several GWA consortia use the same cohorts and study samples; therefore, some GWA summary statistics may have overlapping samples. In our simulations, we evaluated the effect of perfectly overlapping samples on the power of our global test across a range of scenarios. A reduction in power for the detection of horizontal pleiotropy was observed in the model with perfectly overlapping samples. Finally, because MR-PRESSO requires simulations, the processing time to apply the method can be greater than the time required for other methods.

In summary, we have shown through a series of analyses that horizontal pleiotropy is pervasive in MR testing between complex traits and diseases, highlighting the need to employ approaches that minimize horizontal pleiotropy. Rigorous analysis and careful interpretation of causal inference testing in MR is warranted as a result of these observations.

**URLs** The MR-PRESSO software and full set of results can be found at https://github.com/rondolab/MR-PRESSO.

**Methods**

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0099-7.

Received: 26 June 2017; Accepted: 1 March 2018; Published online: 23 April 2018
References

1. Vasan, R. S. Biomarkers of cardiovascular disease: molecular basis and practical considerations. *Circulation* **113**, 2335–2362 (2006).

2. Ebrahim, S. & Davey Smith, G. Mendelian randomization: can genetic epidemiology help to redress the failures of observational epidemiology? *Hum. Genet.* **123**, 15–33 (2008).

3. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).

4. Evans, D. M. & Smith, G. D. Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Ann. Rev. Genomics Hum. Genet.* **16**, 327–350 (2015).

5. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* **28**, 30–42 (2017).

6. Burgess, S. et al. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–552 (2015).

7. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).

8. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

9. Sivakumaran, S. et al. Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).

10. Gerring, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.* **8**, 78 (2016).

11. Parkes, M., Cortes, A., van Heel, D. A. & Brown, M. A. Genetic insights into common and complex diseases. *Eur. J. Epidemiol.* **30**, 2926–2940 (2015).

12. Bowden, J. et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statist. Med.* https://doi.org/10.1002/sim.7221 (2017).

13. Bowden, J. et al. Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NO-ME assumption. *bioRxiv* https://doi.org/10.1101/159442 (2017).

14. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

15. Bowden, J. et al. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int. J. Epidemiol.* **45**, 1961–1974 (2016).

16. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).

17. Corbin, L. et al. Body mass index as a modifiable risk factor for type 2 diabetes: refining and understanding causal estimates using Mendelian randomisation. *Diabetes* **65**, 3002–3007 (2016).

18. Fox, J. & Long, J. S. *Modern Methods of Data Analysis* (Sage Publications, Thousand Oaks, CA, USA, 1990).

19. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

20. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyx102.

21. MacArthur, J. et al. The new NHGRI–EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

22. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbis, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).

23. Warren, H. R. et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.* **49**, 403–415 (2017).

24. Ehrdt, G. B. et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat. Genet.* **48**, 1171–1184 (2016).

25. Liu, C. et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet.* **48**, 1162–1170 (2016).

26. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

27. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* https://doi.org/10.1093/ije/dyx102.

28. MacArthur, J. et al. The new NHGRI–EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

29. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbis, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).

30. Warren, H. R. et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.* **49**, 403–415 (2017).

31. Ehrdt, G. B. et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat. Genet.* **48**, 1171–1184 (2016).

32. Liu, C. et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet.* **48**, 1162–1170 (2016).
Methods

General assumptions of Mendelian randomization. MR relies on genetic variants that are robustly associated with the tested exposure. A variant can be used as an IV through its effect size on the exposure and its proportional effect on the outcome. The causal estimate is an inferred estimate of the predicted response of the outcome caused by a modification of the exposure. Supplementary Fig. 3 illustrates a standard MR framework (Supplementary Note 1). The single-instrument approach can be extended to multiple instruments. Multi-instrument MR can be performed using an inverse-variance-weighted, fixed-effects meta-analysis (IVW meta-analysis). IVW meta-analysis consists of fitting a weighted linear regression with a fixed intercept of 0 between the set of effect sizes on the outcome (either continuous response or dichotomous response) and the effect sizes on the exposure (either continuous predictor or dichotomous predictor), with the inverse of the variance of the effect sizes on the outcome as weights. When the intercept is fixed to 0, the slope of the regression then provides an estimate of the causal effect of the exposure on the outcome. The validity of MR analysis relies on three assumptions (Supplementary Fig. 3): (1) the variant (i.e., IV) is associated with the exposure; (2) the variant is independent of all confounders of the exposure-outcome relationship; and (3) the variant is independent of the outcome conditional on the exposure and all confounders of the exposure-outcome association (i.e., exclusion restriction criterion). Violation of the third assumption, the exclusion restriction criterion, is a direct consequence of horizontal pleiotropy (Supplementary Fig. 1).

Existing methods to detect and correct for horizontal pleiotropy in MR. Several existing methods have been developed to detect horizontal pleiotropy in MR. Methods to detect horizontal pleiotropy include the Q* test and the Q score test, as well as the recently proposed modified versions of these tests', which are traditionally used to identify over dispersion and have been applied in the context of MR'. Several existing MR methods have also been designed to correct for horizontal pleiotropy. Methods that we evaluated in the current study are described below.

Outlier methods. Traditional methods to detect outliers have been applied in the context of MR to detect and remove invalid IVs. We applied Cook’s distance to identify outliers from the list of genetic variants. We also applied Studentized residuals where we performed Student tests on the Studentized residuals and defined outliers using a Bonferroni-corrected P-value threshold (0.05 divided by the number of variants). Finally, the Q (modified) outlier test and the Q* (modified) outlier test use a χ² distribution with one degree of freedom to test each variant separately in association with a Bonferroni-corrected P-value threshold (0.05 divided by the number of variants).

Correction of the average horizontal pleiotropic effect methods. The intercept of MR-Egger regression can correct for the average horizontal pleiotropic effect across all IVs. The 1/(Ω²) index, which encapsulates the collective strength of a set of variants, can be used to inform on the use of MR-Egger regression. Multivariable MR (MVR) takes into account known causal exposures to the outcome (Ei, in Supplementary Fig. 1b). If the IV is subject to horizontal pleiotropy and has a significant effect on not only the tested exposure E1 but also a second exposure E2, then MMR allows adjustment for this second horizontal pleiotropic effect. In practice, MMR is implemented by fitting a weighted linear regression by regressing Ω on both E1 and E2 (with the inverse of the variance of effect sizes on the outcome as weights). The βj are the genetic effects of the variants on the known secondary causal exposure E2. The model can be extended to include multiple exposures. In this study, a fixed intercept of 0 was used in all MMR models.

Outlier-robust methods. Additional methods are naturally robust to horizontal pleiotropic outlier variants in MR. The weighted median estimator provides a consistent estimate of the causal effect when up to 50% of genetic variants are invalid instruments. The mode-based estimate (MBE) method uses the property that valid instruments should provide the largest number of similar individual-instrument causal estimates even if the majority of instruments is invalid.

MR-PRESSO: MR pleiotropy residual sum and outlier. We developed the MR pleiotropy residual sum and outlier (MR-PRESSO) test to detect and correct for horizontal pleiotropic outliers. MR-PRESSO comprises three components (Fig. 1): (a) detection of horizontal pleiotropy (and violation of the exclusion restriction assumption, Supplementary Fig. 1) in MR (global test); (b) correction by removal of offending IVs that are due to horizontal pleiotropy (outlier test); (c) testing of significant differences in the causal estimates before and after outlier removal (distortion test).

MR-PRESSO extends the framework of IVW meta-analysis and is based on the following rationale. IVW meta-analysis consists of fitting a weighted linear regression between the set of effect sizes on the outcome and the effect sizes on the exposure. The slope of the regression provides an estimate of the causal effect of the exposure on the outcome. Under the null hypothesis where horizontal pleiotropy does not exist, all variants are expected to be close to the regression line (i.e. to have small residuals in the regression). However, when a variant is subject to horizontal pleiotropy, the effect size on the outcome can be larger or smaller than the effect size mediated by the exposure in question and therefore the variant can deviate from the true slope of the regression line. MR-PRESSO is designed to detect whether a subset of variants is significantly deviating from the regression line. The MR-PRESSO outlier test requires that at least 50% of the variants are valid instruments and relies on the InSIDE (instrument strength independent of direct effect) condition, which states that the effect sizes of the variants on the exposure should not depend on the horizontal pleiotropic effects on the outcome.

MR-PRESSO global test. The MR-PRESSO global test evaluates for the presence of horizontal pleiotropy and is made up of four steps (Fig. 1a): (1) For each variant j, we removed the variant in question and refit a IVW regression. This allowed us to calculate the slope of the regression line on the remaining variants, denoted β̂−j, which represents the causal estimate without variant j. (2) The estimated causal effect (slope) without variant j was used to predict the expected effect size on the outcome as the product of β̂−j and the effect size of the same variant on the exposure y. Then we calculated the observed residual sum of squares (RSS) as the difference between the observed effect size of the variant on the outcome (ŷ) and the predicted effect size of the same variant on the outcome RSSobs(j) = (ŷ - β̂−j y). The global observed RSS was then obtained by summing over j RSSobs(j): RSSobs = ∑ RSSobs(j) = RSSexp(j)RSSR SS ()RSS () (obtained in step 2 of the global test) with the expected effect size on the outcome as the product of β̂−j and the trend test. (3) Given the observation that valid instruments should provide the largest number of similar individual-instrument causal estimates even if the majority of instruments is invalid, we developed the MR-PRESSO global test. The MR-PRESSO distortion test quantifies the distortion in the causal estimate due to significant horizontal pleiotropic outlier.
variants (Fig. 1c). Distortion ($D$) is defined as the percentage of the causal estimate that is due to significant horizontal pleiotropic outlier variants. It is calculated as

$$D = 100 \times \frac{\hat{\beta}_{\text{outlier}} - \hat{\beta}_{\text{causal}}}{\hat{\beta}_{\text{causal}}}$$

with $\hat{\beta}_{\text{outlier}}$ the original causal effect estimated using all variants and $\hat{\beta}_{\text{causal}}$ the corrected causal estimate obtained after removing outliers identified by MR-PRESSO. Normalizing by the absolute value of the corrected causal estimate provides a direction for the magnitude of $D$. To test for statistical significance of $D$, we calculated an empirical $P$ value by generating a null distribution (the null hypothesis corresponds to the expected distortion due to a random set of variants). We defined $n_{v}$ as the number of variants detected as outliers by the MR-PRESSO outlier test and $n_{r}$ as the total number of variants robustly associated with the exposure. The null distribution is generated by substituting $n_{v}$ variants detected as outliers by the MR-PRESSO outlier test with $n_{r} - 2n_{v}$ non-outliers, which are drawn with replacement from the entire set of non-outlier variants. This results in the total number of variants being fixed at $n_{r} - n_{v}$. We repeated this procedure $K$ times to generate the null distribution. An empirical $P$ value is then calculated as the number of times that the observed distortion is greater than the expected distortion under the null hypothesis divided by $K$.

Simulation framework. We performed simulations to evaluate the statistical properties (false-positive rate and power) to detect and correct for horizontal pleiotropy. We simulated the standard MR framework shown in Supplementary Fig. 1b with an outcome as well as two exposures $E$, and $E_j$. A total of 50 variants was simulated per case. Horizontal pleiotropy was induced by a certain percentage of the 50 variants $G$, having a significant effect on both $E_i$ (through $Y_j$) and $E_j$ (through $Y_j$). We varied the following parameters:

- The main causal effect (of exposure 1 on the outcome), $\hat{\beta}_{\text{causal}} = 0$ (no main causal effect) or $\hat{\beta}_{\text{causal}} = 0.1, 0.2, 0.5$ whereas the causal effect of exposure 2 was always set to $\hat{\beta}_{\text{causal}} = 0.1$,
- Percentage of horizontal pleiotropy (0, 2, 4, 10, 50, 90);
- Type of horizontal pleiotropy: positive (all $Y_j$ positive) or balanced (approximately half of the $Y_j$ positive and half negative);
- Verification of the InSIDE$^\gamma$ (instrument strength independent of direct effect) assumption (Supplementary Note 1).
- For each scenario, 10,000 simulations were performed.

Collection of genome-wide association (GWA) summary statistics. We retrieved publicly available genome-wide association (GWA) summary statistics data for 82 complex traits and diseases (Supplementary Note 1). We performed the following steps to ensure that all datasets were uniform and standardized. For each, we retrieved the appropriate variant annotation (build, raid, chromosome, position, reference and alternate alleles) and summary statistics (effect size, standard errors, $P$ values and sample size of the study). All variant coordinates (chr, pos) were lifted over to hgl9 using the UCSC Genome Browser LiftOver Tool. We imputed z-scores of variants using Impr$^\gamma$ using 1000 Genomes Phase 3 European panel$^\delta$ ($n=503$) as a reference panel. Effect sizes, standard errors and $P$ values were then calculated using the variance of the trait estimated from genotyped variants and allele frequencies calculated on the same subset of individuals from the 1000 Genomes reference panel. Sets of GWA-significant variants were manually retrieved from the corresponding GWA manuscripts. In total, we retrieved GWA summary statistics for 82 traits and diseases (Supplementary Table 14).

Detection and correction of horizontal pleiotropy using MR-PRESSO or covariate adjustment in MMR. We applied MR-PRESSO to all possible exposure–outcome pairs of 82 traits and diseases and then compared the results of that test to those obtained by other MR methods. In total, we performed 4,250 MR tests. Only 53 distinct traits had a sufficient set of genome-wide significant variants that could be used as IVs. This led to 53 × (82 – 1) = 4,293 possible exposure-outcome pairs. The remaining pairs were removed because of missing values in the summary data, which led to a total of 4,250 MR tests. We compared those results with results obtained by other approaches, including the Q (modified) test$^\gamma$ and Q$^\gamma$ (modified) test$^\gamma$, as well as MR-Egger regression$^\gamma$. Next, we evaluated five strategies to correct for significant horizontal pleiotropy detected from our MR-PRESSO global test. The first approach included covariates in our MMR model, either one by one or all at the same time. We considered only covariates with a statistically significant causal effect (causal estimate of the IVW meta-analysis using a Bonferroni-corrected cut-off). Furthermore, to account for co-linearity, we also included only covariates with a correlation coefficient of <0.3. We note that not all pairs were eligible for MMR analysis due to either a lack of relevant covariates to adjust on in the one-by-one model or too many covariates to adjust on in the full model. The second approach corrected for horizontal pleiotropy by removing offending variants that were statistically significant outliers according to the MR-PRESSO outlier test. The MR-PRESSO global test was performed on the adjusted MR models to determine if there was any remaining horizontal pleiotropy. Finally, in a similar fashion, the Q (modified) outlier test and Q$^\gamma$ (modified) outlier test were used to correct for horizontal pleiotropy by removing significant outliers, and the Q (modified) test and Q$^\gamma$ (modified) test, respectively, were applied to test for any remaining horizontal pleiotropy. 1,000,000 simulations were performed to calculate the empirical $P$ values for the MR-PRESSO distortion test.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code and data availability statement. We implemented MR-PRESSO using the R Project for Statistical Computing (version 3.3.1, Vienna, Austria). The MR-PRESSO software and full set of results are available online (https://github.com/rondoloh/MR-PRESSO). A list of the GWA study datasets used is in Supplementary Table 14.

References
38. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet. Epidemiol. 37, 658–665 (2013).
39. Pasaniuc, B. et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatic 30, 2906–2914 (2014).
40. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).
Experimental design

1. Sample size
   Describe how sample size was determined.
   The study utilizes summary association statistics datasets from genome-wide association (GWA) studies. Therefore, the sample size for these datasets was determined by the total number of participants in these GWA studies.

2. Data exclusions
   Describe any data exclusions.
   None

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   Experimental replication was not attempted.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   Randomization is not relevant to this study because we only used summary association statistics and did not enroll any samples/organisms/participants.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   Blinding is not relevant to this study because we only used summary association statistics and did not enroll any samples/organisms/participants.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters
   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

   □ n/a  Confirmed
   □ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   □ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   □ A statement indicating how many times each experiment was replicated
   □ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
   □ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   □ The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
   □ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   □ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.
Software

Describe the software used to analyze the data in this study.

We have developed a method called Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MR-PRESSO). We have made this software publicly available. The link to download the software is provided in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No restrictions

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in the current study.

No eukaryotic cell lines were used in the current study.

No eukaryotic cell lines were used in the current study.

No eukaryotic cell lines were used in the current study.

Animals and human research participants

Provide details on animals and/or animal-derived materials used in the study.

No animals were used in the current study.

The study did not involve human research participants.