Diagnostics of Pleiotropy in Mendelian Randomization Studies: Global and Individual Tests for Direct Effects

Web Material

Web Appendix 1: If $\beta_{2j} = 0$ for all j then $\gamma_{2j} = 0$ for all j.

Without loss of generality, assume all G_js are centered at zero so that $\mathbb{E}(G_j) = 0$ for all j. Suppose $X = \mathcal{R}(\mathbf{G}) + \epsilon$, where $\mathcal{R}(\mathbf{G}) = \sum \alpha_j G_j$ and ϵ is the error that is independent of $\mathbf{G}$. Plugging $\mathcal{R}(\mathbf{G})$ and ϵ into the data-generating model for Y and integrating out ϵ and U, we obtain

$$\log\{\Pr(Y = 1|X, U, \mathbf{G})\} = \beta_0^* + \beta_1 \mathcal{R}(\mathbf{G}) + \sum \beta_{2j} G_j,$$

whose design matrix is not full-rank since $\mathcal{R}(\mathbf{G})$ is linear combination of G_j. Suppose a marginal direct effect model for G_1 is defined as

$$\log\{\Pr(Y = 1|X, U, \mathbf{G})\} = \gamma_{01} + \gamma_{11} \mathcal{R}(\mathbf{G}) + \gamma_{21} G_1.$$

With some algebra, we show next γ_{21} is the sum of β_{21} and an additional term that involves other β_{2j}s. Let $\mathcal{K}_{-1}(\mathbf{G}) = \sum_{j \neq 1} \beta_{2j} G_j$. We proceed by decomposing $\mathcal{K}_{-1}(\mathbf{G})$ to linear combinations of $(\mathcal{R}(\mathbf{G}), G_1)$ and independent errors, the latter of which can be integrated out. Observe that

$$\mathcal{K}_{-1}(\mathbf{G}) = b_{11} \mathcal{R}(\mathbf{G}) + b_{12} G_1 + \varepsilon_1,$$

where b_{11} and b_{12} are regression coefficients that can be derived by ordinary least squares,

$$\begin{pmatrix} b_{11} \\ b_{12} \end{pmatrix} = \begin{bmatrix} \mathbb{E}\{\mathcal{R}(\mathbf{G})\mathcal{R}(\mathbf{G})\} & \mathbb{E}\{\mathcal{R}(\mathbf{G})G_1\} \\ \mathbb{E}\{\mathcal{R}(\mathbf{G})G_1\} & \mathbb{E}\{G_1 G_1\} \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}\{\mathcal{K}_{-1}(\mathbf{G})\mathcal{R}(\mathbf{G})\} \\ \mathbb{E}\{\mathcal{K}_{-1}(\mathbf{G})G_1\} \end{bmatrix}$$

Because G_js are mutually independent, $\mathbb{E}\{\mathcal{K}_{-1}(\mathbf{G})G_1\} = 0$. Some algebra leads to

$$b_{12} = -\frac{\alpha_1 \mathbb{E}(G_1 G_1) \sum_{j>1} \alpha_j \beta_{2j} \mathbb{E}(G_j G_j)}{\{\sum \alpha_j^2 \mathbb{E}(G_j G_j)\}\{\mathbb{E}(G_1 G_1)\} - \{\alpha_1 \mathbb{E}(G_1 G_1)\}^2},$$

and

$$\gamma_{21} = \beta_{21} + b_{12}.$$

This derivation shows that, if $\beta_{2j} = 0$ for all j, then $\gamma_{21} = 0$, and $\gamma_{2j} = 0$ for all j. Indeed as expected, if $\beta_{2j} = 0$ for all $j \neq j'$, then $\gamma_{2j'} = \beta_{2j'}$. In general, $\beta_{2j} \neq \gamma_{2j}$. However, it can be deduced that $b_{12} \approx 0$ when the majority of $\beta_{2j} \approx 0$, or $\alpha_j \beta_{2j}$ can be positive for some js or negative for other js.

Web Appendix 2: The distribution of $\hat{\gamma}_{2j}$s is degenerate with rank $m-1$

Under the global null hypothesis, the estimating function from a single observation for the j^{th} model is $S_j = \mathbf{X}_j(Y - \mu)$, where $\mathbf{X}_j = (1, \mathcal{R}(\mathbf{G}) = \sum_j \alpha_j G_j, G_j)^T$ and $\mu = \exp\{\beta_0 + \beta_1 \mathcal{R}(\mathbf{G})\}$. Without loss of generality, the asymptotic null distribution for $(\hat{\gamma}_{01}, \hat{\gamma}_{11}, \hat{\gamma}_{21})$ is

$$\sqrt{n}\begin{pmatrix} \hat{\gamma}_{01} - \beta_0^* \\ \hat{\gamma}_{11} - \beta_1 \\ \hat{\gamma}_{21} - \beta_2 \end{pmatrix} = \{\mathbb{E}(\mathbf{X}_1^T \mathbf{X}_1 \mu)\}^{-1} S_1 + o_p(1) = \frac{1}{\det\{\mathbb{E}(\mathbf{X}_1^T \mathbf{X}_1 \mu)\}} \begin{pmatrix} \mathrm{A}_1 & \mathrm{D}_1 & \mathrm{J}_1 \\ \mathrm{B}_1 & \mathrm{E}_1 & \mathrm{H}_1 \\ \mathrm{C}_1 & \mathrm{F}_1 & \mathrm{I}_1 \end{pmatrix} S_1 + o_p(1),$$

where

$$\begin{aligned} \mathrm{C}_1 &= \mathbb{E}(\mu^2 \textstyle\sum \alpha_j G_j)\mathbb{E}(G_1 \mu^2 \sum \alpha_j G_j) - \mathbb{E}(\mu G_1)\mathbb{E}\{(\mu \sum \alpha_j G_j)^2\} \\ \mathrm{F}_1 &= \mathbb{E}(\mu G_1)\mathbb{E}(\mu \textstyle\sum \alpha_j G_j) - \mathbb{E}(\mu^2 G_1 \sum \alpha_j G_j) \\ \mathrm{I}_1 &= \mathbb{E}\{\mu^2 (\textstyle\sum \alpha_j G_j)^2\} - \mathbb{E}\{\mu \sum \alpha_j G_j\}^2 \end{aligned}$$

Let $\Delta_1 = \det\{\mathbb{E}(\mathbf{X}_1^T \mathbf{X}_1 \mu)\}$, the influence function for $\hat{\gamma}_{21}$, denoted by $\mathcal{U}_1$, is written as

$$\frac{1}{\Delta_1}(\mathrm{C}_1 + \mathrm{F}_1 \textstyle\sum \alpha_j G_j + \mathrm{I}_1 G_1)(Y - \mu) = \widetilde{\mathbf{X}}_j (Y - \mu),$$

where $\tilde{\mathbf{X}}_j$ is the design matrix component in the influence function. Observe that

$$\begin{aligned} \textstyle\sum \Delta_j \alpha_j \dfrac{\mathrm{C}_j}{\Delta_j} &= \mathbb{E}(\mu^2 \textstyle\sum \alpha_j G_j)\mathbb{E}\{\mu^2 (\sum \alpha_j G_j)^2\} - \mathbb{E}(\mu^2 \sum \alpha_j G_j)\mathbb{E}\{\mu^2 (\sum \alpha_j G_j)^2\} = 0 \\ \textstyle\sum \Delta_j \alpha_j \dfrac{\mathrm{F}_j \sum \alpha_j G_j}{\Delta_j} &= \textstyle\sum \alpha_j G_j \mathbb{E}(\mu \sum \alpha_j G_j)\mathbb{E}(\mu \sum \alpha_j G_j) - \sum \alpha_j G_j \mathbb{E}\{\mu^2 (\sum \alpha_j G_j)^2\} \\ \textstyle\sum \Delta_j \alpha_j \dfrac{\mathrm{I}_j G_j}{\Delta_j} &= \textstyle\sum \alpha_j G_j \mathbb{E}\{\mu^2 (\sum \alpha_j G_j)^2\} - \sum \alpha_j G_j \mathbb{E}(\mu \sum \alpha_j G_j)\mathbb{E}(\mu \sum \alpha_j G_j), \end{aligned}$$

So that $\sum_j \alpha_j \Delta_j \widetilde{\mathbf{X}}_j = 0$, which means the design matrix component of the influence function for m $\hat{\gamma}_{2j}$s is linearly dependent with rank $m-1$. So the corresponding asymptotic distribution is degenerate with rank $m-1$.

Web Appendix 3: A parametric simulation procedure to obtain the null distribution of ordered p-values $(p_{(1)}, ..., p_{(m)})$

Following the derivation of **Web Appendix 2**, let the influence function for $(\hat{\gamma}_{21}, ..., \hat{\gamma_{2m}})$ be $(\mathcal{U}_1, ..., \mathcal{U}_m)$, the asymptotic null distribution is multivariate normal, expressed as $\mathcal{N}(\mathbf{0}, \mathcal{U}^T\mathcal{U})$, and the null distribution of the associated z-scores $(z_1, ..., z_m)$ can be similarly derived. Both distributions are degenerated with rank $m-1$. We first simulate the null distribution by first simulating the first $m-1$ z_j from the full rank multivariate normal distribution. The distribution of z_m conditional on $(z_1, ..., z_{m-1})$ is degenerated to a fixed point, and so we set z_m to be the expected mean of the z_m on $(z_1, ..., z_{m-1})$ computed by the joint multivariate normal null distribution. The corresponding $(p_1, ..., p_m)$ are computed in each simulated dataset, and ordered to obtain the quantiles. All quantiles $(p_{(1)}, ..., p_{(m)})$ are therefore obtained from simulated datasets. Typically, more than 10^4 simulated datasets are needed to be generated to obtain a reliable distribution for the tail of the quantiles, for example the minimum p-value.

Web Appendix 4: Sensitivity analysis for validity of the two methods when G_js and U are correlated

A simulation study was conducted to assess the sensitivity of the GLIDE method to violation of the independence assumption between G_js and U. The parameter setting is identical to the null simulations presented in Table 1, except that in model (4) $U = \sum_j \phi_j + \epsilon_1$ and ϕ_js are sampled from Uniform(0,1,0.2). Web Table 1 shows the empirical type I error when there is correlation between G_js and U. Neither MR-Egger nor GLIDE provide a valid test of pleiotropy, with more inflated type I error rates from MR-Egger.

Web Table 1: The type I error rate for the proposed GLIDE test and the Egger test when the nominal p-value is 0.05 and G_js are correlated with U.

Sample size	β_1	GLIDE	MR-Egger
500 cases/500 controls	0	0.1964	0.5290
	0.5	0.1208	0.2772
2500 cases/2500 controls	0	0.7916	0.9464
	0.5	0.6020	0.8000

Web Appendix 5: Parameter settings for simulation experiments in Figure 2

In Figure 2(A) γ_js were randomly sampled from a uniform distribution from 0 to 0.6θ, with θ increasing from 0 to 1 to create a ladder of effect sizes for direct effects. In this setting, the direct effects are all positive. Figure 2(B) shows the scenario where the direct effect γ_js were generated from the uniform distribution $(0.4\theta, 0.3\theta)$ with θ increasing from 0 to 1. In this scenario, some direct effects are negative and some are positive, representing "balanced pleiotropy". Figures 2(C) examines the scenario where a subset of variants (fifteen out of the twenty five) have pleiotropic effects, which can be more plausible in biology than all variants having direct effects. In this scenario the pleiotropic effects are all positive. Figure 2(D) shows the scenario where there is correlation between α_j and γ_j, and the correlation is positive. Specifically, α_js were sampled from Uniform(0.1,0.2) + Uniform(0,0.4θ), and γ_js were sampled from Uniform(0, θ). Therefore in this scenario the InSIDE condition is violated.

Web Appendix 6: Estimation of individual direct effects

A simulation study was conducted to evaluate the bias of the estimated surrogate direct effects as estimates of true individual direct effects. The degree of approximation between γ_{1j} and β_{2j} is critical to the use of the q-q plot as a means to identify individual SNPs with pleiotropic effects. In this set of simulations, we varied the number of SNPs m from 25, 50, to 100 and let $\phi_j = 0$ for all j. We assigned the direct effect (either 0 or 0.3) to the m^{th} SNP and evaluated the bias of the estimated surrogate direct effect for the direct

effect of this particular SNP. For the rest of $m1$ SNPs, γ_j is generated from either one-sided pleiotropy (uniform distribution between 0 and 0.3) or balanced pleiotropy (uniform distribution between 0.2 and 0.2). The bias was evaluated in 2000 simulated datasets, each with case-control sampling of approximately 500 cases and 500 controls. Web Table 2 shows the bias of the estimated surrogate direct effect for the m^{th} SNP in a number of settings. Clearly if direct effects are balanced in positive and negative signs, there is little bias when using the surrogate direct effect as an estimate of the true direct effect. When direct effects are one-sided, the bias reduces with an increasing number of SNPs and an increasing portion of null SNPs. For SNPs studied in Mendelian randomization, the direct effects can be negative or positive, just as the instrumental strength parameter α_j has been observed in our studies to be positive for some SNPs and negative for others. For the numbers of SNPs we will investigate next in the BMI and height analyses for GECCO (77 and 696, respectively), the results in Web Table 2 suggest that the bias for individual estimated direct effects in our investigation can be largely negligible.

Web Table 2: The bias of the estimated surrogate direct effect relative to the true direct effect for the m^{th} SNP when some of the rest of (m-1) SNPs may have direct effects.

	% null in the	Balanced Pleiotropy			One-sided Pleiotropy		
	rest of (m-1) SNPs	m=25	m=50	m=100	m=25	m=50	m=100
$\beta_{2m} = 0$	20%	7e-4	0.010	−0.007	0.062	0.006	−0.008
	50%	−0.012	0.009	−0.005	−0.064	0.002	0.004
	80%	−0.029	5e-4	0.007	−0.007	0.002	0.005
	100%	0.001	0.001	0.002	0.001	0.001	0.002
$\beta_{2m} = 0.3$	20%	0.005	0.010	−0.003	0.071	0.010	−0.009
	50%	0.010	0.015	−0.004	−0.060	0.006	−0.003
	80%	0.027	0.005	0.012	0.005	0.005	0.011
	100%	2e-4	6e-4	0.004	2e-4	6e-4	0.004

Web Appendix 7: Description of the eleven studies in the data analysis

The 11 GECCO studies used in our analysis have been previously described (ref. 1), including the Health Professionals Follow-up Study (HPFS, ref. 2); Nurses Health Study (NHS, ref. 3); Physicians Health Study (PHS, ref. 4); Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO, ref. 5); VITamins and Lifestyle Study (VITAL, ref. 6); Womens Health Initiative (WHI, ref. 7); the Colon-Cancer Family Registry (C-CFR, ref. 8); Ontario Familial Colon Cancer Registries (OFCCR, ref. 9); Diet, Activity and Lifestyle Survey (DALS, ref. 10-11); Postmenopausal Hormone Study (PMH-CCFR, ref. 12); and Darmkrebs: Chancen der Verhtung durch Screening (DACHS, ref. 13). There were no overlap of participants between the 11 studies.

GECCO: National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088; R01 CA059045; U01 CA164930).

CCFR : This work was supported by grant UM1 CA167551 from the National Cancer Institute and through cooperative agreements with the following CCFR centers: Australasian Colorectal Cancer Family Registry (U01 CA074778 and U01/U24 CA097735); Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (U01/U24 CA074800); Ontario Familial Colorectal Cancer Registry (U01/U24 CA074783); Seattle Colorectal Cancer Family Registry (U01/U24 CA074794); University of Hawaii Colorectal Cancer Family Registry (U01/U24 CA074806); USC Consortium Colorectal Cancer Family Registry U01/U24 CA074799); The Colon CFR GWAS was supported by funding from the National Cancer Institute, National Institutes of Health (U01 CA122839 and R01 CA143237 to Graham Casey). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Colon Cancer Family Registry (CCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CCFR.

DACHS: German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4 and CH 117/1-1), and the German Federal Ministry of Education

and Research (01KH0404 and 01ER0814).

DALS: National Institutes of Health (R01 CA48998 to M. L. Slattery)

HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, R01 CA137178, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003), NHS by the National Institutes of Health (R01 CA137178, P01 CA087969, UM1 CA186107, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003,) and PHS by the National Institutes of Health (R01 CA042182).

OFCCR: National Institutes of Health, through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783); see CCFR section above. Additional funding toward genetic analyses of OFCCR includes the Ontario Research Fund, the Canadian Institutes of Health Research, and the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation.

PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Additionally , a subset of control samples were genotyped as part of the Cancer Genetic Markers of Susceptibility (CGEMS) Prostate Cancer GWAS (Yeager, M et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007 May;39(5):645-9), CGEMS pancreatic cancer scan (PanScan) (Amundadottir, L et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet. 2009 Sep;41(9):986-90, and Petersen, GM et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet. 2010 Mar;42(3):224-8), and the Lung Cancer and Smoking study (Landi MT, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. 2009 Nov;85(5):679-91). The prostate and PanScan study datasets were accessed with appropriate approval through the dbGaP online resource (http://cgems.cancer.gov/data/) accession numbers phs000207.v1.p1 and phs000206.v3.p2, respectively, and the lung datasets

were accessed from the dbGaP website (http://www.ncbi.nlm.nih.gov/gap) through accession number phs000093.v2.p2. Funding for the Lung Cancer and Smoking study was provided by National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01 HG 004438. For the lung study, the GENEVA Coordinating Center provided assistance with genotype cleaning and general study coordination, and the Johns Hopkins University Center for Inherited Disease Research conducted genotyping.

PMH: National Institutes of Health (R01 CA076366 to P.A. Newcomb).

VITAL: National Institutes of Health (K05 CA154337).

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

References

[1] Peters U, Jiao S, Schumacher FR, et al. Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology.* 2013;144:397403.

[2] Giovannucci E, Rimm EB, Stampfer MJ, et al. Aprospective study of cigarettesmoking andrisk of colorectal adenoma and colorectal cancer in U.S. men. *J Natl Cancer Inst.* 1994;86:18391.

[3] Giovannucci E, Colditz GA, Stampfer MJ, et al. A prospective study of cigarette smoking and risk of colorectal adenoma and colorectal cancer in U.S. women. *J Natl Cancer Inst.* 1994;86:1929.

[4] Christen WG, Gaziano JM, Hennekens CH. Design of Physicians' Health Study IIa randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention

of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann Epidemiol.* 2000; 10:12534.

[5] Prorok PC, Andriole GL, Bresalier RS, et al. Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control Clin Trials.* 2000;21(6 Suppl):273S309S.

[6] White E, Patterson RE, Kristal AR, et al. VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *Am J Epidemiol* 2004;159:8393.

[7] Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials.* 1998;19:61109.

[8] Newcomb PA, Baron J, Cotterchio M, et al. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev* 2007;16:233143.

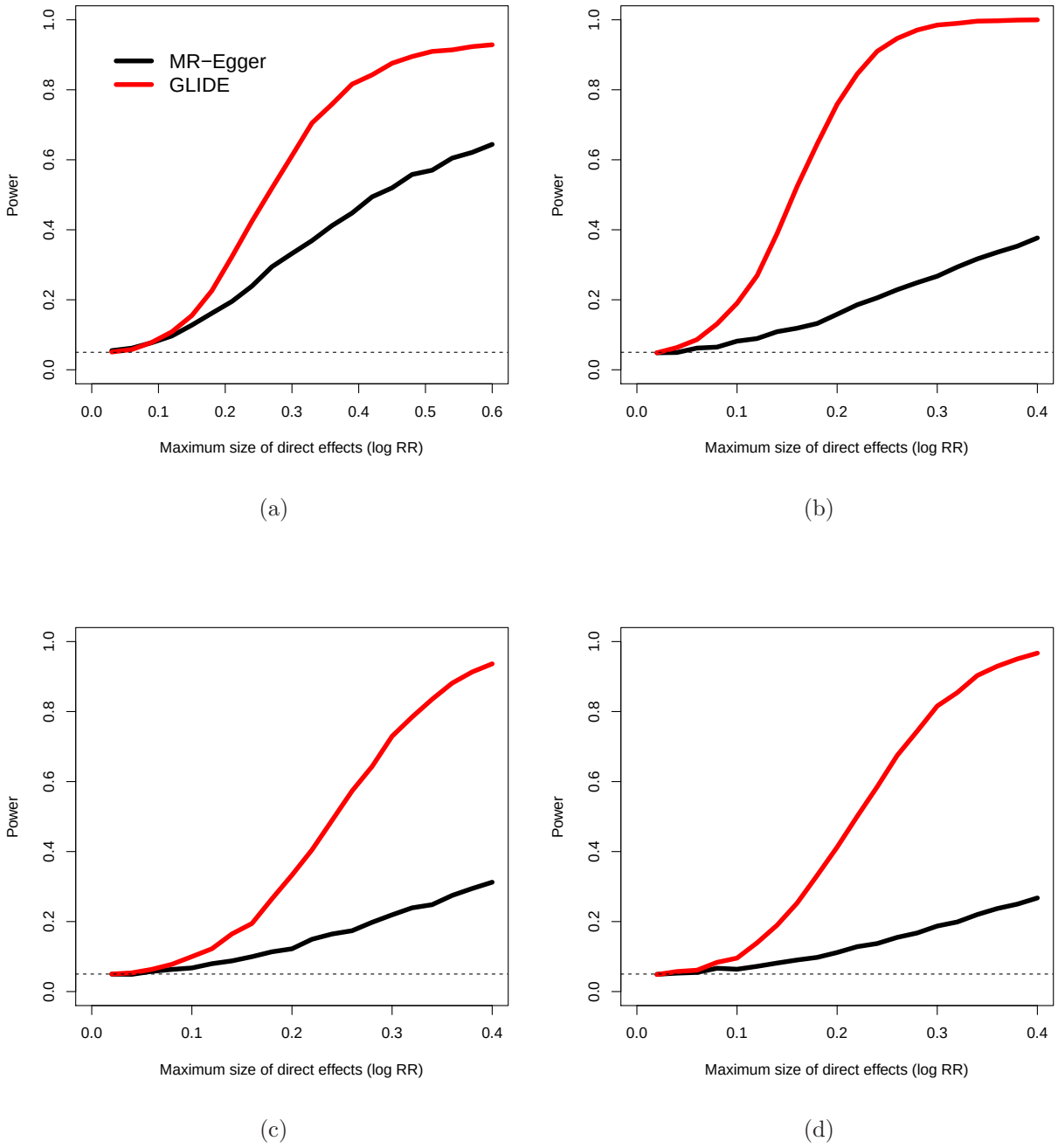
[9] Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2007;39:98994.

[10] Slattery ML, Friedman GD, Potter JD, Edwards S, Caan BJ, Samowitz W. A description of age, sex, and site distributions of colon carcinoma in three geographic areas. *Cancer* 1996;78:166670.

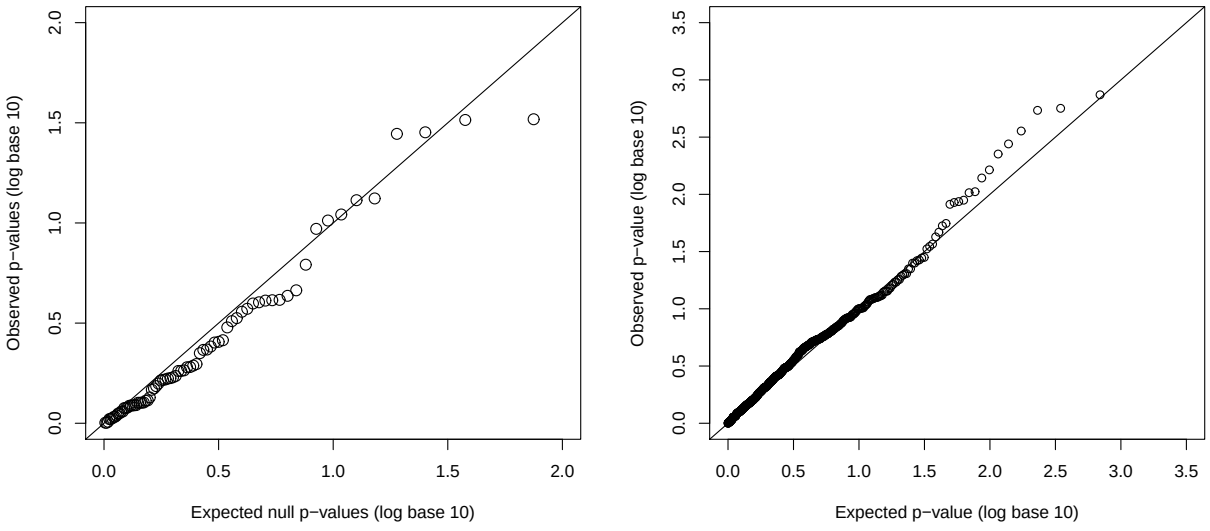
[11] Slattery ML, Potter J, Caan B, et al. Energy balance and colon cancerbeyond physical activity. *Cancer Res* 1997; 57:7580.

[12] Newcomb PA, Zheng Y, Chia VM, et al. Estrogen plus progestin use, microsatellite instability, and the risk of colorectal cancer in women. *Cancer Res* 2007; 67:75349.

[13] Brenner H, Chang-Claude J, Seiler CM, et al. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Ann Intern Med* 2011;154:2230.



Web Figure 1: Statistical power of GLIDE and MR-Egger in simulations to test the global null hypothesis, that there is no direct effect for any genetic variant. There is no causal effect from the exposure to the disease outcome. (a) The direct effects are all positive; (b) Some direct effects are positive and some are negative; (c) A proportion of SNPs (60%) have pleiotropic effects; (d) All SNPs have direct effects that are correlated with the genetic associations with the intermediate exposure.



(a) (b)

Web Figure 2: The q-q plots for p-values of the refined set of SNPs as instrumental variables after removing SNPs with evidence of pleiotropy in Figure 1. (a) 75 BMI-associated SNPs. (b) 693 height-associated SNPs.