International Journal of Epidemiology, 2016, 1600–1616
doi: 10.1093/ije/dyw088
Advance Access Publication Date: 24 June 2016
Original article

Mendelian randomization

Selecting instruments for Mendelian randomization in the wake of genome-wide association studies

Daniel I Swerdlow,[1,2,*] Karoline B Kuchenbaecker,[3] Sonia Shah,[1] Reecha Sofat,[1,4] Michael V Holmes,[1,5] Jon White,[1] Jennifer S Mindell,[6] Mika Kivimaki,[6] Eric J Brunner,[6] John C Whittaker,[7,8] Juan P Casas,[7] and Aroon D Hingorani[1]

[1]Institute of Cardiovascular Science, University College London, London, UK, [2]Department of Medicine, Imperial College London, London, UK, [3]Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK, [4]Centre for Clinical Pharmacology and Therapeutics, University College London, London, UK, [5]Clinical Trial Service Unit and Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, Oxford, UK, [6]Research Department of Epidemiology & Public Health, University College London, London, UK, [7]Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK and [8]Genetics Division, Research and Development, GlaxoSmithKline, NFSP, Harlow, UK

*Corresponding author. Institute of Cardiovascular Science, University College London, Gower Street, London, WC1E 6BT, UK. E-mail: d.swerdlow@ucl.ac.uk

Accepted 30 March 2016

Abstract

Mendelian randomization (MR) studies typically assess the pathogenic relevance of environmental exposures or disease biomarkers, using genetic variants that instrument these exposures. The approach is gaining popularity—our systematic review reveals a greater than 10-fold increase in MR studies published between 2004 and 2015. When the MR paradigm was first proposed, few biomarker- or exposure-related genetic variants were known, most having been identified by candidate gene studies. However, genome-wide association studies (GWAS) are now providing a rich source of potential instruments for MR analysis. Many early reviews covering the concept, applications and analytical aspects of the MR technique preceded the surge in GWAS, and thus the question of how best to select instruments for MR studies from the now extensive pool of available variants has received insufficient attention. Here we focus on the most common category of MR studies—those concerning disease biomarkers. We consider how the selection of instruments for MR analysis from GWAS requires consideration of: the assumptions underlying the MR approach; the biology of the biomarker; the genome-wide distribution, frequency and effect size of biomarker-associated variants (the genetic

architecture); and the specificity of the genetic associations. Based on this, we develop guidance that may help investigators to plan and readers interpret MR studies.

Key words: Mendelian randomization, genome-wide association study, biomarkers, causal inference

Key Messages

- MR offers novel opportunities for reliable causal inference within the framework of observational research designs.
- The findings from an MR analysis can provide insight into the pathophysiology of complex disease and have translational relevance, including the prioritization of drug targets.
- The emerging genetic architecture of disease biomarkers now allows more informed selection of genetic variants for MR studies than was hitherto possible.
- As the number of biomarker-associated variants grows though genome-wide association studies and, more recently, metabolomics and proteomics, selection of the most appropriate instruments for MR analysis will become an increasingly important issue.
- We have proposed a set of principles that should inform the selection process to aid the design, analysis and interpretation of MR studies.

Introduction

Adverse environmental influences, such as smoking and alcohol consumption, are associated with a higher risk of many chronic, non-communicable diseases. Individuals at higher risk also exhibit alterations in numerous quantitative biological traits (also known as disease biomarkers or intermediate phenotypes), years before disease onset (Supplementary Table 1, available as Supplementary data at *IJE* online). These associations have been identified mainly through non-genetic observational studies. However, observational epidemiological studies of this type can be subject to a variety of biases. Importantly, it can be difficult to separate causal associations from those that arise from confounding or reverse causation. Effect estimates from such studies may also be prone to regression dilution bias[1] and errors in the measurement of the biomarker for technical or biological reasons.[1]

Mendelian randomization (MR) is an evolving paradigm in which genetic variants (usually single nucleotide polymorphisms, SNPs) are used to help distinguish causal from non-causal associations between environmental exposures or biomarkers and disease outcomes.[2] Two unique attributes of genotype make this possible. First, the random allocation of parental alleles to zygotes at meiosis, independent of environmental exposures, reduces the potential for confounding in genetic association studies in the same way as randomized treatment allocation in clinical trials[3,4] (Figure 1a). Second, the invariant nature of the DNA sequence and unidirectional flow of biological information, from gene sequence through intermediate phenotypes to disease, avoids reverse causation, though it should not be taken to imply a stability of genetic effect which in theory could be modified in a context-dependent fashion.[5]

An MR study typically considers three types of association: (i) the association of a biomarker (or environmental exposure) with the disease outcome; (ii) the association of a genetic variant with biomarker or environmental exposure; and (iii) the association of the same variant with disease risk[6] (Figure 1b). Provided certain assumptions are met (Figure 1), consistency in direction and magnitude of the three estimates provides evidence on causal relevance of the environmental exposure or biomarker. The causal effect can be quantified within a formal statistical framework, using instrumental variables methods which have been adopted and adapted from the econometric literature.[7,8] Some illustrative examples of the early use of MR are outlined in Box 1 and Table 1, and more recent examples that have exploited certain enhancements to the MR approach, are described in more detail later in this. It is notable that several important MR studies of certain disease biomarkers have identified inconsistency between effect estimates obtained in non-genetic observational studies and those through MR analysis that have altered thinking on the causal relevance of those biomarkers, as we describe later.

A systematic review (see Supplementary methods for details, available as Supplementary data at *IJE* online) reveals a 10-fold increase in MR studies published between 2004 and 2015 (Supplementary Figure 1, available as Supplementary data at IJE online). The majority have been in the fields of cardiovascular disease and diabetes (51% of published studies); other disease areas including cancer

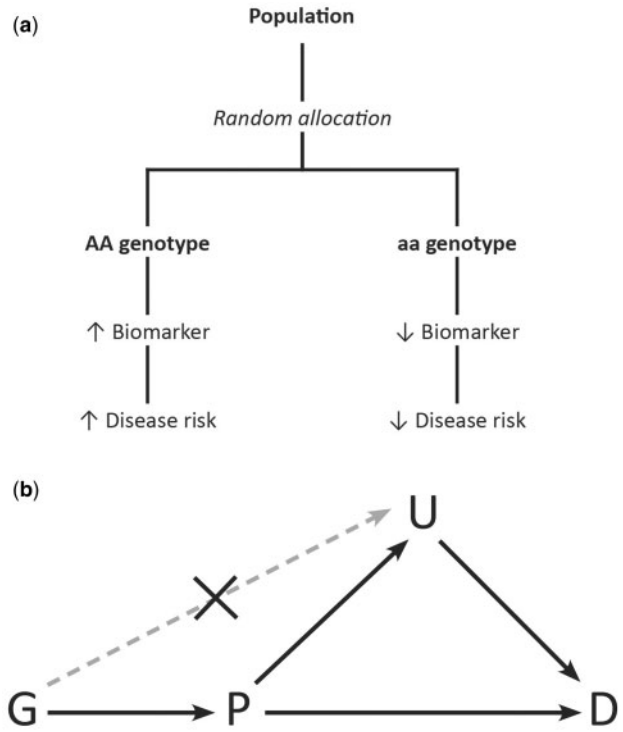

Figure 1. A: Mendelian randomization is a natural analogue of the classical randomized controlled trial (RCT). Random allocation of alleles at conception and the unidirectional flow of information from DNA sequence to endogenous biomarker phenotype allow causal inference of the type possible within the RCT framework. Genotype is generally unrelated to environmental exposures, thus reducing confounding. B: the Mendelian randomization model: the causal role of an exposure, P, on a disease state, D, is being evaluated. A genetic variant, G, is associated with biomarker P but not with confounders, U. Variant G is also associated with disease D and acts only through its effects on biomarker P. The model rests on three core assumptions: (i) the genetic instrument (G) is associated with the exposure or biomarker of interest (P); (ii) the genetic instrument (G) is independent of potential confounding factors (U) in the relationship between the exposure/biomarker (P) and the outcome (D); (iii) the outcome (D) is associated with the genetic instrument (G) only through the effect of the exposure/biomarker (P), and is in all other respects independent.

(10%); and mental health (10%). Most MR studies (86%) have been of disease biomarkers (defined in Box 2) such as blood lipids, body mass index (BMI) or blood pressure, and 50% have used a candidate gene approach to identify suitable instruments (Table 1). However, genome-wide association studies (GWAS) of disease biomarkers are providing a new source of instruments for MR analysis. Of the 2111 GWAS listed in the NIH National Human Genome Research Institute (NHGRI) GWAS catalogue[9,10] [http://www.genome.gov/gwastudies, as of 23 August 2015], 672 (32%) concern genetic variants associated with 520 disease biomarkers, with some variants exhibiting associations with more than one biomarker. Other studies based on high-density locus-centric SNP arrays such as Metabochip[11] and Immunochip[12], designed based on GWAS findings in cardiometabolic and autoimmune/inflammatory disorders respectively, have reported many additional genotype-biomarker associations. Many MR studies ($n = 211$) were published after a GWAS of their corresponding biomarker; of those studies, 61% ($n = 129$) used the preceding GWAS to inform the selection of the instruments.

Many of the early reviews in the field that covered the concept, applications[1,2,7,13–15] and analytical aspects of the MR technique[16–18] preceded the surge of GWAS. Thus, the question of how best to select instruments for MR studies, given the now extensive pool of available variants, has received insufficient attention. In this article we focus on the most common category of MR studies—those concerning disease biomarkers (see Box 2). We show that using GWAS as a source of instruments for MR analysis requires consideration of the assumptions underlying the MR approach, the biological nature of the biomarker of interest, the distribution of SNP-biomarker associations at the genome-wide and regional levels, the genetic effect sizes and specificity of associations.

BOX 1. Applications of Mendelian randomization

MR analysis has been applied to assess whether CRP, a circulating marker of inflammation, plays a true causal role in the development of CHD. Despite the robust association of CRP level with CHD in observational studies, *CRP* variants used to instrument long-term elevations in CRP concentration did not provide evidence of a causal role for this biomarker in the development of CHD, based on meta-analysis of up to 47 studies including 46 557 cases.[78,82,125] The observational association between CRP and CHD is more likely explained by confounding or reverse causation. HDL-cholesterol (HDL-C) exhibits an inverse association with CHD risk in observational studies, but whether this association is causal has been in dispute. An MR study used variants in the *LIPG* gene, encoding hepatic lipase, as an instrument for HDL-C and examined its relationship with myocardial infarction (MI) risk.[104] Although higher HDL-C is observationally associated with lower MI risk, MR analysis based on *LIPG* variants, both alone and within allele scores to instrument HDL-C concentration, did not find evidence for a causal role for HDL-cholesterol in CHD.

Table 1. Illustrative examples of different types of MR study. Examples are provided of MR studies of exogenous exposures, *cis*-MR for drug target validation, and disease biomarker MR analysis

Author (year)	Location	Date of relevant GWAS	Exposure	Endpoint	Sample characteristics	Source of variant(s)	No. of variants	Genes	Hypothesized effect shown?	Formal MR methods	Meta-analysis	Total *n* (cases/controls)
Bech (2006)[117]	Denmark	2011	Caffeine intake	Stillbirth	Pregnant women	Candidate gene	3	*NAT2*, *CYP1A2*, *GSTA1*	Yes	No	No	299 (142/157)
Holmes (2014)[76]	UK	2011	Alcohol intake	CHD	General population	Candidate gene	1	*ADH1B*	Yes	Yes	Yes	261991 (20259/168731)
Sofat (2010)[118]	UK	–	CETP inhibition	Blood pressure	General population	pQTG	2	*CETP*	No	Yes	Yes	58948
Swerdlow (2014)[91]	UK	–	HMG-CoA reductase inhibition	Type 2 diabetes	General population	Candidate gene	2	*HMGCR*	Yes	No	Yes	223,463 (26236/164842)
Swerdlow (2012)[90]	UK	–	Interleukin-6 signalling	CHD	General population	pQTG	3	*IL6R*	Yes	No	Yes	133449 (25458/100740)
Rasmusen-Torvik (2010)[119]	USA	2008	Fasting glucose	Carotid IMT	General population	GWAS	5	*GCKR*, *G6PC2*, *GCK*, *SLC30A8*, *MTNR1B*	Yes	Yes	No	7260
Elliott (2009)[82]	UK	2008	CRP	CHD	General population/ case-control	GWAS	1	*IL6R*	Yes	Yes	Yes	46434 (14365/32069)
Giltay (2009)[120]	Netherlands	2008	Cholesterol	Depressive symptoms	Elderly men	Candidate gene	1	*APOE*	No	No	No	1089
Lim (2009)[121]	Singapore	2006	Obesity	Cataract	General population	Candidate gene	1	*FTO*	No	No	No	3000 (1339/1661)
Linsel-Nitschke (2008)[122]	Germany	2008	LDL-C	CHD	General population	Candidate gene	1	*LDLR*	Yes	Yes	No	7579 (1324/6255)
Perry (2009)[123]	UK	–	Beta-carotene	Diabetes mellitus	Case-control	Candidate gene	1	*BCMO1*	No	Yes	No	10128 (4549/5579)
Trompet (2009)[124]	Netherlands	2008	Cholesterol	Cancer	Elderly population	Candidate gene	1	*APOE*	No	No	No	2913 (290/2623)

pQTG, protein quantitative trait gene; LDL-C, Low density lipoprotein cholesterol; IMT, intima-media thickness.

BOX 2. A hierarchy of biomarkers for Mendelian randomization studies based on the central dogma

For the purposes of this review, we separate exposures that might alter disease risk that are external (exogenous) to the body (e.g. cigarette smoke or socioeconomic position) from those that are internal to the body (endogenous), which we refer to as disease biomarkers. We recognize a hierarchy of disease biomarkers that reflects the central dogma—the unidirectional information flow from gene through mRNA to protein. The influence of genetic variation is initially on mRNA sequence or level, and then on the function or amount of the encoded protein. Such alterations in proteins then lead to the downstream biochemical or structural alterations, including changes in more complex phenotypes (e.g. blood pressure) that affect disease risk. Among these endogenous exposures, we draw a natural distinction between proteins and more downstream biomarkers because proteins usually represent products of individual genes and are the most proximal, widely measured consequence of natural genetic variation (Figure 2).

Assumptions underlying MR analysis

The MR approach, as classically described, rests on the assumption that any disease association of a genetic variant employed as an instrument because it proxies the biomarker of interest should be both unconfounded and explained exclusively through an effect on the biomarker (Figure 1b).[15] A potential violation of these assumptions occurs when an SNP associates with several biomarkers, only one of which is of causal interest. The association of a genetic variant with more than one phenotype is commonly referred to as pleiotropy. When pleiotropy is observed, two of the three critical assumptions of an MR analysis may be called into question. However, as we show later, a pleiotropic variant need not necessarily be excluded as an instrument, provided careful consideration is given to the mechanism giving rise to the pleiotropy and to the nature of the biomarker of interest; specifically, whether or not this is a protein. We also evaluate a number of enhancements to the basic MR design, based on multiple instruments which have since been developed partly to enhance power of MR studies, and partly to overcome some of the challenges imposed by pleiotropic instruments.

Disease biomarkers and their position in the putative disease pathway

Interest in some disease biomarkers is in their performance as predictors of disease risk.[1] For this application, it is not essential that the biomarker-disease association is causal; merely that there is a demonstrable and consistent association of the biomarker with the disease, that is of sufficient magnitude to make it a useful predictor. However, if there is interest in the potential aetiological role of a biomarker that might be amenable to modification by public health measures or drug treatment, evidence on a causal association is essential. Thus, reliable demonstration of even a modest causal effect through genetic association analysis could still be important because of the potential to develop interventions with a much larger effect on the same biomarker.[2,16]

Disease biomarkers are biologically diverse, encompassing circulating proteins (e.g. fibrinogen, C-reactive protein or interleukin-6), low molecular weight metabolic intermediates (e.g. homocysteine and uric acid) and complex physiological phenotypes such as blood pressure (Supplementary Table 1). Most biomarkers are continuous traits with genetic and environmental determinants. Many follow an approximately normal (or log-normal) distribution, and show a linear (or log-linear) association with disease risk. As we show later in a detailed discussion of potential reasons for genetic pleiotropy, the position of the biomarker of interest in the pathway connecting genetic variation to disease risk has an important bearing on the design, interpretation and validity of an MR study. In particular, we show why MR analysis of protein biomarkers instrumented by SNPs in the encoding gene has certain advantages over other categories of MR analysis.

Genetic architecture of SNP-biomarker associations

The wealth of GWAS findings allows some observations to be made about the genetic architecture of different disease biomarkers, which has bearing on the selection of SNPs for MR analysis of these traits. However, it must be borne in mind that most previous GWAS have utilized genotyping arrays that have a bias towards common variants, so that there is less information on alleles of lower frequency and their potential role as instruments in MR analysis.

C-reactive protein (CRP), an acute-phase protein associated in observational studies with cardiovascular disease (CVD) risk, provides an illustrative example (Box 1).

Three 'Manhattan' plots (Supplementary Figure 2, available as Supplementary data at *IJE* online) depict genetic associations with CRP: the first is based on findings in 5000 participants from the Whitehall II study,[19] genotyped using a gene-centric 50000 -SNP array (IBC HumanCVD BeadChip 'Cardiochip') covering 2100 genes implicated in CVD;[20] the second is from a GWAS in 6345 participants from the Women's Genome Health Study;[21] and the third is from a subsequent meta-analysis of GWAS of CRP including 82 725 participants from 15 studies.[22] The findings illustrate some general features of genomic associations with biomarkers.

First, genetic associations with mRNA expression or protein biomarkers such as CRP may be detected with smaller sample sizes when compared with studies of disease endpoints, presumably because the level or function of a protein biomarker is a comparatively proximal consequence of genetic variation, with fewer biological steps between DNA sequence variation and protein synthesis and a larger signal-to-noise ratio.[23] For more distal biomarkers such as blood metabolites or complex physiological phenotypes such as blood pressure, larger samples have typically been required. Nevertheless, regardless of the type of biomarker, increasing sample size, usually through meta-analysis, leads to identification of additional associated variants. Low-frequency variants, such as those identified by newer exome and whole genome sequencing studies sometimes of larger effect than common variants studied in GWAS, but common alleles can also on occasion produce large effect sizes. However, whole genome arrays are mainly populated by common alleles and even imputation against the 1000 Genomes reference panel most efficiently captures information on other common rather than rare alleles. Therefore, the new loci detected later in larger GWAS datasets tend to also harbour common variants but with smaller effects than the loci identified by earlier, smaller studies. For example, when 25 independent GWAS of CRP were pooled by meta-analysis, with an aggregate sample of 82 725 individuals, 12 additional loci were identified beyond the 7 reported by an earlier, smaller study. The effect sizes at each of these new loci were generally smaller than in the sentinel study (Supplementary Figure 3, available as Supplementary data at *IJE* online).[22] Meta-analyses of GWAS of blood lipids,[24] BMI,[25] blood pressure[26] and other disease biomarkers have also led to the identification of new loci also generally of smaller phenotypic effect, undetected by earlier, smaller GWAS.

Second, loci containing genetic variants associated with CRP are scattered throughout the genome. This appears to be a general feature of loci associated with disease biomarkers, such as circulating metabolites (e.g. homocysteine[27] and uric acid[28,29]), lipoproteins,[30–32] metabolomic profiles[33–36] and the more complex physiological phenotypes such as blood pressure[26,37,38] and BMI.[25,39–46] For protein biomarkers like CRP, a natural and important distinction emerges between two categories of genetic variants that might be used for MR analysis. The first are those variants acting in *cis*, located in the vicinity of the encoding gene (in this case *CRP*, chr1q23.2), which are potentially coincident with *cis*-eQTLs (expression quantitative trait loci) influencing mRNA expression. GWAS of mRNA expression profiles and protein biomarker concentrations indicate that *cis*-acting variants are a common feature of the genome.[47] The second category contains those acting in *trans*, i.e. located outside the gene encoding the protein biomarker of interest, often on a different chromosome. A variant at one locus associated with an effect on expression of a distant gene may operate via chromosomal conformational mechanisms, through microRNAs that alter mRNA stability of a range of distant target genes,[48] or because they are located in genes encoding transcription factors that regulate expression of other physically distant genes[49] or by downstream biochemical mechanisms. It is SNPs of this type that can often be pleiotropic.

Third, it is typical not only to identify associations with biomarkers at widely separated genomic locations, but also to observe multiple biomarker-associated SNPs at each locus. Although multiple independent causal variants may be present at a single locus, the multiplicity of associations commonly arises due to linkage disequilibrium (LD) between SNPs, only a subset or one of which may be functional. In order to use a SNP as an instrument in MR analyses, it is not necessary to prove the SNP itself is the causal variant, provided that its association with the biomarker of interest arises from LD with a causal variant within the same locus. Moreover, and importantly, there must be no additional LD with other nearby variants that might influence the expression or activity of a different protein. Were that the case, LD would lead to confounding and violate a key assumption of the MR paradigm. Both local LD (i.e. in the immediate vicinity of a given SNP) and distant LD (i.e. elsewhere on the same chromosome) can be ascertained using web-based tools such as the SNP Association and Proxy Search (SNAP) resource at [http://www.broadinstitute.org/mpg/snap/[50]]. In the CRP example, the *CRP* gene[51–55] is isolated by two recombination hotspots (Supplementary Figure 4a, available as Supplementary data at *IJE* online) with no evidence for LD with SNPs in the adjacent *DUSP23* and *APCS* genes. This substantially reduces the risk that confounding by LD would compromise MR analysis using SNPs in the *CRP* gene. SNP selection for an MR analysis becomes more challenging where multiple SNPs are in LD, all associate with the biomarker of interest and the associations span

several genes in close physical proximity. For example, a 74kb region of chromosome 1 (chr1p36.2) contains the *MTHFR*, *NPPA* and *NPPB* genes and includes SNPs associated with circulating concentrations of homocysteine, atrial- and brain–type natriuretic peptide, each of which have been implicated as causal factors in cardiovascular disease[56–58] (Supplementary Figure 4b). Statistical methods for prioritizing SNPs in such circumstances, such as conditional analysis or variable selection, are available and have been described elsewhere.[59,60] Recent developments include a Bayesian statistical test to quantify the probability that associations observed at the same locus with a range of outcomes (e.g. mRNA expression, blood a biomarker disease outcome) can be explained by the same causal variant,[61] which may help map association signals from a GWAS to the responsible gene. However, functional annotations or experimental evidence may be required in some cases to support the selection of instruments.

Genetic effect size

SNP arrays deployed in GWAS contain common variants (minor allele frequency, MAF, > 5%), which tend to have small to moderate effect sizes.[62] Statistical analyses in GWAS set stringent significance thresholds (typically *P*-value $< 5 \times 10^{-8}$) in order to reduce the number of false-positive associations arising from the vast number of statistical tests performed. For this reason, and because false-positive associations were a feature of the era of candidate gene association studies,[63] much attention in a GWAS is correctly on the reliability of any genetic association, based on the *P*-value.

Provided an association is identified robustly, the size of the genetic effect gains importance when prioritizing SNPs for use as MR instruments, with SNPs of larger effect preferred because they increase statistical power provided the minor allele frequency is sufficiently high.[7] In a study with a fixed sample size, the *P*-value for the SNP-biomarker association provides an indirect measure of the effect size, but this is also influenced by the frequency at which the variant occurs in the sample, and the LD relationship between the typed variant and the causal variant (if they differ).

Specific metrics of effect size can be used to inform the selection of SNPs as instruments in an MR analysis. The most commonly used indicators of effect are: (i) the beta-coefficient from a linear regression of each additional minor allele of an SNP locus with the trait of interest, which equates to the absolute difference in concentration of biomarker for each additional allele, expressed on the native or standardized scale; (ii) the proportion of the phenotypic variance explained by the SNP in the sample (R^2); and (iii) the F-statistic from the linear regression model of the genetic instrument with the biomarker. Both R^2 and the F-statistic are influenced by the minor allele frequency, and the F-statistic is additionally affected by the sample size,[64] For the F-statistic, an arbitrary threshold value of $F > 10$ has been proposed for determining suitability of SNPs for MR analysis,[7] to avoid weak-instrument bias. However, investigators should be cautious about the use of an arbitrary F-statistic threshold for the selection of instruments, particularly where the estimate of the F-statistic comes from a single small study. As reported by Burgess and Thompson,[65] F-statistic estimates can be inflated by chance in small studies. This is because 'confounders may not be perfectly balanced between genotypic subgroups in finite samples'.[65] Under such circumstances the chance difference in confounders may explain more of the difference in the biomarker of interest between the genotypic groups than the instrument itself. As a corollary, the estimate of the causal association will be inflated towards that of the biased observational association between the biomarker and disease outcome. Since the F-statistic is related to the proportion of the variance in the biomarker explained by the genetic variants, the sample size and the number of instruments, Burgess and Thompson suggest three ways in which this effect can be mitigated: by increasing the sample size and/or by combining genotype biomarker associations across studies by meta-analysis; by increasing the number of instruments; and by adjusting for measured covariates. Many of these approaches are now routinely applied in contemporary MR analysis

Returning to the CRP example, Supplementary Figure 5 (available as Supplementary data at *IJE* online) illustrates how the choice of effect metric affects the ranking of potential SNPs that might be used as instruments in an MR analysis. In general, low-frequency variants with large effects tend to rank highly when assessed using the beta-coefficient, but diminish in priority when ranked by the proportion of variance explained (R^2) or F-statistic, because the latter penalize low allele frequency. In general, we have found the proportion of variance explained (R^2) to be the most useful metric of effect when planning SNP selection for MR analysis. For these reasons, most successful MR analyses to date have relied mainly on common variants as instruments. Rare variants are of value in other types of study design, including recall-by-genotype studies.

Though common SNPs typically explain only a small proportion of the variance in a trait, the value of R^2 should be placed in context. For example, a common SNP (rs1205) in the vicinity of the CRP gene (MAF = 0.34) explains only 0.7% of the variance in this trait, but the difference in CRP concentration per allele (beta-coefficient: -0.15mg/l log CRP) is similar in magnitude to the

difference in CRP value between treatment and control groups in a randomized clinical trial of rosuvastatin, a potent statin drug which lowers CRP in addition to its effect on blood lipids.[66]

The degree to which loci contribute to biomarker variance may also vary. For some biomarkers, a single locus may dominate (e.g. *LPA* associated with lipoprotein(a) concentration).[67,68] In other cases, gene-centric and genome-wide analyses of uric acid[29] and HDL-cholesterol[31,32,69–71] indicate that SNPs at multiple loci contribute to the variance in each trait but certain loci harbour variants of large effect [*SLC2A9* and *CETP*, respectively, (Supplementary Figure 6, available as Supplementary data at *IJE* online)]. In our experience, SNPs acting in *cis* with effects on mRNA transcription level and protein concentration are often, but not invariably, those with the largest effect[21].

Specificity of the genetic association

The assortment of alleles at the time of gamete formation is independent of environmental factors. This is why genetic variants associated with disease biomarkers generally exhibit no association with behavioural, dietary and lifestyle factors, even though the biomarkers they instrument frequently do.[72] However, we note that certain genetic variants have been identified that influence habitual behaviours such alcohol and coffee consumption or smoking.[73–75] In such cases, these associations arise not because of non-random assortment, but rather because there is a mechanistic explanation: the variants influence the expression or function of genes involved in the handling of, or response to, chemical constituents of these exposures leading to an alteration in smoking or drinking behaviour. Such variants have in fact served as useful instruments to evaluate the causal influence of such exposures on the risk of common disease[76,77] (Table 1).

A previous MR analysis of CRP (Box 1) using *cis*-acting SNPs in the CRP gene as instruments, had a particularly straightforward interpretation because variants in the gene were associated exclusively with the encoded CRP protein but none of the very wide range of other biomarkers with which CRP itself is associated.[78] Similarly, specific genotype-biomarker associations were reported in an MR analysis of fibrinogen levels using SNPs in the *FGB* gene related to fibrinogen levels.[79] However, because of the complex biological inter-relationships between the widely measured circulating biomarkers,[80] biomarker-associated SNPs rarely exhibit the degree of specificity that was fortuitously observed with SNPs in the *CRP* gene; it is more common to find that SNPs identified for an association with one biomarker are also associated with several others. Speculatively, this issue is likely to become more prominent as a wider range of biomarkers are more routinely measured using new proteomic and metabolomic technologies.

For example, variants in *LIPC* (rs4775041, chr15q21.3) are associated with both HDL-cholesterol and triglycerides,[81] and variants at the *APOE* cluster (rs4420638, chr19q13.32) with LDL-cholesterol, HDL-cholesterol, triglycerides,[69] CRP,[82] Lp(a) levels[83] and lipoprotein-associated phospholipase A2 activity.[84] At first glance, this lack of specificity might be interpreted as irrevocably violating one of the principal assumptions of the MR paradigm. For example, an association of the rs662799 SNP in the *APOA5* gene (chr11q23.3) with triglyceride level and coronary heart disease (CHD) risk[85] may be taken to indicate a causal role in CHD for triglycerides. However, since this SNP is also associated with HDL-cholesterol, it is uncertain whether the CHD association of this SNP reflects its effect on triglycerides, HDL-cholesterol or some other consequence of variation in this gene.

The mechanisms responsible for pleiotropy may be varied, but have been incompletely characterized. However, based on available understanding of genomic organization and gene regulation, established or theoretical reasons for an association of a SNP with several biomarkers include (Figure 2):

i. an effect on expression of an alternatively spliced gene leading to two distinct protein products with different actions (Figure 2a);
ii. linkage disequilibrium between SNPs spanning nearby genes at the same locus (i.e. the problem of confounding by LD) (Figure 2b);
iii. an effect in *cis* on expression and in *trans* (either directly or through LD with an adjacent SNP) on a physically distant gene mediated through chromosomal conformational effects (Figure 2c);
iv. an effect on expression of a microRNA that regulates the stability of transcripts from multiple target genes (Figure 2d);
v. an effect on expression of a transcription factor (e.g. hepatocyte nuclear factor-1α[86]) that regulates several distant target genes (Figure 2e);
vi. and residency of an SNP in a gene encoding a single protein whose activity influences several downstream biomarkers, some of which lie on the causal pathway to a disease outcome and some which may not (Figure 2f, g).

Two assumptions of an MR analysis are that the instrument should not be associated with any confounder of the biomarker disease association and that the association of the genetic instrument with the disease outcome should be mediated solely through the biomarker of interest. When a

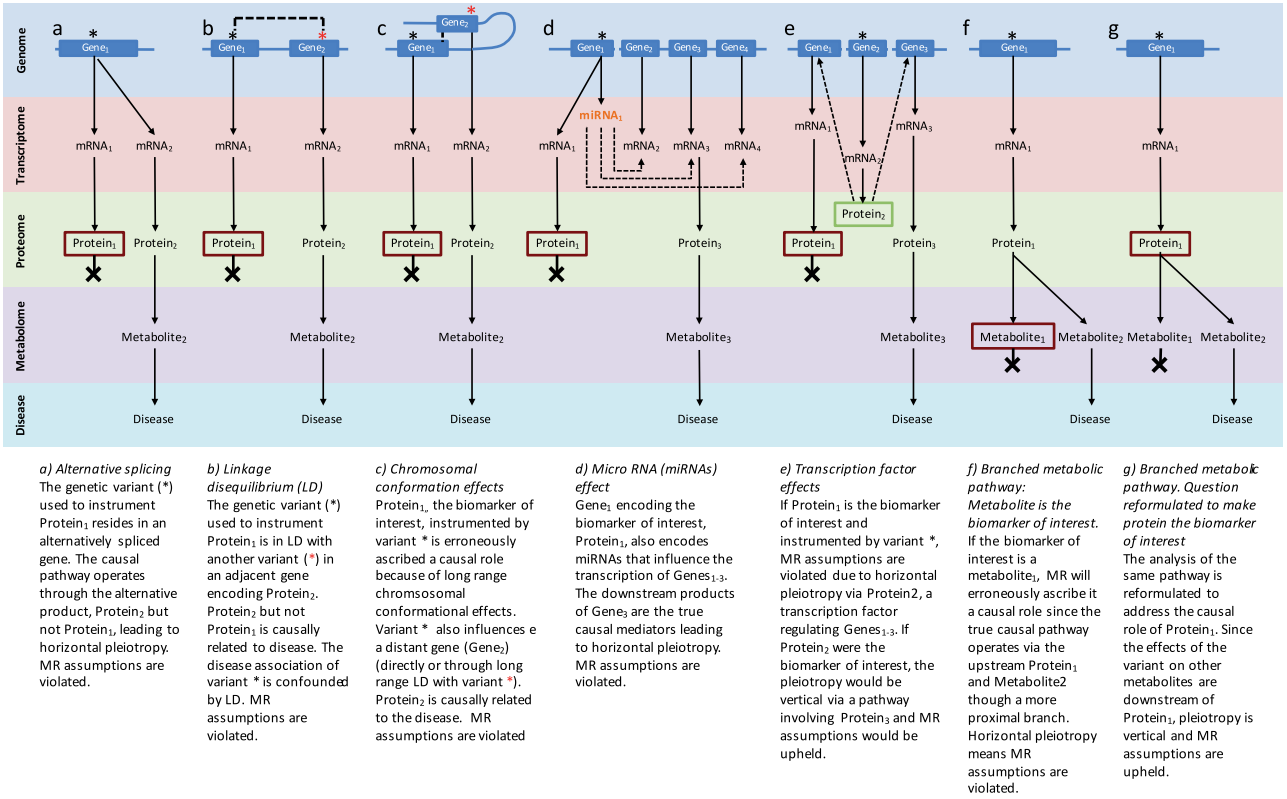

Figure 2. Mechanisms that may give rise to genetic pleiotropy and implications for MR analysis.

genetic variant is associated with several biomarkers, including the biomarker of interest, the assumptions of an MR analysis will be violated if the explanation for the pleiotropy is a disease pathway that branches more proximally to the biomarker of interest. This has been termed horizontal pleiotropy.[87] By contrast, the assumptions of an MR analysis hold if the associations of a genetic variant with several biomarkers arise because of the serial and sequential effects of the biomarker of interest on others residing more distally on the same causal pathway to disease. This has been termed vertical pleiotropy. In Figure 2, we explain in more detail how the different established and putative mechanisms listed above could give rise either to horizontal or vertical pleiotropy, and the implications for MR analysis.

Thus, when faced with a candidate instrument that exhibits genetic pleiotropy, a critical issue for MR analysis is the likelihood that this is vertical rather than horizontal in nature. Confidence in a vertical explanation for pleiotropy may be high when there is good pre-existing functional insight. For example, the association of obesity-related gene variants with a range of cardiometabolic traits has been interpreted as evidence of the causal effect of adiposity on these other risk factors.[88] In other cases, however, understanding of the functional relationships between the myriad circulating biomarkers may not be deep, and it may be difficult to exclude the possibility of horizontal pleiotropy. Moreover, the extent of horizontal pleiotropy may be underestimated because of the relatively modest number of biomarker measures that are currently available in epidemiological studies. The availability of new nuclear magnetic resonance[89] and mass spectrometry-based lipidomic and metabolomics analysis will soon allow more comprehensive assessment of horizontal pleiotropy. However, such technologies also offer the enticing prospect of ascertaining genetic instruments that instrument certain circulating biomarkers more precisely. For example, the major blood lipid fraction HDL-C actually represents the cholesterol content of a wide range of high-density lipoprotein particles which each may have a different aetiological relationship with other lipids and metabolites and with disease risk.

The variety of mechanisms by which horizontal pleiotropy may arise are diminished the closer the biomarker of interest lies (in a functional sense) to the genetic variant which is acting as the instrument, hence the importance of considering the nature of the biomarker of interest in an MR analysis. According to the central dogma of molecular biology, there is a unidirectional flow of information from

genetic sequence variation through mRNA, protein and thence through myriad downstream metabolic changes en route to disease events. In essence, invariant sequence variation in DNA can encode downstream perturbations in the transcriptome, proteome, metabolome and, in some instances, disease risk, whereas these perturbations cannot, to the best of our knowledge, alter DNA sequence. Sequence variation can therefore be envisaged as producing a series of sequential perturbations of the transcriptome, proteome and then metabolome. Proteins are the most widely measured proximal circulating biomarkers of interest for MR, separated from the genetic sequence only by mRNA. Thus, when a protein biomarker is instrumented in an MR analysis by *cis*-acting variants in the vicinity of the encoding gene, the likelihood of horizontal pleiotropy is diminished, though it is still possible (e.g. by alternative splicing of mRNA species; see Figure 2). If alternative splicing of the mRNA, the presence of a local miRNA encoding site and confounding by local and long-range LD can be reliably excluded (e.g. based on widely available, detailed, open access bioinformatic data), any pleiotropy observed of a *cis*-SNP instrumenting its encoded protein is more likely to be vertical than horizontal in origin. For this reason, MR analysis of protein biomarkers, based on *cis*-SNPs, forms a privileged category of MR analysis—which we term '*cis*-MR'. Proteins form the targets of most drugs, and several recent examples have demonstrated that variants in genes encoding a drug target mimic the mechanism-based consequences of modifying the same target pharmacologically,[90–92] confirming the validity of the assumption of vertical pleiotropy and exemplifying the utility of *cis*-MR. This observation is motivating a particular use of *cis*-MR: for drug target selection and characterization, with applications in drug development.[93] SNPs acting in *cis* could also be used as instruments to assess the causal relevance for disease of epigenetic marks such as DNA methylation[94] or an even more proximal consequence of sequence variation, mRNA level.[23]

Handling non-specific SNP associations in MR analysis of non-protein biomarkers

The lack of specificity of genetic associations poses greater difficulty when the biomarker of interest is not a protein but a more distal biomarker, for example a lipid particle (such as HDL-cholesterol) or a metabolite (e.g. uric acid). In such cases, the distinction of *cis*-SNPs from other categories of instrument is redundant. Moreover, because of limited functional understanding, it may be difficult to distinguish which of the several biomarkers associated with an SNP lies proximal to the biomarker of interest (and which could then influence disease independently of it, violating one of the MR assumptions), and which might lie distal to it on the causal pathway to disease (Figure 2). In effect, under such circumstances it can be difficult to distinguish horizontal from vertical pleiotropy. How can the problem of non-specificity of the available instruments be addressed in such situations? Three complementary approaches are considered, which harness the knowledge base of genome-wide associations with disease biomarkers or recent methodological developments.

Demonstration of the consistency of SNP-biomarker-disease associations, regardless of the genetic instrument employed

The first option is to compare the effect on disease risk of genetic variants from different locations, each exhibiting a shared association with the biomarker of interest but with a different repertoire and pattern of effects on other biomarkers. Here, causality for the biomarker of interest would be inferred from a consistent association of the different instruments with both the biomarker and the disease outcome. For example, SNPs in *LDLR*, *PCSK9*, *APOE* and *SORT1*[31–33,69,70,95,96] have a distinct repertoire of effects on other biomarkers but all associate with LDL-cholesterol and also with the risk of CHD events, in proportion to their effect of LDL-cholesterol, as carefully shown by Ference and colleagues.[97] This consistency provides strong support for the causal role of LDL-cholesterol in the pathogenesis of CHD (Supplementary figure 7, available as Supplementary data at *IJE* online). By analogy, blood pressure was confirmed to be a causal factor in CHD because the many different blood pressure-lowering drugs tested in RCTs (including diuretics, beta-blockers and calcium channel blockers) each reduced CHD risk despite different mechanisms of action and different effects on other variables such as serum potassium, glucose and uric acid.

Multi-locus approaches

A second approach, whose use has been growing,[88,98–100] is to derive a new genetic instrument that incorporates information from multiple loci. The instrument is composed of SNPs selected from across the genome on the basis of a genome-wide significant association with a trait of interest, recognizing that some may exhibit associations with additional biomarkers. The most conservative approach is to select a single, strongly associated SNP from each locus; however, approaches that incorporate several SNPs at each associated locus where these are independent of one another, to a whole genome approach, including SNPs whose associations are below genome-wide levels of significance,

have also been explored.[101] The potential benefits are 2-fold. The first is an increase in the variance in the trait of interest explained by the genetic instrument to improve the power of the MR analysis. The second is a possible dilutional effect on pleiotropy, since SNPs selected on the basis of an association with one biomarker should not systematically be associated with other biomarkers unless one or more of these is in a related biological pathway. Under those circumstances, it would not be possible to eliminate pleiotropy entirely. The stability of the causal estimate based on a multi-locus gene score, to the exclusion of subsets of SNPs drawn at random, can be used as an adjunct means to evaluate bias in the causal estimate that may arise from the potential pleiotropic influence of a subset of SNPs.

As discussed previously, SNPs associated with a particular biomarker tend to be distributed across many independent, biologically distinct loci (e.g. at least 36 loci associate with LDL-cholesterol, 47 with HDL-cholesterol, 32 with triglycerides[30] and 23 with blood pressure[26,37,38,102]). It is therefore possible to assign to each individual in a dataset a score based either on a simple count of the number of trait-raising alleles carried, or a score where the allele count is weighted by the per-allele biomarker effect size.[103] The set of SNPs used for calculating such scores should have minimal redundancy so that each SNP is independent in its trait effect, a simple approach is to select a single SNP from each locus. In theory, associations that arise because of horizontal pleiotropy at one locus should then be independent of horizontal pleiotropic effects at other loci, and these smaller, unsystematic horizontal pleiotropic associations should be diluted relative to associations with the trait of interest. Supplementary Figure 8 (available as Supplementary data at *IJE* online) illustrates this effect using gene scores for HDL-cholesterol and triglycerides in a sample of 5000 men and women in the Whitehall II study.[19] The scores were constructed using variants identified by one of the largest GWAS of lipids published to date[30] and are robustly associated with their cognate lipid fractions. In each case, the score exhibits a considerably stronger association and greater specificity than any individual SNP (Supplementary Table 2, available as Supplementary data at *IJE* online), which has also been demonstrated in an analysis of allele scores for three clinically important biomarkers.[101] A simple, unweighted score is justifiable if the component SNPs all exhibit similar effect sizes. However, where a small number of loci have a dominant effect on the trait of interest (as is the case with uric acid, for example), a weighted score may be preferable. For weighted scores, the effect size should ideally be calculated in a dataset independent from that used for the MR analysis, to reduce bias as a consequence of over-fitting.[17]

A further enhancement of the multi-locus approach has been to use information from multiple SNPs but to treat them as individual instrumental variables in a multivariable model, (see Palmer *et al.*[17]). Approaches that allow the incorporation of summary genetic effect estimates have also been developed, obviating the need to have access to participant-level data.[104,105] Techniques have also been developed to accommodate the situation where genotype-biomarker associations are available in datasets distinct from, or only partially overlapping with, those in which genotype-disease associations are estimated.[106]

Despite the attraction of multi-locus approaches, it can still prove difficult to develop a truly specific genetic instrument. For example, a multi-locus MR analysis of the causal relevance of the three major lipid fractions was unable to identify instruments that were truly specific for each lipid fraction, the development of specific instruments for HDL-cholesterol and triglycerides being particularly problematic. Approaches developed to deal with residual pleiotropy include dropping the most pleiotropic SNPs from the instrument (with a corresponding reduction in power) or adjusting for residual pleiotropy in the analysis, which requires access to participant-level data and is unsatisfying conceptually as it returns to a standard observational approach that it was hoped would be rendered unnecessary by MR analysis.[100,107,108]

A further development, referred to as multivariable MR analysis, allows for vertical or horizontal pleiotropic associations among a pre-specified, measured set of risk factors.[109–111] The assumptions of this approach are: that the genetic variants used as instruments are associated with at least one of a pre-specified set of risk factors, including the risk factor(s) of primary interest, but not with any others that might confound the association of the biomarker(s) of interest with the disease outcome; and that none has an effect on disease outcome except through the set of pre-specified risk factors. The approach has been applied to dissect the causal relevance of HDL-cholesterol and triglycerides for CHD, using summary effect estimates from previous GWAS.[107] Although a clear advance, the approach can only allow for biomarkers that have been measured in the dataset. Horizontal pleiotropy due to unmeasured biomarkers may still undermine causal interpretation as with other types of MR analysis. The approach also focuses on the causal relevance of the biomarker of interest on the disease outcome independent of other biomarkers, which may underestimate the total causal effect in the preence of vertical pleiotropy operating through another biomarker in the pre-specified set.

To address the issue of unmeasured pleiotropy, Bowden *et al.*[112] recently reported that Egger regression, originally

developed to quantify small-study bias in meta-analysis of randomized trials, can be adapted and applied to provide an unbiased estimate of the causal effect of a biomarker on disease outcome even in the presence of invalid genetic instruments. Briefly, the unbiased causal effect of a biomarker on disease outcome is estimated as the slope of the regression line from a plot of the genotype-disease against genotype-biomarker association for a set of variants selected for an association with the biomarker of interest. By contrast to the more usual two-stage least squares regression, the Egger regression line is not constrained to pass through the origin. The intercept of the line provides an estimate of the extent of unmeasured pleiotropy. The approach is attractive but suffers from a reduction in power compared with the other methods. The reader is referred to the original paper for more details. Sensitivity analysis, in which effect estimates from standard two-stage least squares instrumental variable analysis, multivariable MR analysis and MR-Egger are compared, may help better judge the causal relevance of any given biomarker. This approach is illustrated in a recent MR analysis of uric acid in CHD.[113]

Importantly, regardless of the strengths and weaknesses of each of these approaches, and bearing in mind there may be no perfect solution to the problem of pleiotropic instruments in the MR analysis of non-protein biomarkers, all approaches can be considered to be a substantial advance over non-genetic observation studies.

Reformulating the study question as a *cis*-MR analysis

A third approach to addressing pleiotropy is to reframe the research question so as to make a protein the primary 'exposure' of interest. This allows the investigator to harness the advantages of *cis*-MR. Since *cis*-acting regulatory variants in the vicinity of genes that influence mRNA and protein expression appear to be a consistent feature of the genome, the genetic tools for *cis*-MR analyses of this type should generally be available. Moreover, since more than 90% of drug targets are proteins,[114] the analysis is likely to have translational relevance, as *cis*-MR analysis has a role as a means for drug target validation. For example, a question on the causal role of HDL-cholesterol in CHD could be reformulated as: 'what is the likely therapeutic benefit of targeting a specific protein (e.g. cholesteryl ester transfer protein, CETP) that influences HDL-cholesterol concentration?' Though the causal relevance of HDL-C in CHD is not directly answered by an analysis of this type because SNPs in the *CETP* gene also influence other major blood lipids and lipoproteins,[30] these SNPs can help address the specific and important question of whether pharmacological modification of CETP to raise HDL-cholesterol will help prevent CHD events.[115,116]

A guide to the selection of instruments for MR analysis of disease-associated biomarkers

Figure 3 summarizes some of the decisions to be made in the selection of instruments for MR analyses of disease biomarkers, based on the principles described in this review. These serve as a guide, but we emphasize that each MR analysis deserves thorough consideration on a case-by-case basis, with due attention paid to any underlying biological knowledge that may inform the design, analysis, reporting and inferences drawn.

For example, variants in the *IL6R* gene are associated with directionally opposite effects on CRP and interleukin-6, which may confuse interpretation of MR analysis using such instruments to evaluate the causal relevance of these two biomarkers. Insight comes from a comparison of the effect of pharmacological interleukin-6 receptor blockade on these biomarkers. This clearly shows that such variants mimic the effect of interleukin-6 receptor blockade and are variants optimally suited to a *cis*-MR of this receptor, with application in drug development.

The motivating factor for many MR analyses is the association between a biomarker and a disease outcome detected in an observational study. The next issue is whether any genetic variant(s) associated with the biomarker of interest have been identified that might serve as an instrument in an MR analysis. If the biomarker is a protein and SNPs can be identified in the encoding gene which influence its level or function, then a single locus *cis*-MR may be possible, provided confounding by LD and horizontal pleiotropy due to alternative splicing or miRNA effects can be confidently excluded, and the effect size is sufficiently large for an adequately powered analysis.

If the biomarker is not a protein and SNPs from multiple independent loci contribute to its variance, a multi-locus multi-instrument MR analysis may be possible, but the instrument is more likely to be affected by horizontal pleiotropy. The recent methodological advance of multivariable MR-Egger analysis may help deal with this. Alternatively, it may be possible to refocus the research question on variants influencing one or more of the proteins encoded by the loci influencing the biomarker of interest, that is reformulating the question as a *cis*-MR analysis.

Regardless of the approach used, consideration should be given to maximizing the sample size through the use of

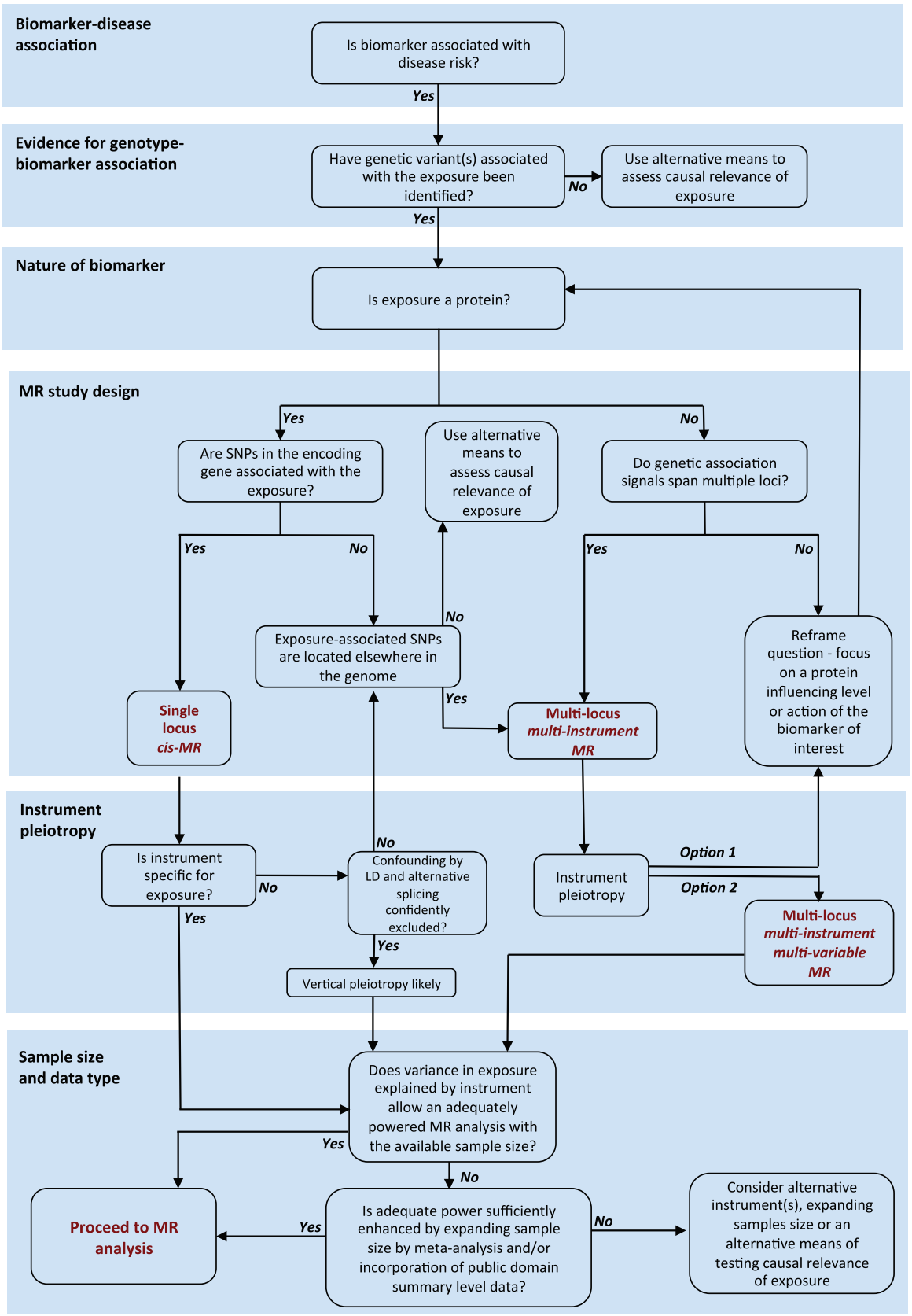

Figure 3. Illustrative guide to some of the key decisions in selecting instruments for MR analysis of disease biomarkers, based on the principles outlined in this review. The figure is intended to help plan a Mendelian randomization study of a disease-associated biomarker and should not be viewed as an inflexible decision tree. For additional considerations and details, please refer to the main text.

meta-analysis and the incorporation of public domain summary level estimates where possible.

Conclusions

MR offers novel opportunities for reliable causal inference within the framework of observational research designs. The findings can provide insight into the pathophysiology of complex disease and have translational relevance, including the prioritization of drug targets. The emerging genetic architecture of disease biomarkers now allows more informed selection of genetic variants for MR studies than was hitherto possible. As the number of biomarker-associated variants grows, selection of the most appropriate instruments for MR analysis will become an increasingly important issue. We have proposed a set of principles that should inform the selection process to aid the design, analysis and interpretation of MR studies.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

D.I.S. is supported by a National Institute of Health Research Academic Clinical Fellowship. R.S. has been supported by a British Heart Foundation (Schillingford) Clinical Training Fellowship (FS/07/011). M.V.H. has been supported by a Medical Research Council Population Health Scientist Fellowship (G0802432). M.K. is supported by the National Institute on Aging (AG034454), the Medical Research Council (K013351), the National Heart, Lung and Blood Institute (HL036310) and the NordForsk. J.P.C. and A.D.H. are supported by University College London National Institute for Health Research Biomedical Research Centre. E.J.B. is supported by a British Heart Foundation programme grant (RG/13/2/30098) and the MooDFOOD Collaborative Project (FP7 grant 613598).

Conflict of interest: D I Swerdlow has been a consultant to Pfizer for work unrelated to this paper. John C Whittaker is employed by and holds stock in GSK.

References

1. Davey Smith G, Ebrahim S. What can mendelian randomization tell us about modifiable behavioural and environmental exposures? *BMJ* 2005;**330**:107679.
2. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22.
3. Mendel G. Experiments in plant hybridisation. *Proceedings of Brunn Natural History Society, Brunn, 8 February and 8 March 1865*. Brunn, Germany: Natural History Society of Brunn, 1866.
4. Hingorani A, Humphries S. Nature's randomized trials. *Lancet* 2005;**366**:1906–08.
5. Crick F. Central dogma of molecular biology. *Nature* 1970;**227**:561–63.
6. Williamson EJ, Aitken Z, Lawrie J, Dharmage SC, Burgess JA, Forbes AB. Introduction to causal diagrams for confounder selection: Causal diagrams. *Respirology* 2014;**19**:303–11.
7. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med* 2008;**27**:1133–63.
8. Bautista LE, Smeeth L, Hingorani AD, Casas JP. Estimation of bias in nongenetic observational studies using 'mendelian triangulation'. *Ann Epidemiol* 2006;**16**:675–80.
9. Hindorff LA, Sethupathy P, Junkins HA *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits (http://www.genome.gov/gwastudies). *Proc Natl Acad Sci USA* 2009;**106**:9362–67.
10. Hindorff LA, Junkins HA, Hall P, Mehta JP, Manolio TA. *A Catalog of Published Genome-Wide Association Studies*. www.genome.gov/gwastudies (23 August 2015, date last accessed).
11. Voight BF, Kang HM, Ding J *et al*. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* 2012;**8**:e1002793.
12. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* 2011;**13**:101.
13. Little J, Khoury MJ. Mendelian randomization: a new spin or real progress? *Lancet* 2003;**362**:930–31.
14. Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004;**33**:30–42.
15. Tobin MD, Minelli C, Burton PR, Thompson JR. Commentary: development of Mendelian randomization: from hypothesis test to 'Mendelian deconfounding'. *Int J Epidemiol* 2004;**33**:26–29.
16. Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am J Epidemiol* 2006;**163**:397–403.
17. Palmer TM, Lawlor DA, Harbord RM *et al*. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat Methods Med Res* 2012;**21**:223-42.
18. Pierce BL, Ahsan H, Vanderweele TJ. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol* 2011;**40**:740-52.
19. Marmot M, Brunner E. Cohort Profile: The Whitehall II study. *Int J Epidemiol* 2005;**34**:251–56.
20. Keating BJ, Tischfield S, Murray SS *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS One*. 2008;**3**:e3583.
21. Ridker PM, Pare G, Parker A *et al*. Loci related to metabolic-syndrome pathways including LEPR, HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am J Hum Genet* 2008;**82**:1185–92.
22. Dehghan A, Dupuis J, Barbalic M *et al.* Meta-analysis of genome-wide association studies in > 80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* 2011;**123**:731–38.

23. Holmes MV, Exeter HJ, Folkersen L *et al*. Novel genetic approach to investigate the role of plasma secretory phospholipase A2 (sPLA2)-V isoenzyme in coronary heart disease: modified Mendelian randomization analysis using PLA2G5 expression levels. *Circ Cardiovasc Genet* 2014;**7**:144–50.
24. Willer CJ, Schmidt EM, Sengupta S *et al*. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;**45**:1274–83.
25. Speliotes EK, Willer CJ, Berndt SI *et al*. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010;**42**:937–48.
26. Levy D, Ehret GB, Rice K *et al*. Genome-wide association study of blood pressure and hypertension. *Nat Genet* 2009;**41**: 677–87.
27. Lange LA, Croteau-Chonka DC, Marvelle AF *et al*. Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults. *Hum Mol Genet* 2010;**19**:2050–58.
28. Kolz M, Johnson T, Sanna S *et al*. Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet* 2009;**5**:e1000504.
29. Wallace C, Newhouse SJ, Braund P *et al*. Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am J Hum Genet* 2008;**82**:139–49.
30. Teslovich TM, Musunuru K, Smith AV *et al*. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;**466**:707–13.
31. Kathiresan S, Melander O, Guiducci C *et al*. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 2008;**40**:189–97.
32. Kathiresan S, Willer CJ, Peloso GM *et al*. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 2009;**41**:56–65.
33. Sabatti C, Service SK, Hartikainen A-L *et al*. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 2009;**41**:35–46.
34. Chambers JC, Elliott P, Zabaneh D *et al*. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nat Genet* 2008;**40**:716–18.
35. Saxena R, Hivert M-F, Langenberg C *et al*. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* 2010;**42**:142–48.
36. Dupuis J, Langenberg C, Prokopenko I *et al*. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010;**42**:105–16.
37. Wang Y, O'Connell JR, McArdle PF *et al*. From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc Natl Acad Sci USA* 2009;**106**:226–31.
38. Newton-Cheh C, Johnson T, Gateva V *et al*. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 2009;**41**:666–76.
39. Croteau-Chonka DC, Marvelle AF, Lange EM *et al*. Genome-wide association study of anthropometric traits and evidence of interactions with age and study year in Filipino women. *Obesity (Silver Spring)* 2011;**19**:1019-27.
40. Liu JZ, Medland SE, Wright MJ *et al*. Genome-wide association study of height and body mass index in Australian twin families. *Twin Res Hum Genet* 2010;**13**:179–93.
41. Johansson A, Marroni F, Hayward C *et al*. Linkage and genome-wide association analysis of obesity-related phenotypes: association of weight with the MGAT1 gene. *Obesity (Silver Spring)* 2010;**18**:803–08.
42. Thorleifsson G, Walters GB, Gudbjartsson DF *et al*. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet* 2009;**41**:18–24.
43. Willer CJ, Speliotes EK, Loos RJF *et al*. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 2009;**41**:25–34.
44. Loos RJF, Lindgren CM, Li S *et al*. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet* 2008;**40**:768–75.
45. Fox CS, Heard-Costa N, Cupples LA, Dupuis J, Vasan RS, Atwood LD. Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. *BMC Med Genet* 2007;**8 (Suppl 1)**:S18.
46. Frayling TM, Timpson NJ, Weedon MN *et al*. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007;**316**:889–94.
47. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010;**6**: e1000888.
48. Small EM, Olson EN. Pervasive roles of microRNAs in cardiovascular biology. *Nature* 2011;**469**:336–42.
49. Dunham I, Kundaje A, Aldred SF *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
50. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de. Bakker PIW SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008;**24**:2938–39.
51. Goldman ND, Liu T, Lei KJ. Structural analysis of the locus containing the human C-reactive protein gene and its related pseudogene. *J Biol Chem* 1987;**262**:7001–05.
52. Szalai AJ, VanCott JL, McGhee JR, Volanakis JE, Benjamin WH Jr. Human C-reactive protein is protective against fatal Salmonella enterica serovar typhimurium infection in transgenic mice. *Infect Immun* 2000;**68**:5652–56.
53. Cao H, Hegele RA. Human C-reactive protein (CRP) 1059G/C polymorphism. *J Hum Genet* 2000;**45**:100–01.
54. Russell AI, Cunninghame Graham DS, Shepherd C *et al*. Polymorphism at the C-reactive protein locus influences gene expression and predisposes to systemic lupus erythematosus. *Hum Mol Genet* 2004;**13**:137–47.
55. Brull DJ, Serrano N, Zito F *et al*. Human CRP gene polymorphism influences CRP levels: implications for the prediction and pathogenesis of coronary heart disease. *Arterioscler Thromb Vasc Biol* 2003;**23**:2063–69.
56. Wald DS, Law M, Morris JK. Homocysteine and cardiovascular disease: evidence on causality from a meta-analysis. *BMJ* 2002;**325**:1202.

57. Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. *JAMA* 2002;**288**:2015–22.
58. Wang TJ, Larson MG, Keyes MJ, Levy D, Benjamin EJ, Vasan RS. Association of plasma natriuretic peptide levels with metabolic risk factors in ambulatory individuals. *Circulation* 2007;**115**:1345–53.
59. Ayers KL, Cordell HJ. SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol* 2010;**34**:879–91.
60. Verzilli C, Shah T, Casas JP *et al*. Bayesian meta-analysis of genetic association studies with different sets of markers. *Am J Hum Genet* 2008;**82**:859–72.
61. Giambartolomei C, Vukcevic D, Schadt EE *et al*. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 2014;**10**:e1004383.
62. McCarthy MI, Abecasis GR, Cardon LR *et al*. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008;**9**:356–69.
63. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet* 2003;**361**:865–72.
64. Park J-H, Gail MH, Weinberg CR *et al*. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A* 2011;**108**:18026–31.
65. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol* 2011;**40**:755–64.
66. Ridker PM, Danielson E, Fonseca FAH *et al*. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med* 2008;**359**:2195–207.
67. Clarke R, Peden JF, Hopewell JC *et al*. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med* 2009;**361**:2518–28.
68. Ober C, Nord AS, Thompson EE *et al*. Genome-wide association study of plasma lipoprotein(a) levels identifies multiple genes on chromosome 6q. *J Lipid Res* 2009;**50**:798–806.
69. Aulchenko YS, Ripatti S, Lindqvist I *et al*. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* 2009;**41**:47–55.
70. Waterworth DM, Ricketts SL, Song K *et al*. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* 2010;**30**:2264–76.
71. Ridker PM, Paré G, Parker AN, Zee RYL, Miletich JP, Chasman DI. Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among 18 245 initially healthy women from the Women's Genome Health Study. *Circ Cardiovasc Genet* 2009;**2**:26–33.
72. Lewis SJ, Davey Smith G. Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiol Biomarkers Prev* 2005;**14**:1967–71.
73. Liu JZ, Tozzi F, Waterworth DM *et al*. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010;**42**:436–40.
74. Cornelis MC, Monda KL, Yu K *et al*. Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption. *PLoS Genet* 2011;**7**:e1002033.
75. Frank J, Cichon S, Treutlein J *et al*. Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. *Addict Biol* 2012;**17**:171–80.
76. Holmes MV, Dale CE, Zuccolo L *et al*. Association between alcohol and cardiovascular disease: Mendelian randomization analysis based on individual participant data. *BMJ* 2014;**349**:g4164.
77. Rode L, Bojesen SE, Weischer M, Nordestgaard BG. High tobacco consumption is causally associated with increased all-cause mortality in a general population sample of 55 568 individuals, but not with short telomeres: a Mendelian randomization study. *Int J Epidemiol* 2014;**43**:1473-83.
78. CRP CHD Genetics Collaboration. Collaborative pooled analysis of data on C-reactive protein gene variants and coronary disease: judging causality by Mendelian randomization. *Eur J Epidemiol* 2008;**23**:531–40.
79. Keavney B, Danesh J, Parish S *et al*. Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'. *Int J Epidemiol* 2006;**35**:935–43.
80. Visscher PM, Montgomery GW. Genome-wide association studies and human disease: from trickle to flood. *JAMA* 2009;**302**:2028–29.
81. Willer CJ, Sanna S, Jackson AU *et al*. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 2008;**40**:161–69.
82. Elliott P, Chambers JC, Zhang W *et al*. Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA* 2009;**302**:37–48.
83. Khan TA, Shah T, Prieto D *et al*. Apolipoprotein E genotype, cardiovascular biomarkers and risk of stroke: systematic review and meta-analysis of 14,015 stroke cases and pooled analysis of primary biomarker data from up to 60,883 individuals. *Int J Epidemiol* 2013;**42**:475–92.
84. Suchindran S, Rivedal D, Guyton JR *et al*. Genome-wide association study of Lp-PLA(2) activity and mass in the Framingham Heart Study. *PLoS Genet* 2010;**6**:e1000928.
85. Sarwar N, Sandhu MS, Ricketts SL *et al*. Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *Lancet* 2010;**375**:1634–39.
86. Armendariz AD, Krauss RM. Hepatic nuclear factor 1-alpha: inflammation, genetics, and atherosclerosis. *Curr Opin Lipidol* 2009;**20**:106–11.
87. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014;**23**:R89-98.
88. Holmes MV, Lange LA, Palmer T *et al*. Causal effects of body mass index on cardiometabolic traits and events: a mendelian randomization analysis. *Am J Hum Genet* 2014;**94**:198-208.
89. Würtz P, Kangas AJ, Soininen P *et al*. Lipoprotein subclass profiling reveals pleiotropy in the genetic variants of lipid risk factors for coronary heart disease: a note on Mendelian randomization studies. *J Am Coll Cardiol* 2013;**62**:1906–08.
90. Swerdlow DI, Holmes MV , Kuchenbaecker KB *et al*. The interleukin-6 receptor as a potential target for coronary heart

disease prevention: evaluation using Mendelian randomization. *Lancet* 2012;**379:**1214-24.
91. Swerdlow DI, Preiss D, Kuchenbaecker KB *et al*. HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomized trials. *Lancet* 2015;**385:**351-61.
92. Holmes MV, Simon T, Exeter HJ *et al*. Secretory Phospholipase A2-IIA and Cardiovascular Disease: a Mendelian randomization study. *J Am Coll Cardiol* 2013;**62:**1966-76.
93. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov* 2013;**12**:581–94.
94. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol* 2012;**41**:161–76.
95. Erdmann J, Grosshennig A, Braund PS *et al*. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet* 2009;**41**:280–82.
96. Samani NJ, Erdmann J, Hall AS *et al*. Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007 2;**357**:443–53.
97. Ference BA, Yoo W, Alesh I *et al*. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: *a Mendelian randomization analysis. J Am Coll Cardiol* 2012;**60**:2631–39.
98. Ripatti S, Tikkanen E, Orho-Melander M *et al*. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet* 2010;**376**:1393–400.
99. Elks CE, Loos RJF, Sharp SJ *et al*. Genetic markers of adult obesity risk are associated with greater early infancy weight gain and growth. *PLoS Med* 2010;**7**:e1000284.
100. Holmes MV, Asselbergs FW, Palmer TM *et al*. Mendelian randomization of blood lipids for coronary heart disease. *Eur Heart J*. 2015;**36:**539-50
101. Evans DM, Brion MJA, Paternoster L *et al*. Mining the human phenome using allelic scores that index biological intermediates. *PLoS Genet* 2013;**9**:e1003919.
102. Levy D, Larson MG, Benjamin EJ *et al*. Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Med Genet* 2007;**8(Suppl 1)**:S3.
103. Burgess S, Thompson SG. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol* 2013;**42**:1134–44.
104. Voight BF, Peloso GM, Orho-Melander M *et al*. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomization study. *Lancet* 2012;**380:**572-80.105.
105. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* 2013;**37**:658–65.
106. Pierce BL, Burgess S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am J Epidemiol* 2013;**178**:1177–84.
107. Do R, Willer CJ, Schmidt EM *et al*. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet* 2013;**45**:1345–52.
108. Johnson T. *Efficient Calculation for Multi-SNP Genetic Risk Scores*. Stevenage, UK: Glaxosmithkline, 2012.
109. Burgess S, Freitag DF, Khan H, Gorman DN, Thompson SG. Using multivariable mendelian randomization to disentangle the causal effects of lipid fractions. *PLoS One* 2014;**9**:e108891.
110. Burgess S, Thompson SG. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* 2015;**181**:251–60.
111. Burgess S, Dudbridge F, Thompson SG. Re: 'Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects'. *Am KJ Epidemiol* 2015;**181**:290–91.
112. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015;**44**:512–25.
113. White J, Sofat R, Hemani G *et al*. Plasma urate concentration and risk of coronary heart disease: a Mendelian randomization analysis. *Lancet Diabetes Endocrinol* 2016;**4:**327-36.
114. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* 2006;**5**:821–34.
115. Brousseau ME, Schaefer EJ, Wolfe ML *et al*. Effects of an inhibitor of cholesteryl ester transfer protein on HDL cholesterol. *N Engl J Med* 2004;**350**:1505–15.
116. Cannon CP, Shah S, Dansky HM *et al*. Safety of anacetrapib in patients with or at high risk for coronary heart disease. *N Engl J Med* 2010;**363**:2406–15.
117. Bech BH, Autrup H, Nohr EA, Henriksen TB, Olsen J. Stillbirth and slow metabolizers of caffeine: comparison by genotypes. *Int J Epidemiol* 2006;**35**:948–53.
118. Sofat R, Hingorani AD, Smeeth L *et al*. Separating the mechanism-based and off-target actions of cholesteryl ester transfer protein inhibitors with CETP gene polymorphisms. *Circulation* 2010;**121**:52–62.
119. Rasmussen-Torvik LJ, Li M, Kao WH *et al*. Association of a fasting glucose genetic risk score with subclinical atherosclerosis: The Atherosclerosis Risk in Communities (ARIC) study. *Diabetes* 2011;**60**:331–35.
120. Giltay EJ, van Reedt Dortland AKB, Nissinen A *et al*. Serum cholesterol, apolipoprotein E genotype and depressive symptoms in elderly European men: the FINE study. *J Affect Disord* 2009;**115**:471–77.
121. Lim LS, Tai E-S, Aung T *et al*. Relation of age-related cataract with obesity and obesity genes in an Asian population. *Am J Epidemiol* 2009;**169**:1267–74.
122. Linsel-Nitschke P, Götz A, Erdmann J *et al*. Lifelong redu_ction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease - a Mendelian randomization study. *PLoS One* 2008;**3**:e2986.
123. Perry JRB, Ferrucci L, Bandinelli S, *et al*. Circulating beta-carotene levels and type 2 diabetes-cause or effect? *Diabetologia*. 2009 Oct;**52**(10):2117–2121.
124. Trompet S, Jukema JW, Katan MB, *et al*. Apolipoprotein e genotype, plasma cholesterol, and cancer: a Mendelian randomization study. *Am J Epidemiol*. 2009 Dec **1;170**(11): 1415–1421.
125. Casas JP, Shah T, Cooper J, *et al*. Insight into the nature of the CRP-coronary event association using Mendelian randomization. *Int J Epidemiol*. 2006 Aug;**35**(4):922–931.