CHAPTER TITLE: Human genomics and drug development

Amand F Schmidt [abc], Aroon D Hingorani [abd], Chris Finan [abcd]

a. Institute of Cardiovascular Science, Faculty of Population Health, University College London, London WC1E 6BT, United Kingdom;

b. UCL British Heart Foundation Research Accelerator;

c. Department of Cardiology, Division Heart and Lungs, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands

d. Health Data Research UK, London

Abstract

Insights into the genetic basis of human disease are helping to address some of the key challenges in new drug development including the very high rates of failure. Here we review the recent history of an emerging, genomics-assisted approach to pharmaceutical research and development, and its relationship to Mendelian randomization (MR), a well-established analytical approach to causal inference. We demonstrate how human genomic data linked to pharmaceutically relevant phenotypes can be used for (a) drug target *identification* (mapping relevant drug targets to diseases), (b) drug target *validation* (inferring the likely effects of drug target perturbation), (c) evaluation of the effectiveness and specificity of *compound-target engagement* (inferring the extent to which the effects of a compound are exclusive to the target and distinguishing between on-target and off-target compound effects), and (d) the selection of end points in clinical trials (the diseases or conditions to be evaluated as trial outcomes). We illustrate how genomics can help identify indication expansion opportunities for licensed drugs and repurposing of compounds developed to clinical phase that proved safe but ineffective for the original intended indication. We outline statistical and biological considerations in using MR for drug target validation (drug target MR) and discuss the obstacles and challenges for scaled applications of these genomics-based approaches.

Genomics-led drug development in context

The modern era of drug development can be traced back to observations that certain natural substances from plants or animals had beneficial effects on the human body, which could be harnessed to treat illness[1]. For example, willow bark extracts had been noted to reduce inflammation as early as the 5th century BC[1]. With advances in organic chemistry came the ability to extract and modify active molecular entities

and use knowledge of their structure to synthesise related chemicals with similar properties[2]. For example, the active ingredient in willow bark extract was eventually shown to be salicylic acid, and it was the acetyl salt of salicylic acid that was developed as aspirin by Bayer. Indeed, much of the early expertise in organic chemistry emerged from German and Swiss companies that applied their knowledge to the development of compounds with medicinal properties.

With the ability to synthesise large numbers of compounds and test them for biological activity in cell, tissue, and organ-based laboratory models of human disease, came the era of 'phenotypic' screens for drug development[3]. Phenotypic screening investigated the impact of compounds on cells, tissues or model organisms and selected efficacious compounds based on their ability to alter some aspect of biochemistry, physiology or pathological process in the model. In this approach a compound could be pursued as a drug despite incomplete knowledge of the molecular basis of its action. Many licensed drugs emerged from this approach and it continues to the present day, recently leading to the development of memantine for Alzheimer's disease and levetiracetam and zonisamide for epilepsy[3]. However, if knowledge of mechanism of action and the therapeutic target remains elusive, it is difficult to anticipate the full repertoire of drug effects and the opportunity to develop improved compounds based on mechanistic knowledge is also constrained.

With increased understanding of the molecular basis of disease, efforts have gradually shifted towards a more *target-based* approach to drug development [4,5]. By contrast with the phenotypic approach to drug development, a potential drug target is selected based on prior evidence that implicates it in the disease process. The target-based approach contrasts with the phenotypic approach by working from the potential mechanism towards the disease as opposed to starting with the disease and working back to a mechanism. The target-based approach to drug development was inspired by the receptor theory of drug action, growth of pharmacology as a scientific discipline, and an expanding knowledge base on the function of key protein classes, particularly receptors and their natural substrates and ligands. Many targets of this type formed the basis for what has also been referred to as rational drug development; with notable examples including beta-adrenoceptor blockers and

agonists and histamine H_1- and H_2-receptor blockers. With these advances came the growing recognition that proteins, encoded by the genome, are not only the major proximal effector molecules in biology, but also comprise the major category of drug targets [6].

Target-based drug development and high rates of drug development failure

In the target-based model, drug development starts by attempting to identify proteins, 'drug targets', causally linked to a disease and whose function is amenable to the therapeutic action of small molecule drugs or monoclonal antibodies[7]. Until recently, programmes based on this paradigm continued to be shaped by proof-of-concept laboratory experiments in (cell, tissue, and animal) disease models focussing on a small set of targets prioritised based on prior knowledge, with the aim of building evidence of a causal relationship between the target and the disease. Successful drugs continue to be developed using this approach, but there is a growing recognition that the process has been inefficient. Numerous reviews of the field have concluded that success rates in target-based drug development remain extremely low. Fewer than 5 in 100 initiated drug development programmes yield a licenced drug; 90% of randomised clinical trials fail; and around a half to two thirds of such failures are due to lack of efficacy in the disease of interest [8–12].

It has become apparent that the problem of late-stage failure in target-based drug development has its roots in the poor predictive utility of laboratory models and observational (nonrandomized) studies for human disease pathogenesis. Work in isolated systems (cells and tissues *ex vivo*) may not be representative of the situation in the whole organism; moreover, work in animal models may not be representative of human pathophysiology[13]. Human observational studies (though set in the right organism) can be affected by confounding and reverse causation. This leads to a high false discovery rate that permeates through (pre-) clinical science, resulting in pursuing drug targets and related compounds with a high probability of failure, increasing the overall cost of drug development if these failures occur at late-stage clinical phase testing[13]. Developing a solution to the problem of high rates of drug development failure has therefore become both a scientific challenge and an economic imperative.

Genome wide association studies for drug target identification

The sequencing of the human genome and identification of all protein coding genes now provides both a comprehensive list of potential key effectors in disease biology as well as potential therapeutic targets. With the development of comprehensive maps of human genetic sequence variation has come the ability to undertake genome-wide association studies (GWAS) in patients and populations. GWAS test relationships between natural sequence variation (genotype) and biomarkers (quantitative biological traits e.g. blood pressure or circulating metabolite concentration) relevant to disease or to clinical end points (phenotype) [14–16]. Proteins mediate the effect of both genetic variation (according to Crick's Central Dogma)[17] and drug action, and independent variation in the genome is inherited at random (according to Mendel's Laws), much like treatment allocation in a clinical trial. Thus, the concept has emerged that variants in a gene encoding a drug target, that alter its expression or function, might be used as a tool to anticipate the effect of drug action on the same target (Figure 1).

This paradigm makes it feasible, for the first time, to match drug targets to disease end points (target identification) systematically and robustly through GWAS. For example, Finan, Gaulton and colleagues showed that GWAS frequently 'rediscover' genes encoding the targets of licensed drugs for the same disease[6]. GWAS test common variation across all genes against a single phenotype, with stringent control over the false positive rate and a philosophy of replication of positive findings to reduce the risk of false discoveries. Phenome wide association analysis (PheWAS – in which associations from a single gene with many diseases and biomarkers are explored[18]) complement GWAS by helping to anticipate the mechanism-based adverse effects of drug action beyond the primary disease indication (target validation). Integrating evidence from GWAS of the proteome, metabolome, physiological and imaging data with disease GWAS can also help map the mediating pathways to disease (Figure 1) and identify biomarkers of target efficacy that can be used as a proxy to evaluate the effects of first-in-class compounds on target engagement and specificity (compound validation). The principle used here is that, in the absence of horizontal pleiotropy, a specific compound with no off-target effects should share the same pattern of effects on biomarkers as variants in a gene encoding the corresponding drug target that affect its expression or activity. An

example of this consistency was demonstrated by Wurtz et al., in the comparison of the effect of variants in the HMGCR gene and the effect of taking a statin on circulating metabolites and lipoproteins[19].

As the number of GWAS on common diseases has grown since 2007, it has become apparent that human genetics can provide a rich resource for the identification of genes encoding potential drug targets matched to the diseases, with potential to increase the success rate of drug development[6]. At least four lines of evidence support this proposition: (1) human genetics has consistently 'rediscovered' known drug target-disease indication pairings across a wide range of disease areas[6]; (2) analysis of drug approvals have shown that approval rates were doubled among drug-indication pairs for which, in retrospect, there was prior genetic support; and (3) there are emerging examples of successful development programmes being seeded by human genetic information e.g. *PCSK9* and *ANGPLT3* for the treatment of coronary heart disease as well as maraviroc for the treatment of HIV disease[20].

The importance of human genetic evidence in informing numerous aspects of drug development is increasingly reflected in the emphasis given to human genetic target validation by the pharmaceutical industry[21].

Limitations of using GWAS as a resource for identification of drug targets to treat disease

Although there are huge opportunities from the use of GWAS as a resource for human drug target identification several challenges are also recognized, some of which have seeded new research initiatives and methodological developments.

Breadth and depth of GWAS studies

One key limitation is that GWAS have yet to be fully exploited as a resource for drug target identification. There are perhaps 10,000 complex diseases (those with both a polygenic and environmental contribution) but approximately only a few hundred have been studied in GWAS[13]. Moreover, although there are many large meta-GWAS studies in common disease (e.g., AFgen, CardiogramplusC4D, or Diagram), sample sizes for many conditions remain modest, limiting the ability to robustly detect the full spectrum of genetic loci influencing a given disorder. This shortfall is

likely to be addressed by many growing initiatives around the globe to undertake GWAS in large prospective cohort studies and national biobanks linked to research-based measures (e.g., biomarkers, proteomics and metabolomics, and imaging data acquisition) as well as through linkage to primary and secondary care health records. Successful examples include UK, Estonian and Japanese Biobanks, as well as the FinnGen study. Comparable genomics initiatives have also been set within healthcare systems including the Million Veterans Programme and Geisinger Health.

GWAS identify genomic regions not causal genes

Currently, most drug targets are proteins and for target identification the genes mediating the association need to be correctly identified (causal genes). Many variants identified through GWAS are located in the vicinity of coding sequences, presumably in regulatory regions. Assignment of the causal gene in the region of a GWAS association can therefore be difficult, particularly when linkage disequilibrium (LD) between variants displaying disease associations in a region extend across several genes. Therefore, complementary approaches have been developed[22] to help prioritise the likely causal gene in a region identified by a GWAS based on 1) fine mapping using dense genotyping arrays; 2) transancestry association studies exploiting smaller LD blocks in certain non-European populations; 3) statistical and pathway methods for gene prioritization (e.g. GoSHIFTER[23], Prix Fixe[24], and FUMA[25]), 4) colocalisation of disease associated signals with mRNA or protein expression (e.g. Coloc[26], Enloc[27], Ecaviar[28]); and 5) incorporation of functional annotation methods (e.g. STOPGAP[29]).

Orthogonal information for gene prioritization from GWAS can also come from comparisons with genes known to be responsible for monogenic disorders that display a similar phenotype to the complex condition being studied. Sequence variation responsible for a monogenic disease more typically resides in the coding rather than the regulatory region of the responsible gene so there is less ambiguity about the location of the causal gene than for a GWAS of a common disorder[30,31]. Assignment of a causal gene may also be aided by comparison with information on sets of highly curated gene functions e.g., genes involved in metabolic processes; or from comparisons with genes encoding drug targets for compounds licensed for the same disease indication as the GWAS. Despite a widely perceived difficulty in

assigning a causal gene from a GWAS signal, a recent survey of different approaches for gene prioritisation from GWAS came to the conclusion that the proximity of a gene to the lead variant was the strongest predictor of the causal gene[32].

Not all genes encode druggable proteins

After identifying a gene that is associated with a therapeutically relevant disease trait using GWAS, a subsequent challenge is understanding if the encoded protein might be amenable to drug development (i.e. “druggable”).

The current Ensembl/Havana35 annotated human genome (GRCh38 v103) contains 22,940 protein coding genes with UniProtKB (v2021_01) SwissProt (manually reviewed entries) containing 20,396 human proteins and TrEMBL (unreviewed) contains 174,126 proteins. Several attempts have been made to define the druggable portion of the human genome[6]: the subset of protein coding genes that encode drugged or potentially druggable proteins, so designated because they share sequence, structural or functional properties with previously drugged proteins. Progressive iterations have seen an expansion in the number of proteins that have been included within the druggable set as more targets yield to drug development. In the latest iteration of the druggable genome, Finan *et al*. included potential targets for monoclonal antibodies for the first time (based on proteins that are secreted or which are targeted to the cell membrane) and removed olfactory receptors and phosphatases based on known difficulties targeting them, taking the set up to 4479 proteins. However, novel targeting modalities such as proteolysis targeting chimeras (ProTacs)[33] which work by marking the target protein for degradation via E3 ligase and the development of antisense technologies to target mRNA species not just proteins, will eventually widen the druggable component of the genome even further[34].

Tractability of an identified drug target

In addition to druggability, a further consideration of the tractability of a potential drug target relates to consideration of the tissues in which the target is expressed, the relevance or not of that tissue to the disease process and whether the therapeutic modality chosen renders the target accessible to the drug. For example, Mendelian

randomization (MR) analyses have suggested that targeting PCSK9 to reduce LDL-cholesterol could lead to a mechanism-based increase in the risk of type 2 diabetes[35,36] and Alzheimer's disease[37,38]. However, these effects might not be observed in practice if PCSK9 is targeted by a monoclonal antibody therapeutic, which may not gain access[6,13,38,39] to the beta cell or the central nervous system, where these effects are possibly mediated.

Therapeutic area

While human genetics for drug development is generally applicable to many therapeutic areas, there are some important exceptions or limitations. Given that genetic target validation of the type described utilises germline genetic variation, it is likely to be more limited for cancer than for other diseases, since many cancers are driven by somatic and not germline mutations. Nevertheless, findings from GWAS of prostate and breast cancer have identified variants in the genes encoding the androgen and oestrogen receptors respectively, both of which are known to be effective therapeutic targets in these diseases, suggesting that GWAS can still play a part in cancer therapeutics where the mechanism of disease initiation is distinct from the treatment[40,41]. GWAS of human subjects is also not capable of identifying targets in an infectious disease pathogen, but may still find use in identifying host response genes that might make a substantial contribution to the infectious disease process[42,43]. Additionally, developmental diseases or those where the pathology is caused by irreversible damage prior to the presentation of disease, for example type 1 diabetes, are less likely to be amenable to target identification through GWAS unless leveraging age-specific data of sufficient sample size.

Mechanistic considerations

Most drugs act by modifying the function of a protein target rather than its level[44], but most variants identified by GWAS are intergenic, rather than protein coding[45]. Nevertheless, genetically validating clinically used drug target/disease mechanisms (positive control examples) have shown that even such variants associated with gene or protein expression (regulatory variants) can model the effects of licensed drugs (see Table[13]).

Surveys of GWAS rediscovery of clinically used drug target/disease parings have indicated such studies are capable of identifying the targets for a variety of drug mechanisms including inhibitors and antagonists, agonists and activators, as well as positive and negative allosteric modulators[6,46]. However, the majority (69%; ChEMBL v27) of drug target/mechanism pairs are the subject of inhibitory or blocking drugs because the effects of activators or agonists tend to be affected by down-regulation of the target[44]. Genetic variants that increase disease risk via reduced expression or function of the encoded protein would naturally point to development of an agonist or activator of the target protein to achieve the desired therapeutic effect, which might be less tractable pharmacologically. In such cases, if the target of interest is inactivated by a second protein and the inactivator is druggable, a therapeutic option might then be to develop an inhibitor of the inactivating protein.

The failure to find significant genetic associations in or near a potential drug target for a disease also does not necessarily exclude the target from consideration[47]. One reason is that the effect of altering the expression or function of a target may only be seen beyond some threshold, such that a genetic variant of weak effect may not adequately model the effect of targeting the protein with a potent drug. The availability of common (weak effect) and rare (large effect) genetic variants in the same gene, that allows the construction of an allelic series (effectively a genetic dose-response curve[39]), may go some way toward mitigating this possibility in specific cases[48]. Another potential cause for not finding a genetic signal in or near a drug target encoding gene may be found in the fact that genetic influences on protein expression or activity are often present from early life. Such early and consistent effect may entrain developmental adaptive compensation (canalization[49]) through changes in other pathways that mitigate any biologically adverse effect on the system as a whole. Furthermore, by design, GWAS minimize the false positive rate (type 1 errors) ensuring that most finding are true positives. At the same time, stringent control of the type 1 error rate increases the false negative (type 2 errors) rate many fold, making GWAS a very poor design to show the absence of an effect[50]. Thus, the lack of genome-wide significant findings of variants in a gene encoding a drug target of interest in a particular disease need not exclude it as a therapeutic target. Next, we discuss how “drug target Mendelian randomization” can

be used to provide a more focused analysis of the likely therapeutic consequence of on-target action of a drug.

Mendelian randomization and drug target validation

Having prioritized a protein as likely causal for a disease, and having confirmed the protein druggable, an essential next step is to gather evidence on the likely range of on-target effects of pharmacologically perturbing the target, a process we refer to as *drug target validation.* Drug target validation has been traditionally addressed in preclinical studies involving cell and tissue experiments, as well as animal models. Additionally, if a potential target can be measured directly in humans, or its activity or expression inferred by measurement of downstream biomarkers, another aspect of target validation might involve assessing the association of the protein or related biomarkers with disease incidence in non-randomized (i.e., observational) studies.

However, the same issues that compromise target identification in cell, tissue and animal models also affect target validation. Moreover, while non-randomized human observational studies on a biomarker association with disease (e.g HDL-C with coronary disease) may provide a start point for a therapeutic hypothesis, most drug targets (e.g., cholesteryl ester transfer protein or lipoprotein lipase) rarely affect a single biomarker. Therefore, a one-to-one assignment of biomarkers to drug targets may lead to oversimplistic, and *overoptimistic*, models of the potential relevance of a specific drug target to a disease. Furthermore, non-randomized studies are also often affected by reverse causation, residual and unmeasured confounding[51], and/or selection bias[52], potentially compromising the inferences that can be drawn from such studies for drug development. Next, we discuss how Mendelian randomization (MR) can address some of the key limitations of preclinical science for drug target validation, and how performing MR analyses where a drug target is the exposure of interest, can strengthen the existing (cumulative) evidence base necessary to make robust choices on which target to pursue for further development.

MR was developed as a novel research design for investigating causal relationships between risk factors and health outcomes using non-randomized data; outside of genetics this is often referred to as instrumental variable analysis[53] or as a quasi-randomized experiment, for example in pharamacoepidemiology[54]. The premise is to

identify one or multiple “instrumental variables” (e.g. genetic variants) that are strongly related to an exposure of interest (e.g., LDL-C), that affect the outcome (e.g. coronary heart disease; CHD) exclusively through the exposure of interest (the exclusion restriction assumption), and that do not share a common cause with the outcome of interest, which would otherwise lead to confounding[49]. This is illustrated in Figure 2, where the risk factor is depicted simply as $\mathbf{X}$, the outcome as $\mathbf{D}$, with their relationship confounded by $\mathbf{U}$ (which may represent multiple common causes), and $\boldsymbol{G}$ (representing a genetic variant or variants) may act as an instrument if $\phi_G = 0$.

By defining confounding as a common cause between an exposure and an outcome, it becomes clear that most genetic associations with traits are relatively protected against this source of bias. While there will be many factors that influence a disease or quantitative trait, hardly any of these factors will, in turn, influence the assignment of one’s genotype. Nevertheless, confounding bias may still occur by for example, inadvertently mixing ethnicities with distinct genotypes within the same study. Because, through shared environment, ethnicities may also differ in disease risk or quantitative trait levels, admixed populations may result in confounding or “population stratification” bias. By reducing the number of potential confounding factors from up to infinity to a much smaller number of ethnic (and familial) variables, genetic analyses can focus on accounting for these specific sources of bias, either analytically, or by design (for example leveraging within-family designs[55]) and produce a result that is often robust to any remaining – that is residual – confounding. Similarly, unlike the directly observed associations between the risk factor and disease, the *genetic associations* are protected from reverse causation, since the presence of the disease does not alter the sequence of the germline. Whether the presence of disease modifies the genetic association with another outcome is, of course, more plausible and the topic of ongoing research [56].

Thus, these relatively robust genetic associations can be used to obtain inference on the effects protein perturbation might have on disease by collecting (aggregated) data on the genetic association with: 1) a relevant drug target related exposure (e.g concentration of the protein forming the drug target, expression of mRNA encoding the target, or a downstream complex biomarker known to be affected by the protein

of interest), and 2) an outcome of interest (e.g., disease incidence or a quantitative trait such as cholesterol) . As an example, PCSK9 was first considered as a drug target after finding that mutations in *PCSK9* were associated with a lower LDL-C concentration (15% reduction), as well as a decreased risk of CHD; hazard ratio 0.50 (95%CI: 0.32 to 0.79)[57].

In the following section, we define "drug target MR" more formally and introduce important inferential considerations, in part based on reference[58].

Genetic weights and the inferential target in drug target MR

To emphasize, the cited *PCSK9* estimates for LDL-C and CHD pertain to the effect of a hypothetical change in one's genotype, which may be (partially) mediated by the encoded protein. For genetic associations to inform drug target validation, we need to be willing to assume that the effect of a variant in a gene (e.g., *PCSK9*) only acts on a disease end-point (e.g. CHD) through its effect on the encoded protein, PCSK9; i.e., assume the absence of horizontal pleiotropy (the exclusion restriction assumption).

Before discussing the PCSK9 example further, we will first formally define our inferential target. Let us denote a single, or multiple, genetic variants (e.g., in *PCSK9*) as $\boldsymbol{G}$, the encoded protein $\boldsymbol{P}$ as the drug target we aim to validate (e.g., PCSK9), an outcome $\boldsymbol{D}$ such as CHD, and an intermediate biomarker $\boldsymbol{X}$ (e.g. LDL-C) potentially affected by $\boldsymbol{P}$ (left diagram of Figure 2)**.** Here the arrows indicate the direction of effect, and the edge labels the effect magnitude in a relevant unit (e.g., as mmol/L for LDL-C or hazard for CHD); we note that effect magnitudes may include zero for no effect. Furthermore, as we discuss below, genetic variants are typically preferentially selected from within or near a protein coding gene of interest acting in *cis.*

Using this diagram, we can formally define the inferential target as "the effect a unit change in protein concentration or activity has on an outcome", and label this as: $\omega = \phi_{\boldsymbol{P}} + \mu\theta$, which itself consists of the following biologically relevant estimates:

- Any *direct effect* the protein has on the outcome, sidestepping biomarker X: ϕ_P.
- The protein effect on the biomarker: μ.
- The biomarker effect on the outcome: θ.

The genetic effect on an outcome such as CHD: $\phi_G + \tilde{\delta}(\phi_P + \mu\theta) = \phi_G + \tilde{\delta}\omega$, is distinct from ω. In the absence of any horizontal pleiotropy via pathways that might occur proximal to protein-translation (e.g., $\phi_G = 0$, referred to as pre-translational pleiotropy[58]) this expression simplifies to $\tilde{\delta}\omega$ which still does not equal ω. Nevertheless, under this no-pre-translational pleiotropy assumption, the genetic association with an outcome such as CHD (e.g., from a GWAS), that is $\tilde{\delta}\omega$, provides clear evidence that the protein likely affects CHD. For example, if $\phi_G = 0$, but $\tilde{\delta}\omega \neq 0$ this implies that $\omega \neq 0$, and hence under the same MR assumptions genetic analyses may provide *indirect* evidence on a drug targets effect on disease.

In this same setting, drug target MR can be used to additional determine the effect direction of ω. Further, if we are willing to assume complete linearity of the drug target effect(s) on disease, as well as strict homogeneity in drug target effect(s) among current and future subjects, we can also obtain an estimate of ω. Specifically, we take the genetic effect on the outcome and the genetic effect on the protein drug target:

$$\frac{\tilde{\delta}(\phi_P + \mu\theta)}{\tilde{\delta}} = \phi_P + \mu\theta, \quad (eq\ 1)$$

$$= \omega.$$

In the absence of linearity and homogeneity, which often may be unlikely, ω represents an *average* effect, and (as we discuss below) might provide robust evidence on the presence of an effect and its direction.

As discussed below, estimates of associations between genetic variants and protein concentration are becoming more widely available. However, in the absence of a direct (and sufficiently strong) genetic association with of P (e.g., PCSK9), drug target MR can utilize the genetic association with X (e.g., LDL-C) as a proxy for

protein concentration and activity. In such a case the denominator in equation 1 includes μ and we are left with:

$$\omega_{bw} = \frac{\tilde{\delta}\left(\phi_P + \mu\theta\right)}{\tilde{\delta}\mu} = \frac{\phi_P + \mu\theta}{\mu}, \quad (eq\ 2)$$

$$= \frac{1}{\mu} \times \omega$$

Hence a biomarker weighted drug target MR analysis does not directly estimate ω but instead provides an estimate on ω_{bw}; with "bw" for biomarker weighted. Notice, that in a drug target MR context, we denote a biomarker as a metabolite level or other quantitative measure such as blood pressure distal in the causal pathway to the encoded protein.

It is important here, and in general, to note that neither $\omega \neq 0$ nor $\omega_{bw} \neq 0$ provide any evidence on the causal effect of $\boldsymbol{X}$ on $\boldsymbol{D}$ (e.g., θ). To see this, set $\theta = 0$, in which case the results from equations 1 and 2 remain unaffected with $\omega = \phi_P + \mu\theta = \phi_P$. As such, and contrary to the PCSK9 example, we could even consider weighting the genetic effect on an outcome by the genetic effect on a biomarker that does not necessarily reside on the causal pathway but is merely affected by the drug target; see for an empirical example[38]. Finally, we also emphasize that the necessary no-horizontal pleiotropy assumption only pertains to pre-translational effects, such as ϕ_G, and that post-translational effects[58] such as ϕ_P are *part of the drug target effect* and hence do not cause bias. This, of course, will change if our inferential target shifts from P to X, in that case the horizontal pleiotropy assumption requires both ϕ_G and ϕ_P to be zero. Hence drug target MRs, by focussing on exposures more proximal to the effect of genetic variation, are more robust to horizontal pleiotropy than MR studies focussing on the causal effect of more distal biomarkers such as X. This also reflects drug targets often affecting multiple downstream biomarkers and disease, which does not violate the horizontal pleiotropy assumption required in drug target MR, which is exclusively concerned with pre-translational pathways[58].

As we discuss below, the impact of horizontal pleiotropy can be further limited by considering the genomic position of variants associated with the exposure of interest which in the case of a protein as the inferential target resolves to variants within or in the proximity of the encoding gene acting in *cis* vs variants located elsewhere (i.e, *cis* vs *trans* acting).

While $\boldsymbol{X}$ can represent any variable that is affected by a protein drug target, and as such is positioned *post-translationally* with respect to ***P***, we can also think of a variable which is instead positioned *pre-translationally*: mRNA expression of the encoding gene. The right diagram of Figure 2 depicts such a situation where, as an example of a pre-translational variable, mRNA expression is represented as $\boldsymbol{E}$. Furthermore, we decomposed the pre-translational horizontal pleiotropy term $\boldsymbol{\phi_G}$, into ϕ_G and ϕ_E, where the former might occur through LD and the latter representing a direct effect of mRNA expression one the outcome, side stepping its effect on protein translation. If we follow the same derivations as above, it is clear we need to assume that ϕ_G and ϕ_E are both zero. Then we find that weighting a drug target MR by the genetic association with mRNA expression we obtain the effect:

$$\omega_{ew} = \frac{\tilde{\delta}_{GE}\tilde{\delta}_{EP}(\phi_P + \mu\theta)}{\tilde{\delta}_{GE}} = \tilde{\delta}_{EP}(\phi_P + \mu\theta),$$

$$= \tilde{\delta}_{EP} \times \omega.$$

This simply reflects the mRNA effect on the outcome. Critically for this effect to provide robust inference on the protein to outcome effect, we need to assume $\phi_E = 0$. That is, we need to be willing to assume that all of the mRNA effect on the outcome acts through its effect on protein expression. As such, drug target MR using pre-translational weights, where the inferential target remains protein P, need to make stronger assumptions (which may be very reasonable) than if we had simply been interested in the mRNA effect on the outcome itself. This further illustrates that the no-horizontal pleiotropy becomes more stringent the further the *inferential target* is positioned from the genetic variants themselves, and that the inferential target does not necessarily match the genetic information used.

In the preceding section we have generally assumed the inferential drug target was a protein, reflecting the majority case. However, it is more accurate to think of the gene product as the inferential target. That is, any mediator from the gene to the protein. For example, if one wanted to assess the potential effects for an antisense drug, altering mRNA expression, the drug-target MR paradigm could be used to assess the effect of modulating transcript level on the outcome, even if the effect is eventually mediated by the effect of transcript modulation on the encoded protein. This conceptualization utilizes Crick's Central Dogma of a unilateral flow of information from DNA to mRNA to protein.

Interpreting drug target MR effect estimates

As detailed in the previous section, if we are willing to assume complete linearity of the drug target effect(s) on disease, as well as strict homogeneity of drug target effect(s) in all patients, the effect magnitude can sometimes provide actionable insights, allowing drug target MR to inform drug development beyond a statistical test of the null hypothesis.

There are some potential caveats that suggest that drug target MR analysis in general (irrespective of the sourced data), may be more useful as a test of effect direction rather than effect magnitude. This is because drugs that inhibit a target do so usually by modifying its function not its concentration, whereas genetic variants used in MR analysis usually affect protein expression and therefore concentration Given the typically non-linear drug dose-response, the often modest explained variance genetic variants have on the level or function of a protein may misrepresent the potential treatment effect of a drug. MR analyses assess the effect of target modulation in any tissue, whereas certain tissues may be inaccessible to a drug either because of its chemistry or anatomical or physiological barriers. Furthermore, randomized controlled trials (RCTs) are closely monitored, and followed for a fixed period, allowing for exploration of induction-times. MR estimates are considered to reflect a life-long exposure, but in the absence of serial assessment, possible changes across age are difficult to explore, as are disease induction-times. For these reasons we suggest that drug target MR offers a robust indication of effect direction

but may not directly anticipate the effect magnitude of pharmacologically interfering with a protein.

Empirical evidence has shown that effect estimates from drug target MRs often differ from effects from drug compounds affecting the same target[59,60]. While various reasons have been proposed to explain this (e.g., life-time exposure by a genetic variant versus a fraction of this in a drug trial[60]), none have actually considered *if* a drug target MR effect and drug compound effect *should* be equal. In the following we provide straightforward mathematical derivations to show that these effects should in fact be expected to differ.

A drug compound (much like a genetic variant) elicits its effect on an outcome through its effect on a drug target, as such we can replace node $\boldsymbol{G}$ in Figure 2, by $\boldsymbol{C}$ representing the drug compounds. The effect of $\boldsymbol{C}$ on a drug target can be indicated by $\tilde{\delta}_{\boldsymbol{C}}$, and any potential off-target effect on outcome $\boldsymbol{D}$ as $\phi_{\boldsymbol{C}}$ (Figure 3). In this case the effect of a drug compound on the outcome is $\phi_{\boldsymbol{C}} + \tilde{\delta}_{\boldsymbol{C}}(\phi_{\boldsymbol{P}} + \mu\theta) = \phi_{\boldsymbol{C}} + \tilde{\delta}_{\boldsymbol{C}}\omega$; which is distinct from ω, the effect the drug target has on the outcome. It becomes straightforward to see that even when the drug compound affects the outcome exclusively through the drug target (e.g., the compound has no off-target effects $\phi_{\boldsymbol{C}} = 0$) its effect is $\tilde{\delta}_{\boldsymbol{C}}\omega$. Now $\tilde{\delta}_{\boldsymbol{C}}\omega$ will only equal ω in the very specific setting when $\tilde{\delta}_{\boldsymbol{C}} = 1$; that is when the drug compound effect on its drug target is one and the drug compound has no off-target effects. Given that there is little reason for $\tilde{\delta}_{\boldsymbol{C}}$ to equate to one, we can conclude the effects from drug compound and drug target effects are distinct and need not agree. Separating a drug compound effect (which may fully or partially act through a drug target) from the drug target effect itself of course does not invalidate the (causal) relevance of either.

While the above derivations clearly show that the difference between a compound effect and drug target effect (assuming the latter has no off-target effects) is determined by $\tilde{\delta}_{\boldsymbol{c}}$, MR analyses are typically -reweighted by the drug target effect on a biomarker such as LDL-C (θ), or use some kind of version of ω_{bw}, in an attempt to compare trial estimates to MR estimates. We show however, that if it is desirable for the drug target and drug compound effect to equate one another, either the drug

compound effect should be divided by its effect on the drug target $\tilde{\delta}_C$, or the drug target effect should be multiplied by the same constant. The former scaling of course is similar to instrumental variable analyses adjusting compound estimates to account for any non-adherence[61]. To reiterate however, such scaling (either by a genetic biomarker effect or by $\tilde{\delta}_C$) is only feasibly when the drug compound has no off-target effect(s); which is unlikely to generally hold.

Using proteomics data as an exposure in drug target MR

Given that the majority of current drug targets are proteins, it seems reasonable to assume have that loci identified from genetic associations with protein concentration (protein quantitative trait loci; pQTLs) will offer an important category of exposures for drug target MR[38].

Current examples of MRs using protein exposures assess the effect of a change in protein level against a disease outcome. However, since most drugs act by altering the function of a protein, using MR of protein level changes implicitly assumes that an effect of variation in protein concentration is directly proportional to variation in protein function. Limited examples of available genetic associations with concentration and activity of the same protein suggest a high correlation between the two, although evidence is mainly for enzymes and the situation could, in theory, be different for other types of protein. The promise of ProTac therapeutics also indirectly supports protein level variation being able to model disease[33].

The main downsides of currently available pQTL data are that these are largely focused on circulating proteins. Among these, secreted proteins that have their action in the circulation could be an important category of drug targets and many are already the targets of monoclonal antibody therapeutics[62]. However, other proteins are present in the circulation due to cell damage or turnover. Although this is not their physiological site of action, it might be assumed that concentration in the blood reflects tissue-specific expression. If so, the utility of blood pQTLs will depend on relative contributions from the disease relevant vs disease irrelevant tissues to the blood pool. Consequently, the relevance of blood pQTLs will vary considerably

between disease areas. In future, comparison of tissue-specific eQTL and blood pQTL data at scale may provide insights into the tissue of origin of proteins detectable in the blood. Additionally, as technologies improve, more tissue-specific pQTL data may become available and be incorporated in MR analyses. Assay heterogeneity is also a potential issue in the use of pQTL data from MR analysis. For example, more research is needed on the agreement between different assays of the same protein based on new proteomics technologies that measure many thousands of proteins in the circulation in a single sample e.g. the Somalogic aptamer based proteomics platform and the O-link antibody based proximity extension assays, as well as with mass-spectrometry and ELISA -based techniques[63].

cis and *trans* instruments in drug target MR

MR analysis of risk factor or biomarker exposures typically incorporate multiple variants selected from throughout the genome and can provide valuable insight on prioritising generic therapeutic strategies e.g., lowering LDL-cholesterol to prevent CHD. Traditionally, drug target MR has preferentially selected instruments from within a small *cis* region known to encode the (protein) drug target, in an attempt to guard against the influence of pre-translational horizontal pleiotropy. One of the advantages of using *cis*-acting variants in a drug target MR, is that there is a more robust hypothesis that these variants act through the target of interest, which can guard against the influence of pre-translational horizontal pleiotropy, although an exception arises if there is LD with a flanking gene that is the true causal gene. With the increase in the more highly powered GWAS of circulating proteins it has become apparent that the circulating concentration of many proteins is influenced by variants outside of the encoding gene region i.e., acting in *trans*. If one assumes that the majority of *trans* associations reflect real biology (not assay artefacts), then the natural question is: could variants acting in *trans* be used as a source of genetic instruments for MR? Whilst a seemingly attractive proposition, we sound some notes of caution. Variants associated in *trans* with the gene encoding the protein of interest, are likely acting in *cis*- for a second gene in their immediate vicinity and through the gene product encoded by that gene. If the gene product from the second gene has a pathway to disease independent of the protein of interest, horizontal pleiotropy occurs and a critical MR assumption is violated[58]. Only in the absence of

such horizontal pleiotropy would the *trans* variants be valid instruments for the protein of interest. However, with *trans* variants, there is no guarantee that the target of interest is on the causal pathway between the *trans*- acting instrument and the disease (Figure 4).

Interpreting the utility of biomarker-weighted drug target MR analyses

Whilst transcript and protein levels can themselves be regarded as biomarkers, in a drug target MR context, we denote a biomarker as a metabolite level or other quantitative measure such as blood pressure distal in the causal pathway to the encoded protein. It is worth contrasting a drug target MR analyses using biomarker exposures, with those using pQTLs and eQTLs. Whereas for eQTLs and pQTLs there is a natural dichotomy into genetic variants acting in *cis*-(in the vicinity of the encoding gene and the gene product) and those acting in trans (distant from the encoding gene), no such natural dichotomy exists for variants influencing downstream traits such as circulating metabolites. However, the lack of a defined *cis*- does not preclude the study of *cis*-variants and cis-MR analysis at a metabolite associated gene with an outcome, where the inference is on the effect of perturbing a protein that has an effect on a downstream metabolite[58,60]. Within the bounds of the MR assumptions, the inference that the gene product is causally associated with the outcome is valid, however, any inference that it is mediated by the biomarker exposure is not (see above). With a biomarker MR, the biomarker acts as a mechanism to indirectly measure the gene product only and associations between the biomarker level and the outcome status can, and probably are in many cases, bystander effects. As discussed above (Figure 2), a biomarker weighted drug target MR does not estimate ω, the effect of the drug target on disease, and instead estimates $\omega_{bw} = \frac{\omega}{\mu}$, where μ represent the drug target effect on the biomarker. While this leads to a valid null-hypothesis test, it is clear that the sign of ω_{bw} may differ from that of ω, and to appropriately interpret the effect direction one needs robust information on the effect direction of μ.

Despite the more challenging inference, there are some good reasons to conduct a biomarker weighted drug target MR, even in the presence of available pQTL data. Chiefly, many non-protein biomarker GWAS are larger than most GWAS of

proteomics to date, and thus are very highly powered with hundreds or thousands of participants using established assays. Additionally, whilst there are no studies directly assessing this, it is likely that biomarker MR estimates will indirectly incorporate the effect of coding sequence variation in a way that eQTLs and pQTLs may not. The reason is that both e/pQTLs influence transcript/protein level and not protein activity, whereas coding sequence variation is more likely to have an impact on protein activity (and not level), an effect which is then captured in the effect on the downstream biomarker (e.g., a metabolite level). That said, coding sequences may, in some cases, also contribute to protein concentration, e.g., where they lead to a large impact on structure that leads to nonsense mediated decay.

Biological relevance of pre-translational pleiotropy for drug target validation

As we addressed before, horizontal pleiotropy in the context of drug target validation, involves an assumption on the absence of pre-translational horizontal pleiotropy; i.e, $\phi_G = 0$, in Figure 2. Since the seminal contribution by Bowden *et al* introduced the MR-Egger method[64] which is appropriate when all instruments are affected by horizontal pleiotropy, there has been a growing body of methods that, under various assumptions, can provide valid MR estimates in the presence of horizontal pleiotropy[49]. Typically, these methods have not considered drug target MRs specifically, *cis* settings, nor pre-translational pleiotropy. For example, a biomarker weighted drug target MR may be highly heterogenous (e.g., a large Q-statistic). Which could either signal the presence of horizontal pleiotropy[65], or may simply be caused by the drug target effecting disease through multiple pathways (i.e., post-translation pleiotropy).. We showed before that such post-translational pleiotropy is in fact part of the drug target effect and therefore does not invalidate drug target MR estimates[58].

Previously, we discussed the concept of horizontal pleiotropy in the context of *trans*-pQTL associations in a drug-target MR. It is worth also considering the mechanisms and implications of pre-translational pleiotropy in the *cis*-setting. This is particularly relevant when performing drug target scans where there may not be a specific understanding of the genomic locus, or a prior hypothesis for the likely effect. LD between *cis*-variants and variants within other genes surrounding the target locus,

provide an obvious source of pre-translation pleiotropy. However, the presence of such LD, while complicating attributing any disease-causing effect to the selected the *cis*-protein under consideration, does provide potentially valuable information for drug development. Further exploration of the LD region might identify the appropriate gene – protein pair, which if druggable, could lead to further target leads.

Pleiotropy is usually inferred by heterogeneity of the MR estimate, but this needs to be considered in the biological context. (Figure 5). For example, if a coding sequence variant is used as an instrument, that influences protein function but not level, and an effect on a disease outcome might be observed when using a downstream biomarker as the exposure but not when using protein or transcript expression as the exposure. Therefore, coding variation that effects the outcome but not the transcript or protein expression will introduce heterogeneity in the analysis. Protein or transcript isoform-level variation is another theoretical mechanism through which heterogeneity can manifest in an MR estimate. For example, genes that have multiple transcripts or multiple protein isoforms, of which only a subset impacts the outcome. Many eQTLs represent associations with an entire transcript pool of the gene and the isoform binding characteristics of protein assays are usually unknown. Given that variants used as instruments might be drawn from the entire genic region, there remains a possibility that a subset of these variants might operate on the outcome, but their impact might not be completely captured via the exposure assay, manifesting in increased heterogeneity of the MR estimate which in this case could be inappropriately interpreted as horizontal pleiotropy (Figure 5).

Scaling drug target validation approaches

Whilst a hypothesis-driven approach investigating a small subset of targets is relatively easy to perform, it is often desirable to investigate the broad landscape of targets against a disease. Scaling up drug target MR involves computational, statistical and methodological coordination. The following sections explore some of the difficulties involved in performing large-scale analysis, with a particular focus on drug target MR.

Publicly available GWAS data

The vast majority of MR studies employ a two-sample MR approach, irrespective of the precise analytical method. The two-sample paradigm uses exposure and outcome datasets derived in different samples and can operate on summary level genetic associations, and ensures any weak instrument bias acts towards a conservative, neutral effect estimate[66]. By accessing data from two (or more) independent sources, many of the obstacles encountered with sharing individual level data are avoided, and will often allow for large, scaled analyses. Researchers, often encouraged by journals, are now frequently sharing summary-level data upon publication, further increasing the potential for two-sample MR. This greater availability of aggregated data, however, also introduces the problem of data handling, where file format and information detail often differ between publications and research laboratories.

In order to efficiently and reliably conduct large-scale scans across multiple targets, GWAS summary datasets require a homogenous structure and a normalisation of genome assembly, effect alleles and genomic coordinate representation (e.g., the difference between Ensembl and VCF representation INDELs) as well as normalisation of variant identifiers. Several projects have attempted to do this[67,68], with some making normalised datasets available[68]. However, the lack of a common data sharing standard for GWAS data can only be viewed as a missed opportunity that has slowed down the pace of large-scale research. Worse, in some cases results are available but they appear to be deliberately obfuscated to limit their use. For example, the AMD consortium[69] has released summary information where the effect size has been dichotomised to ±1, which prohibits MR analyses.

Absence of genetic variation

As discussed, *cis*-MR analyses may be preferred due to the natural robustness against some sources of horizontal pleiotropy, clearly these MR analyses can only be applied in settings when there is variation in the drug target encoding gene. It is theoretically possible to use variation at other loci that also impacts the protein level (*trans*-associations). However, the use of these variants in MR analyses increases the risk of horizontal pleiotropy. Fortunately, GWAS of mRNA and protein expression provide hundreds of empirical examples of genetic variants influencing the expression of nearby genes (acting in *cis*). Recently, the GTEX project and large

pQTL analysis[70] have catalogued *cis*-variants with functional effects for many thousands of genes suggesting that the absence of cis-acting regulatory variants should not generally be a limiting factor for conducting a drug-target MR[70–72]. However, the presence or absence of coding genetic variation in known drug target encoding loci has rarely been systematically explored[47]. Possibly the absence, or presence, of coding genetic variation itself may provide valuable insight on the viability of a drug target for downstream development [47]. If a potential target cannot tolerate natural variation, does that make it a better drug target? i.e., will it elicit a greater effect for a smaller concentration of drug? or will the effect of targeting it be detrimental to the patient? Clearly in absence of genetic variation, drug target MR will simply overlook these targets, and this further illustrates why MR can only be one (likely important) source of evidence in preclinical drug development.

Evidence prioritization

Multiple testing is a particular concern when considering scans of multiple drug target/disease combinations. As described above GWAS are designed to minimize the false positive rate (type 1 errors) at the cost of an increased false negative rate (type 2 errors). As such, GWAS can robustly show that a genetic association is present but are inadequate to rule out the presence of an association.

One could consider a false positive minimization approach for drug target validation. This clearly makes sense if one wants to focus on efficacy, ensuring target perturbation affects the intended trait(s). Especially in pre-clinical settings, however, drug target analyses using human genetics offer the chance to consider safety as well. Unlike validating intended effect, where the aim is to minimize false claims of efficacy, safety is more concerned with not-overlooking potential signals, which requires an optimization of power (and minimization of false negatives). Given that sample size in most drug target MR analyses is fixed (unless, in the rare case where a *de novo* study is designed to genetically validate a drug target), stringent control of false positives will decrease power and often greatly limit the potential to detect safety signals. Hence, depending on the aim of drug target validation, researchers may wish to find a balance between stringent multiple testing control and sufficient power to offer an appropriate level of sensitively to detect important (safety) signals;

with similar considerations for the identification of repurposing opportunities, for which there might be many.

It may also be beneficial to position genetic drug target validation within a programme of pre-existing preclinical experiments on cell, tissue and animal models. Alignment of human genomic and standard preclinical evidence can be used to attempt to replicate or falsify findings and offer an efficient solution to building confidence in a target and disease. When scanning multiple drug targets, further gains might be made by looking for internal consistency and considering that proteins may be grouped by shared pathways, which might be anticipated to result in consistent MR estimates on a disease.

Due to the growing amount of available GWAS data, we are now able to independently replicate MR findings. For a type 1 error rate α, and m independent replications the false positive rate becomes α^m; for example, for $\alpha = 0.05$, and $m = 3$, type 1 error rate is $0.05^3 = 0.000125$. Of course, alternatively, one may decide to forgo replication and optimize power instead by meta-analysing independent GWAS. Due to the often, cumulative nature of GWAS, where newer publications typically meta-analyse previous GWAS findings, completely independent data is rarely available. An ideal scenario would be a study repository that would enable users to de-convolute the cohorts used in a study and identify truly independent cohorts. Building such a resource would be time consuming unless studies were required to register their cohort specific data prior to publication.

Finally, a shift in inferential perspective maybe required when considering results from drug target scans. As suggested before, a drug target scan should be viewed as one component in a body of evidence. With this in mind, under the null distribution, the distribution of p-values is expected to follow a uniform distribution[73], rather than attempting to differentiate true and false positive association based on a p-value cut off (.e.g., 0.05), testing the whole set of p-values for a significant deviation from the uniform distribution gives an indication of how different the results in the set are from the null distribution (Figure 6). For example, in a pre-clinical setting one may wish to identify targets with a strong cardiometabolic fingerprint,

where 30+ traits (including lipid and glucose measurements) might be relevant. In such a setting the above detailed procedure may be employed to prioritize targets for general cardiometabolic enrichment, before considering individual disease associations.

Examples of how drug target MR can increase drug development yield

In the previous sections we have discussed the rationale, and the biological and methodological underpinnings of utilizing human genetics for drug target identification and validation, with specific focus on drug target MR. Next, we present specific cases where human genetics has supported clinical trial findings.

Pre-clinical drug target prioritization on anticipated in-human effects

For those first-in-class drug molecules in early clinical phase, MR studies could inform “stop or go” decisions (e.g., whether to proceed to human phase I-III clinical trial evaluations).

Two prior examples illustrate the concept. In the first study, Holmes and colleagues evaluated whether secretory phospholipase A2 (PLA2) is a valid therapeutic target for CVD management[74]. This MR study was conducted because a first in-class sPLA2 inhibitor (varespladib) was already in advanced clinical development based on observational association between sPLA2-IIA mass and/or activity and incident CVD events in observational studies. For the MR study, Holmes and colleagues used variants in the gene (*PLA2G2A*) that encodes secretory sPLA2-IIA and showed that these variants did not meaningfully affect CVD, despite a large effect on sPLA2-IIA levels. Consistent with this, the VISTA-16 randomised trial evaluating the effect of a sPLA2 inhibitor (varespladib) on CVD outcomes was stopped prematurely for lack of efficacy. In a second example of an MR study in this category, Casas and colleagues showed that variants in the *PLA2G7* gene which encodes the distinct drug target lipoprotein-associated PLA2 (Lp-PLA2), that were associated with differences in the circulating concentration of this marker, were not associated with altered risk of CVD[75]. In agreement with these findings, a subsequent study that used variants in the *PLA2G7* gene that associated with Lp-PLA2 levels, which again did not show a convincing CHD effect[76]. These MR analyses implied that reverse causation, confounding, or both affect the non-randomized (observational) studies

that before reported a risk increasing association of Lp-PLA2 mass or activity with CVD. Consistent with the MR analysis, an Lp-PLA2 inhibitor, darapladib, also failed to demonstrate efficacy in two phase III randomised controlled trials, one in patients with stable coronary artery disease, the other in patients who had recently suffered an acute coronary syndrome[77,78].

Identification of safety and efficacy phenotypes for evaluation clinical trials

The ability to undertake drug target MR analysis for many disease outcomes makes it feasible to anticipate the effect of perturbing a drug target on a wide range of traits. Theoretical arguments detailing the number of diseases likely to be influenced by any gene or protein, suggest that perturbation of any given drug target is likely to influence the risk of several diseases, and that the profile of effects is target specific. This is backed up by empirical observations of genetic pleiotropy (variants in the same gene being associated with several diseases[79,80]) and the parallel observation that same drug class can be effective in different diseases (therapeutic pleiotropy[80 81]).

Using genomics to pre-specify the outcomes in clinical trials (see ref[82]) that are anticipated to be affected by pharmacological action on a particular target (*target-specific outcomes* of both efficacy) would represent a departure from the current norm where end points in a particular therapeutic area tend to be uniform regardless of the target being evaluated[83]. For example, genetic information led to the discovery of atrial fibrillation as a mechanism based adverse effect of ivabradine, licensed for angina and heart failure[83]. This information could also help reduce the risk of an effective drug and target failing to demonstrate efficacy in a clinical trial because a clinical end-point unaffected by target perturbation has been included in a composite outcome, thereby diluting the observed treatment effect. Furthermore, early information on a target's potential adverse effects, can ensure clinical trials are specifically designed to rule out such an adverse effect (e.g., through non-inferiority designs[84]).

Indication expansion and repurposing opportunities

Large-scale use of a new drug post-licensing sometimes provokes debate on possible unexpected benefits in different disease areas or unexpected harms.

However, this evidence is often insufficient to draw firm conclusions. This is either because it comes from non-randomized (phase IV) pharmacoepidemiological studies (that are prone to biases by for example immortal time bias[85], and severe confounding by indication[51]) or from trials where the outcome of concern was not a primary end-point, which increases false positive (e.g., for efficacy) and negative (e.g., for safety) rates. An example of how drug target MR can provide additional evidence comes from evaluations of whether interleukin-6 receptor (IL-6R) is a valid drug target for prevention of coronary events[86,87]. IL-6R is currently the target for the therapeutic monoclonal antibody tocilizumab, licensed for rheumatoid arthritis. Demonstration of a role for this target in coronary disease would provide an opportunity to expand the indications for this already licensed medication. 'Swerdlow and colleagues showed that variants in the *IL6R* gene exhibited effects on a wide range of inflammation and other markers that were consistent in profile with that seen during tocilizumab treatment of patients with rheumatoid arthritis. The same genetic variants were also associated with a reduced risk of CHD, a finding that has received independent corroboration[86]. These findings implicate IL-6R as a valid target for coronary disease prevention and suggest that tocilizumab might be repositioned perhaps initially as an adjunctive infusional therapy in acute coronary syndrome. This question is currently being addressed in ongoing clinical trials[88,89]. More recent genetic studies have also implicated IL6R in abdominal aortic aneurysm[90], atrial fibrillation and inflammatory bowel diseases[91], flagging additional indication expansion opportunities. The same principle could be applied to develop genetically supported repurposing hypotheses for many first-in-class compounds and targets that proved safe in humans but which failed in their originally intended indication. A series of compounds and related targets have been the subject of asset sharing initiatives developed by the US National Institutes of Health and UK Medical Research Council[92].

Delineating compound specificity

Proving that a drug engages its therapeutic target in humans and with no off-target actions has been a major challenge in clinical phase drug development. When a first-in-class molecule causes an adverse effect in clinical phase trials, or post-marketing, it can be difficult to decide if this is mechanism-based or "off-target" (specific to the drug molecule and unlikely to be shared by other class members). In the past,

clarification has been provided by undertaking a randomized trial of other (chemically dissimilar) agents from the same class, but this risks further harm if the adverse effect is mechanism-based rather than agent specific. Variants in a gene that encode a drug target and that affect its expression or activity should lead to metabolic and physiological changes that profile the effects of a clean drug with no off-target actions[93]. Hence comparison of the effects of variants in a gene encoding a drug target in population studies, and a drug targeting the same protein in a trial could help distinguish on- from off-target effects. For example when torcetrapib, a first-in-class cholesteryl-ester transfer protein (CETP) inhibitor, was developed to raise HDL-cholesterol with the aim of preventing cardiovascular events, an unexpected blood-pressure-elevating effect was detected during a large phase-III randomised trial. To disentangle whether this effect was an on or off-target effect, variants in the CETP gene were used to reproduce the effect of CETP inhibition on the major blood lipid fractions (on-target actions) as well as reduce the risk of CHD, which showed a blood pressure *decreasing* effect[38]. These results argued that the blood pressure elevating effect of torcetrapib, and subsequent compounds evacetrapib and anacetrapib, are off-target and should not be shared by CETP inhibitors that are sufficiently different[38]. Indeed the chemically distinct CETP inhibitor Dalcetrapib showed the same blood pressure decreasing effect as reported by the CETP MR[38].

The same principle was used to demonstrate that variants in the *HMGCR* gene that encode HMG-coA reductase, the target for statins, reproduce the effects of statin treatment on multiple metabolites and lipoprotein lipid subclasses measured by NMR spectroscopy[19]. The same genetic variants were used to show that the modestly increased risk of type 2 diabetes among statin users in clinical trials (which does not offset their clinical benefit in reducing coronary heart disease events, even among patients with diabetes) is an on-target action, potentially mediated in part by increases in body weight and waist circumference [94].

With the ability to undertake GWAS of numerous proteomics and metabolomics blood biomarkers, as well as physiological and imaging measures in large cohort studies, comes the ability to infer the effects of drug target perturbation not only on disease end-points but also a vast range of variables that could serve as readouts for adequate and specific target engagement in early phase I-II clinical trials.

Concluding remarks

The approaches we have described necessitate linking genotype to phenotype at scale. This is exposing the need for greater partnership between academia, healthcare systems, technology developers and providers, as well as the pharmaceutical industry. This is because large population and patient cohorts (and associated phenotypic and disease outcome data) reside mainly in the public sector, while new technologies, compound development expertise and financial resources reside mainly in the commercial sector. Developing governance frameworks that allow equitable partnership among these players, and which recognize the contribution made by different stakeholders to value creation, is challenging.

Different models are emerging. Primarily publicly funded, academia-led initiatives include UK Biobank, the EMERGE Consortium, All Of Us and the Million Veteran Program, as well as the recently completed UK 100,000 genomes project. Some studies (e.g. UK Biobank) have subsequently secured additional industry investment for sequencing or proteomics analysis. In other cases, a commercial vehicle has emerged from an academic one (e.g. Nashville Biosciences from Vanderbilt University BioVU). Initiatives with a major pharmaceutical industry component include Amgen purchase of Decode Genetics, the Regeneron partnership with Geisinger Health, and the recent GSK investment in the consumer genetic testing company with a huge client database, 23andMe. The origin, funding, generation, and control of, and access to, the linked genotype and phenotype data differ among these initiatives. Should such resources result in new drugs, it remains unclear if, and how, the contribution of the public sector or private citizens will be recognized. For instance, in the direct-to-consumer genomics arena, an opaque relationship exists between citizens (who pay consumer genomics providers for personal genome analysis for ancestry and disease risk information), and the pharmaceutical industry (who pay the same providers for access to aggregated personal genetic data linked to self-reported health outcomes). In this scenario, citizens risk paying twice: once for a genetic profile provided by the personal genomics company and then again later for any drug developed by its Pharma partners through insights generated from the aggregated genetic and health data to which citizens have already contributed at their own cost. In addition, in some healthcare settings, many

of the citizens that contributed their data to the development might not be able to benefit from any developed therapeutics due to their cost. These emerging requirements for genomic and healthcare data for new drug development are likely to force a rethink of the social contract and economic model for drug development.

Finally, throughout this contribution we have discussed how human genetics and drug target Mendelian randomization can complement existing sources of (pre-) clinical evidence on anticipating the likely effect of drug target perturbation in humans to increase the yield of drug development. We see this as an adjunct in the drug development process but not as a substitute for randomised trials. These will continue to be required because (1) as we explained here drug compound and drug target are distinct exposures, (2) any new compound could have off-target actions that cannot be modelled genetically; (3) drug target effects modelled through MR, even tissue specific analyses, reflect biological consequences of target perturbation, which does not guarantee (irrespective of absence or presence of off-target effects) that a drug compounds will affect the target in the same manner in the same tissues (some of which are notorious difficult to access using a drug); (4) by starting follow-up immediately after randomization drug trials are protected against many of the biases that may still affect genetic studies[49] which often only enroll subjects decades after gamete forming. Nevertheless, by providing *early, in-human,* evidence on the anticipated on-target effects of drug target perturbation against an array of *clinically* relevant traits, drug target MR is likely to provide additional *accurate* information on which target to pursue for clinical phase development. As such integration of human genetics in (preclinical) drug development is expected to decrease cost and increase yield.

Acknowledgments

AFS is supported by BHF grant PG/18/5033837 and the UCL BHF Research Accelerator AA/18/6/34223. CF and AFS received additional support from the National Institute for Health Research University College London Hospitals Biomedical Research Centre. ADH is an NIHR Senior Investigator, and additionally by a UKRI-NIHR grant MR/V033867/1 for the Multimorbidity Mechanism and Therapeutics Research Collaborative.

References

1. Pina, A. S., Hussain, A. & Roque, A. C. A. An historical overview of drug discovery. *Methods Mol Biol* **572**, 3–12 (2009).
2. Drews, J. Drug discovery: a historical perspective. *Science* **287**, 1960–1964 (2000).
3. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat Rev Drug Discov* **16**, 531–543 (2017).
4. Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br J Pharmacol* **162**, 1239–1249 (2011).
5. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nat Rev Drug Discov* **10**, 507–519 (2011).
6. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Science Translational Medicine* **9**, (2017).
7. Lindsay, M. A. Target discovery. *Nat Rev Drug Discov* **2**, 831–838 (2003).
8. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. *Clinical development success rates for investigational drugs*. vol. 32 (2014).
9. Munos, B. *Lessons from 60 years of pharmaceutical innovation*. (2009). doi:10.1038/nrd2961.
10. Pammolli, F., Magazzini, L. & Riccaboni, M. The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery* **10**, 428–438 (2011).
11. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* **9**, 203–214 (2010).
12. Harrison, R. K. Phase II and phase III failures: 2013–2015. *Nature Reviews Drug Discovery* **15**, 817–818 (2016).
13. Hingorani, A. D. *et al.* Improving the odds of drug development success through human genomics: modelling study. *Scientific Reports* **9**, 18911 (2019).
14. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

15. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356–369 (2008).
16. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5–22 (2017).
17. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
18. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet* **17**, 353–373 (2016).
19. Würtz, P. *et al.* Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *J Am Coll Cardiol* **67**, 1200–1210 (2016).
20. Wheeler, J., McHale, M., Jackson, V. & Penny, M. Assessing theoretical risk and benefit suggested by genetic association studies of CCR5: experience in a drug development programme for maraviroc. *Antiviral Therapy* 14 (2007).
21. GSK and 23andMe sign agreement to leverage genetic insights for the development of novel medicines | GSK. https://www.gsk.com/en-gb/media/press-releases/gsk-and-23andme-sign-agreement-to-leverage-genetic-insights-for-the-development-of-novel-medicines/.
22. Richardson, T. G., Zheng, J. & Gaunt, T. R. Computational Tools for Causal Inference in Genetics. *Cold Spring Harb Perspect Med* **11**, a039248 (2021).
23. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *The American Journal of Human Genetics* **97**, 139–152 (2015).
24. Taşan, M. *et al.* Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature Methods* **12**, 154–159 (2015).
25. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications* **8**, 1826 (2017).
26. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics* **16**, e1008720 (2020).

27. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genetics* **13**, e1006646 (2017).
28. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* **99**, 1245–1260 (2016).
29. Shen, J., Song, K., Slater, A. J., Ferrero, E. & Nelson, M. R. STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. *Bioinformatics* **33**, 2784–2786 (2017).
30. Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature genetics* **34**, 154–156 (2003).
31. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161–165 (2005).
32. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3 (2019).
33. Sakamoto, K. M. *et al.* Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *PNAS* **98**, 8554–8559 (2001).
34. It's all druggable. *Nature Genetics* **49**, 169–169 (2017).
35. Schmidt, A. F. *et al.* PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *The Lancet Diabetes & Endocrinology* **5**, 97–105 (2017).
36. Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *New England Journal of Medicine* **375**, 2144–2153 (2016).
37. Williams, D. M., Finan, C., Schmidt, A. F., Burgess, S. & Hingorani, A. D. Lipid lowering and Alzheimer disease risk: A mendelian randomization study. *Annals of Neurology* **87**, 30–39 (2020).
38. Schmidt, A. F. *et al.* Cholesteryl Ester Transfer Protein as a Drug Target for Cardiovascular Disease. *medRxiv* 27 (2020).

39. Plenge, R. M., Scolnick, E. M. & Altshuler, D. *Validating therapeutic targets through human genetics*. (2013). doi:10.1038/nrd4051.

40. Coignard, J. *et al.* A case-only study to identify genetic modifiers of breast cancer risk for BRCA1/BRCA2 mutation carriers. *Nat Commun* **12**, 1078 (2021).

41. Sipeky, C., Tammela, T. L. J., Auvinen, A. & Schleutker, J. Novel prostate cancer susceptibility gene SP6 predisposes patients to aggressive disease. *Prostate Cancer Prostatic Dis* 1–9 (2021) doi:10.1038/s41391-021-00378-5.

42. Pietzner, M. *et al.* Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat Commun* **11**, 6397 (2020).

43. Newport, M. J. & Finan, C. Genome-wide association studies and susceptibility to infectious diseases. *Brief Funct Genomics* **10**, 98–107 (2011).

44. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res* **45**, D945–D954 (2017).

45. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012 (2019).

46. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* (2015) doi:10.1038/ng.3314.

47. Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. *Nature* **581**, 459–464 (2020).

48. Dudbridge, F. Polygenic Mendelian Randomization. *Cold Spring Harb Perspect Med* **11**, a039586 (2021).

49. Richmond, B. & Davey Smith, G. Mendelian Randomization: Concepts and Scope. *Cold Spring Harb Perspect Med* **11**, a040501.

50. Altman, D. G. & Bland, J. M. Statistics notes: Absence of evidence is not evidence of absence. *British Medical Journal* **311**, 485–485 (1995).

51. Schmidt, A. F., Klungel, O. H. & Groenwold, R. H. H. Adjusting for Confounding in Early Postlaunch Settings: Going beyond Logistic Regression Models. *Epidemiology* **27**, 133–142 (2016).
52. Hernan, M. A., Hernandez-Diaz, S. & Robins, J. M. A structural approach to selection bias. *Epidemiology* **15**, 615–625 (2004).
53. Schmidt, A. F. *et al.* Comparison of variance estimators for meta-analysis of instrumental variable estimates. *International Journal of Epidemiology* (2016) doi:10.1093/ije/dyw123.
54. Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V. & Klungel, O. H. Instrumental variables: application and limitations. *Epidemiology* **17**, 260–267 (2006).
55. Hwang, L.-D., Davies, N. M., Warrington, N. M. & Evans, D. M. Integrating Family-Based and Mendelian Randomization Designs. *Cold Spring Harb Perspect Med* **11**, a039503 (2021).
56. Patel, R. S. *et al.* Subsequent Event Risk in Individuals With Established Coronary Heart Disease: Design and Rationale of the GENIUS-CHD Consortium. *Circ: Genomic and Precision Medicine* **12**, (2019).
57. Cohen, J. C., Boerwinkle, E., Mosley Jr., T. H. & Hobbs, H. H. *Sequence variations in PCSK9, low LDL, and protection against coronary heart disease*. vol. 354 (2006).
58. Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020).
59. Ference, B. A. How to use Mendelian randomization to anticipate the results of randomized trials. *European Heart Journal* ehx462–ehx462 (2017).
60. Holmes, M. V., Richardson, T. G., Ference, B. A., Davies, N. M. & Davey Smith, G. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat Rev Cardiol* (2021) doi:10.1038/s41569-020-00493-1.
61. Schmidt, A. F. & Groenwold, R. H. H. Adjusting for bias in unblinded randomized controlled trials. *Statistical Methods in Medical Research* **0**, (2016).

62. Attwood, M. M., Jonsson, J., Rask-Andersen, M. & Schiöth, H. B. Soluble ligands as drug targets. *Nat Rev Drug Discov* **19**, 695–710 (2020).

63. Pietzner, M. *et al.* Cross-platform proteomics to advance genetic prioritisation strategies. *bioRxiv* 2021.03.18.435919 (2021) doi:10.1101/2021.03.18.435919.

64. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512–525 (2015).

65. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in medicine* **36**, 1783–1802 (2017).

66. Burgess, S. & Thompson, S. G. Avoiding bias from weak instruments in mendelian randomization studies. *International Journal of Epidemiology* **40**, 755–764 (2011).

67. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics* doi:10.1093/bioinformatics/btz469.

68. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* (2018) doi:10.7554/eLife.34408.

69. Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature Genetics* **48**, 134–143 (2016).

70. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* (2018) doi:10.1038/s41586-018-0175-2.

71. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat Commun* **9**, 3268 (2018).

72. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* **2**, 1135–1148 (2020).

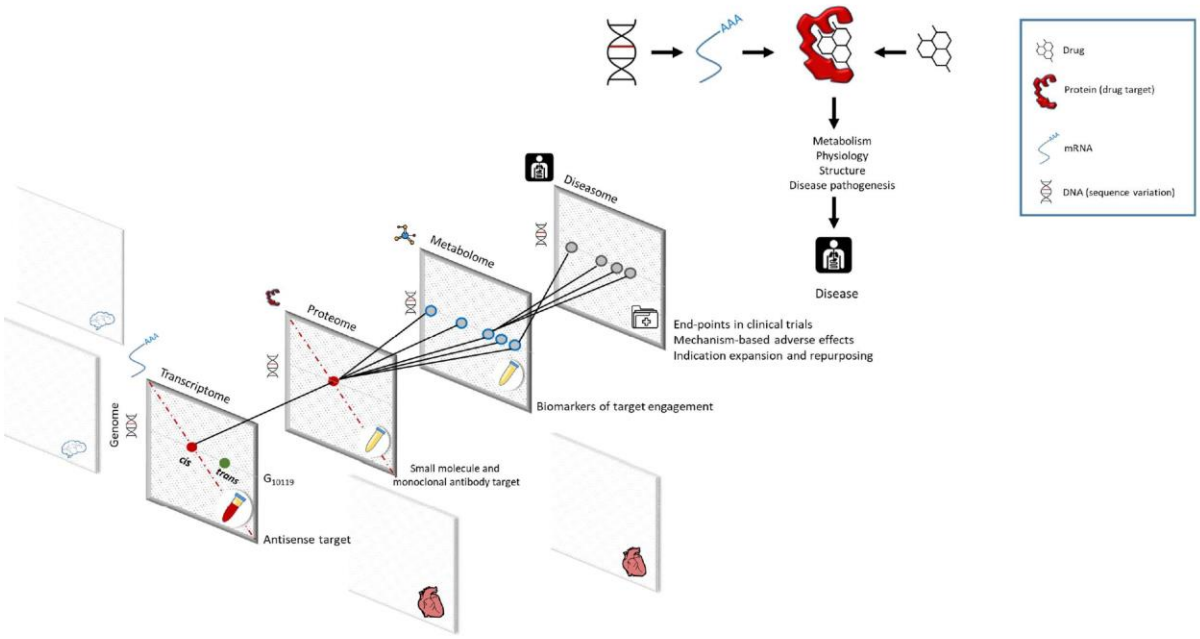
73. Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **64**, 479–498 (2002).

74. Holmes, M. V. *et al.* Secretory phospholipase A2-IIA and cardiovascular disease: A mendelian randomization study. *Journal of the American College of Cardiology* **62**, 1966–1976 (2013).
75. Casas, J. P. *et al.* PLA2G7 Genotype, lipoprotein-associated phospholipase A2 activity, and coronary heart disease risk in 10 494 cases and 15 624 controls of european ancestry. *Circulation* **121**, 2284–2293 (2010).
76. Millwood, I. Y. *et al.* A phenome-wide association study of a lipoprotein-associated phospholipase A2 loss-of-function variant in 90 000 Chinese adults. *International Journal of Epidemiology* dyw087 (2016) doi:10.1093/ije/dyw087.
77. O'Donoghue, M. L. *et al.* Effect of darapladib on major coronary events after an acute coronary syndrome: the SOLID-TIMI 52 randomized clinical trial. *JAMA* **312**, 1006–1015 (2014).
78. STABILITY Investigators *et al.* Darapladib for preventing ischemic events in stable coronary heart disease. *N Engl J Med* **370**, 1702–1711 (2014).
79. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607–618 (2011).
80. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483–495 (2013).
81. Shih, H.-P., Zhang, X. & Aronov, A. M. Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. *Nat Rev Drug Discov* **17**, 19–33 (2018).
82. Ference, B. A., Holmes, M. V. & Smith, G. D. Using Mendelian Randomization to Improve the Design of Randomized Trials. *Cold Spring Harb Perspect Med* **11**, a040980 (2021).
83. Martin, R. I. R. *et al.* Atrial fibrillation associated with ivabradine treatment: meta-analysis of randomised controlled trials. *Heart* **100**, 1506–1510 (2014).
84. Giugliano, R. P. *et al.* Cognitive Function in a Randomized Trial of Evolocumab. *New England Journal of Medicine* **377**, 633–643 (2017).
85. Levesque, L. E., Hanley, J. A., Kezouh, A. & Suissa, S. *Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes*. vol. 340 (2010).

86. Sarwar, N. & Butterworth, A. S. Interleukin-6 receptor pathways in coronary heart disease: A collaborative meta-analysis of 82 studies. *The Lancet* **379**, 1205–1213 (2012).
87. Swerdlow, D. I. *et al.* The interleukin-6 receptor as a target for prevention of coronary heart disease: A mendelian randomisation analysis. *The Lancet* **379**, 1214–1224 (2012).
88. Kleveland, O. *et al.* Effect of a single dose of the interleukin-6 receptor antagonist tocilizumab on inflammation and troponin T release in patients with non-ST-elevation myocardial infarction: a double-blind, randomized, placebo-controlled phase 2 trial. *Eur Heart J* **37**, 2406–2413 (2016).
89. Anstensrud, A. K. *et al.* Rationale for the ASSAIL-MI-trial: a randomised controlled trial designed to assess the effect of tocilizumab on myocardial salvage in patients with acute ST-elevation myocardial infarction (STEMI). *Open Heart* **6**, e001108 (2019).
90. Harrison, S. C. *et al.* Interleukin-6 receptor pathways in abdominal aortic aneurysm. *Eur Heart J* **34**, 3707–3716 (2013).
91. Parisinos, C. A. *et al.* Variation in Interleukin 6 Receptor Gene Associates With Risk of Crohn's Disease and Ulcerative Colitis. *Gastroenterology* **155**, 303-306.e2 (2018).
92. Hayes, A. G. & Nutt, D. J. Compound asset sharing initiatives between pharmaceutical companies, funding bodies, and academia: Learnings and successes. *Pharmacol Res Perspect* **7**, e00510 (2019).
93. Sofat, R. *et al. Separating the mechanism-based and off-target actions of cholesteryl ester transfer protein inhibitors with CETP gene polymorphisms*. vol. 121 (2010).
94. Swerdlow, D. I. *et al. HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: Evidence from genetic analysis and randomised trials*. vol. 385 (2015).

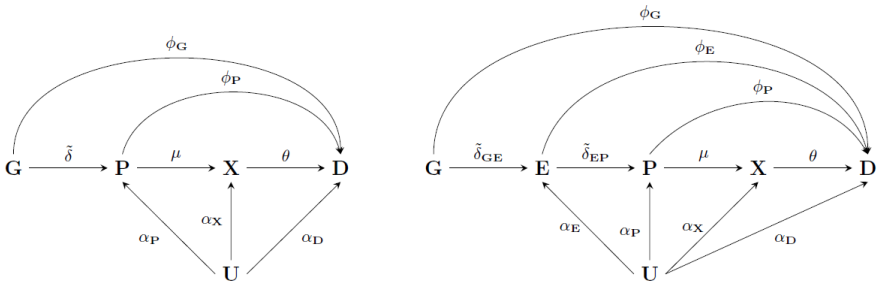
Figure legends

Figure 1 Human genomics and drug development



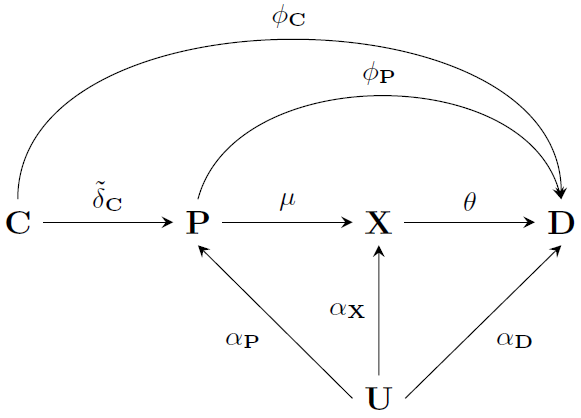
Right hand panel. Relationship between human genomics and drug target identification and validation. Proteins mediate the effect of drugs and natural genetic variation on metabolism, physiology, organs structure and disease pathogenesis. *Left hand panel.* Scalable approaches to interrogate all potential targets and diseases. Mapping the effect of genetic variation (genotype) on gene and protein expression (transcriptome and proteome) in different tissues, on metabolism (metabolome) and disease risk (diseasome) and applying drug target Mendelian randomization can help anticipate the effect of drug action on a target protein. This principle can help support drug target identification, validation, *separation of 'on-'* from 'off-target' effects, end-point selection for clinical trials and indication expansion and repurposing priorities. Information contributing to the different data layers can be summary-level and obtained in independent datasets.

Figure 2 A graph representation of two possible Mendelian randomization scenarios.



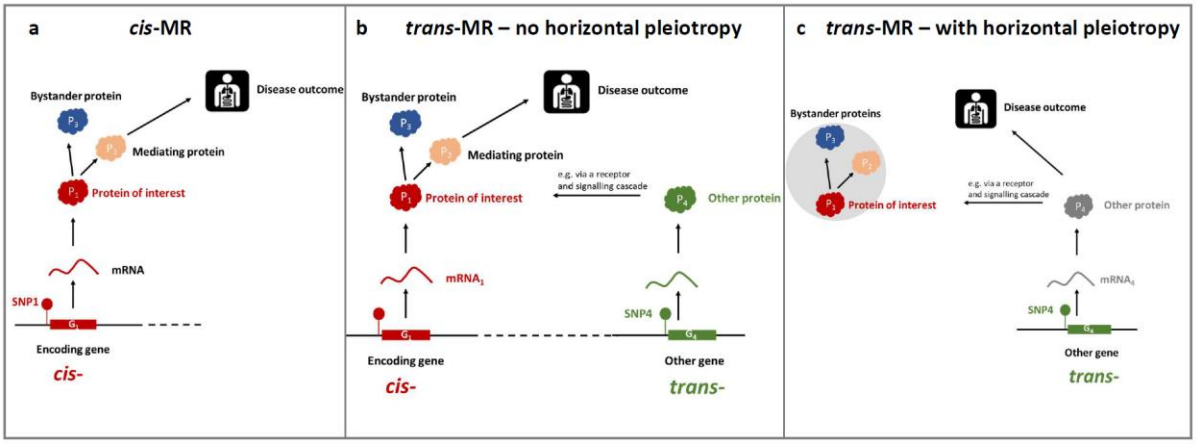
N.b., the *left hand side* of the graph represent a scenario were a single (or multiple) genetic variant is represented by node **G**, which has an effect (indicated by an arrow) on a protein **P**, where the protein affects a downstream biomarker **X**, and all previously defined nodes affect the outcome **D**. Here the effect magnitudes are indicated by arrow labels and may include a null effect (when there is no causal effect between two nodes). The *right hand side* diagram adds a node **E,** between **G** and **P**, reflecting the effect of mRNA expression on **P** and **D**. Finally in both scenarios all nodes, except **G** may be affect by confounding, encoded by common cause **U**; where of course this node most often reflect multiple distinct causes.

Figure 3 A graph representation of a randomized controlled trial of a drug compound.



N.b., Node C represent a drug compound which has an effect (indicated by an arrow) on a protein P, where the protein affects a downstream biomarker X, and all previously defined nodes affect the outcome D. Here the effect magnitudes are indicated by arrow labels and may include a null effect (when there is no causal effect between two nodes). Finally, all nodes, except C (which we assume has been allocated at random) may be affect by confounding, encoded by common cause U; where of course this node most often reflect multiple distinct causes.

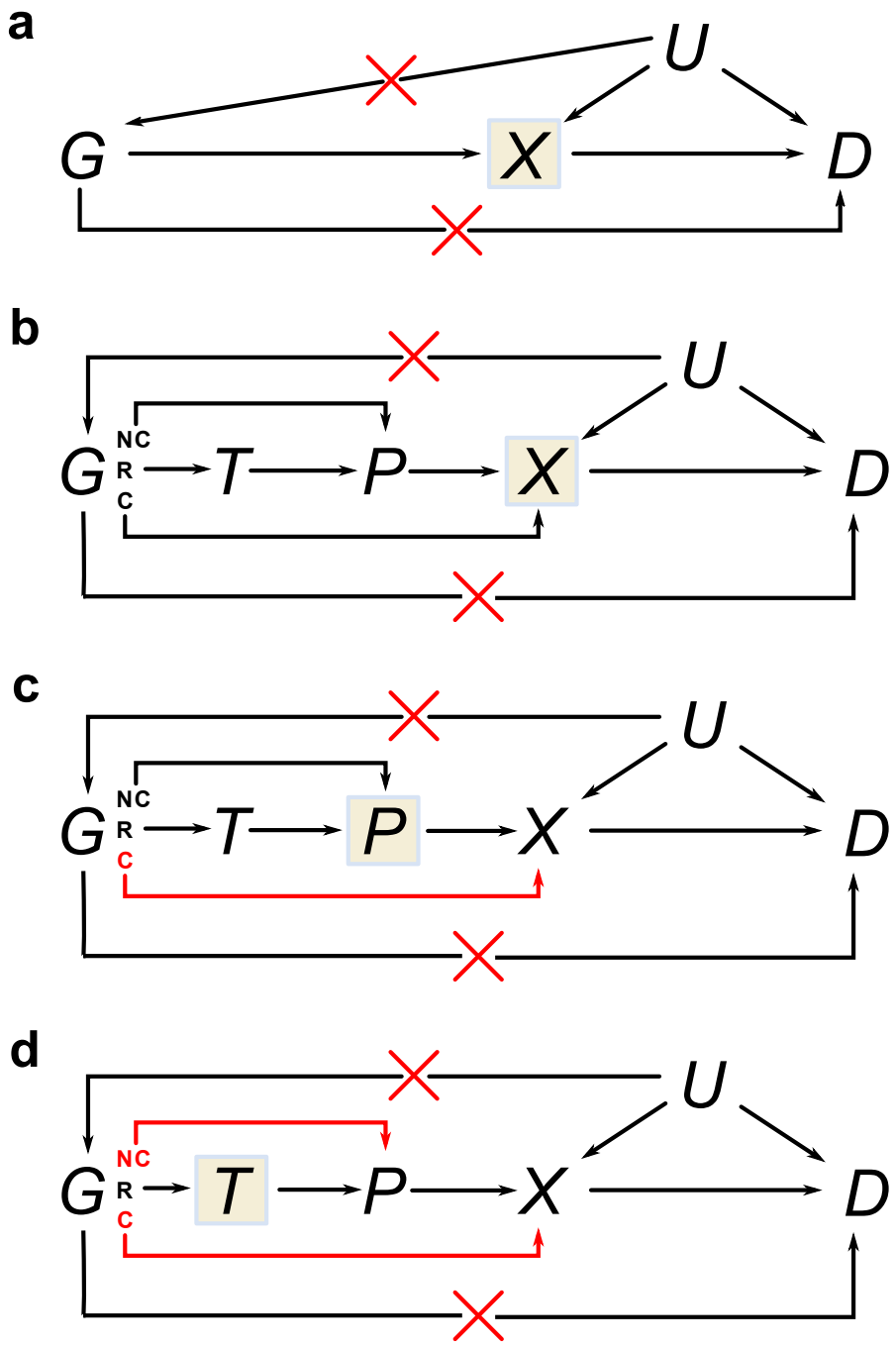
Figure 4 Comparison between cis- vs. trans-Mendelian randomization (MR) for drug target validation.



(a) *cis*-MR A variant in cis- SNP1 in the proximity of gene G1 is used as an instrument for the protein of interest (P1) which is causally linked to a disease. SNP1 is also associated with P2 (a mediator of the effect of P1 on a disease) and with P3 (a bystander protein residing off the causal pathway from P1 to disease. SNP1 is associated with P1, P2, P3 and the disease outcome and is a valid instrument for P1. **(b) *trans*-MR with no horizontal pleiotropy** A variant in another gene (SNP4) which influences expression of P4, upstream in the causal pathway, is used as a trans- instrument for P1. SNP4 associates with P4, P1, P2 and P3 and is a valid instrument for P1 because there is no direct causal pathway from P4 to disease. **(b) *trans*-MR with horizontal pleiotropy** In this scenario, P1 is not causal for disease but is a bystander protein. SNP4 is upstream of P1 and associates with disease

because of a direct pathway through P4 (which may not have been measured). As in scenario (b) SNP 4 associates with P1, P2 and P3 and disease, but it is not a valid instrument for P1 because there is a direct causal pathway from P4 to disease.

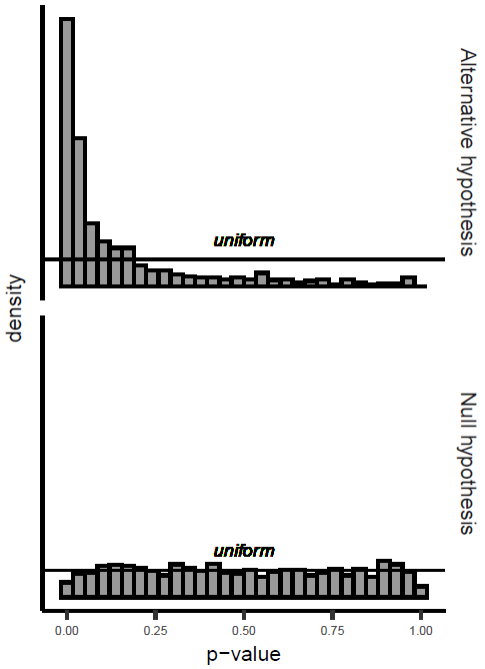
Figure 5 Cis-acting variants with respect to different exposures and perceived heterogeneity



Multiple scenarios where *cis*-acting variants can impact the outcome but not via the exposure. (**a**) A conventional MR graph whereby instrumental variants in gene ***G***are

acting via biomarker exposure ***X*** on disease outcome ***D***. (**b**) Mechanistically the same as part (**a)**but with greater resolution depicting potential (but unused) exposures of transcript level, ***T*** and protein level ***P*** and with the instrumental genetic variants partitioned into non-coding (NC), regulatory (R) and coding (C) variants. Regulatory variants are more likely to impact the transcript level, where as non-coding variants are more likely to impact protein level via translational efficiency and not the transcript level, where as coding variants are more likely to impact on protein function/activity which will alter downstream biomarker level but this effect will not be mediated via transcript or protein level. An exception to these assumptions is nonsense mediated decay where aberrant insertion stop codons will lead to destruction of mRNA. Part (**c**) the exposure is changed from the biomarker to the protein encoded by gene ***G***, and part (**d**) the exposure is changed from the level of protein to transcript level ***T***. In parts (**c**) and (**d**), pathways whereby instrumental genetic variants are impacting the outcome but not via the exposure are highlighted in red.

Figure 6 P-value distributions when the null-hypothesis is false ("alternative hypothesis") and true.



N.b., p-values under the null-hypothesis were generated by sampled from a standard normal distribution, whereas p-values under the alternative distribution were sampled from a normal distribution with mean of 2 and standard deviation of 1.

Table 1 Comparison of the findings from randomised controlled trials and Mendelian randomization trials of the corresponding therapeutic target.

	Orthodox drug development					Mendelian randomization trials (MRT)			
Drug target	**Compound(s) evaluated**	**Developmental stage**	**Therapeutic area**	**Outcomes assessed in preclinical studies or RCTs of selective drug interventions**	**Findings from preclinical studies or RCTs of selective drug interventions**	**Encoding gene**	**Outcomes evaluated in MRTs**	**Findings from MRTs**	**Inferences drawn from comparison of the findings from preclinical studies or RCTs and MRT**
Cholesteryl ester transfer protein[1]	Torcetrapib	Phase III	Cardiovascular disease	Blood lipids (total-, LDL-, and HDL cholesterol, triglycerides); blood pressure; CVD events	HDL-elevation, triglyceride and LDL- reduction. Unintended BP elevation. Unintended increase in CVD events	***CETP***[2]	Blood lipids (total-, LDL-, and HDL cholesterol, triglycerides); blood pressure	Associations with blood lipids consistent with effects in RCTs. No genetic association with BP.	Blood pressure elevating effect of torcetrapib is offtarget
Hydroxy methyl (HMG)-coA reductase[3]	Statins	Phase IV (post-marketing)	Cardiovascular disease	Blood lipid fractions, weight, type 2 diabetes risk	Statin treatment in RCTs linked to increased weight and risk of type 2 diabetes.	***HMGCR***[3]	Blood lipid fractions, anthropometric measures, glucose and insulin, type 2 diabetes risk	*HMGCR* SNPs associated with lower LDL-C, higher weight, fasting glucose and insulin, and type 2 diabetes risk	Increased risk of type 2 diabetes is an unintended on-target effect of statins mediated in part through weight gain
Niemann-Pick C1-like 1 [4]	Ezetimibe	Phase III	Cardiovascular disease	LDL-cholesterol, cardiovascular death, non-fata myocardial infarction, unstable angina requiring hospitalisation and revascularisatio n	Ezetimibe added to statins produces modest additional benefit in cardiovascular outcomes in patients following an acute coronary syndrome	***NPC1L1*** [5]	Plasma lipid levels and risk of coronary heart disease.	Inactivating mutations in NPC1L1 are associated with lower LDL-cholesterol and protection from myocardial infarction risk.	Niemann-Pick C1-like 1 is a validated target for LDL-cholesterol lowering and coronary heart disease prevention.
Proprotein convertase subtilisin/kexin type 9 serine protease [6]	Alirocumab, evolocumab	Phase II	Lipid lowering and cardiovascular disease	LDL-cholesterol	Alirocumab and evolocumab reduce LDL-cholesterol among patients with heterozygous familial or	***PCSK9*** [7]	LDL-cholesterol and rosk of coronary heart disease	Inactivating mutations in PCSK9 associated with reduced LDL-cholesterol and CHD risk	Proprotein convertase subtilisin/kexin type 9 serine protease is a validated target for LDL-cholesterol

					polygenic hypercholesterolaemiaand reduce cardiovascular events in patients with or at high risk of cardiovascular disease				lowering and reduction in cardiovascular risk
Glucagon-like peptide-1 receptor [8]	Liraglutide	Phase III	Diabetes and cardiovascular disease	Death from cardiovascular causes, non-fata myocardial infarction, or non-fata stroke.	Liraglutide reduced risk of death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke among patients with type 2 diabetes mellitus	***GLP1R*** [9]	Body weight, glycaemic traits, lipids, blood pressure, risk of type 2 diabetes and coronary heart disease	A low frequency, coding region missense variant in GLP1R is associated with lower fasting glucose, diabetes risk and risk of coronary heart disease.	GLP1R is a validated target for treatment of diabetes and reducing coronary heart disease risk
Lipoprotein- associated phospholipase A2 (Lp-PLA2) [10,11]	Darapladib	Phase III	Cardiovascular disease	Major cardiovascular events or major coronary events	No reduction in CVD events in patients with stable coronary disease or recent ACS; despite reductions in Lp-PLA2 mass and activity.	***PLA2G7***[12, 13]	Lp-PLA2 concentration, blood lipids, inflammation markers, and CHD events	PLA2G7 variants were not associated with alterations in cardiovascular risk markers or CHD events	Lp-PLA2 is not involved in the development of cardiovascular disease; low priority as therapeutic target for this indication
Interleukin-6 receptor[14]	Tocilizumab	Phase III	Autoimmune disease	Blood lipid fractions and inflammation markers including IL- 6, CRP and fibrinogen	In patients with rheumatoid arthritis, tocilizumab induced alterations in circulating inflammation markers characteristic of IL-6 blockade	***IL6R***[14]	Blood lipid fractions and inflammation markers including iL-6, CRP and fibrinogen. Cardiovascular events including CHD events and abdominal aortic aneurysm	Variants in the *IL6R* gene that recapitulate the biomarker profile of IL6-R blockade were associated with a reduction in CHD events	IL-6 receptor signalling is involved in the development of CHD. The IL-6 receptor blocker tocilizumab could be repurposed for the treatment of CVD
C-reactive protein[15]	No CRP inhibitors yet available for	Preclinical	Cardiovascular disease	Effects of CRP on processes	Observational associations of	***CRP***[16]	Inflammation and coagulation	SNPs in the CRP gene exclusively	CRP is not Causal in CHD

	clinical use.			believed to contribute to atherosclerosis studied *in vitro* or in animals. Associations of CRP with CVD in human observational studies.	CRP with CVD events in humans, but studies prone to confounding. Pro-atherogenic effect of CRP *in vitro* and in animals later proved to be artefactual.		markers, blood lipid fractions, and coronary heart disease events	associated with CRP exhibited no association with CHD. No causal association of CRP with CHD based on instrumental variables analysis.	pathogenesis; priority as a therapeutic target for CHD prevention diminished
Secretory phospholipase A2 (sPLA2)[17]	Varespladib	Phase III	Cardiovascular disease	sPLA2 concentration, blood lipids, inflammation markers, and CVD events	No beneficial effect of varespladib on CVD events in patients with recent acute coronary syndrome (ACS), despite a drug-induced reduction in sPLA2 concentration and activity	***PLA2G2A***[18]	sPLA2 mass and activity and major vascular events (MVE) in general populations and patients with ACS	SNPs in the PLA2G2A gene were associated with substantial alterations in sPLA2 mass and activity but not with MVE	sPLA2 is not involved in the development of cardiovascular disease; dismissed as a therapeutic target in CVD
Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4 [19]	Ivabradine	Phase IV (post-marketing)	Cardiovascular disease	Risk of atrial fibrillation	Developed for angina and heart failure, post-hoc meta-analysis of RCTs (motivated by genetic findings [14, 15], indicated ivabridine treatment is associated with a higher risk of atrial fibrillation.	***HCN4*** [20,21]	Atrial fibrillation (genome wide association analysis)	Variants in the gene *HCN4* encoding the target of ivabridine associate with a higher risk of atrial fibrillation.	Atrial fibrillation is a mechanism-based adverse effect of ivabridine treatment.
TNF receptor 1 and TNF [22 23]	Monoclonal antibodies against tumour necrosis factor-alpha (TNF)	Phase II I and Phase IV	Neurological disease	Multiple sclerosis exacerbations	Multiple sclerosis exacerbations.	***TNFRSF1A*** [24]	Multiple sclerosis	A variant in the TNFRSF1A that encodes the TNF receptor 1 gene indices expression of a	Exacerbation of MS induced by anti-TNF monoclonal antibodies is mechanism

								soluble form of TNFR1 that blocks the effect of TNF, and associates with a higher risk of MS. The mechanism mimics that of monoclonal antibodies against TNF.	based.

Table references:

1. Sofat R, Hingorani AD, Smeeth L, Humphries SE, Talmud PJ, Cooper J, 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. doi:10.1161/CIRCULATIONAHA.109.865444

2. Barter PJ, Caulfield M, Eriksson M, Grundy SM, Kastelein JJP, Komajda M, et al. Effects of Torcetrapib in Patients at High Risk for Coronary Events. N Engl J Med. 2007;357: 2109–2122. doi:10.1056/NEJMoa0706628

3. Swerdlow DI, Preiss D, Kuchenbaecker KB, Holmes MV, Engmann JEL, Shah T, et al. HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. The Lancet. 2014;385: 351–361. doi:10.1016/S0140-6736(14)61183-1

4. Cannon CP, Blazing MA, Giugliano RP, McCagg A, White JA, Theroux P, Darius H, Lewis BS, Ophuis TO, Jukema JW, De Ferrari GM, Ruzyllo W, De Lucca P, Im K, Bohula EA, Reist C, Wiviott SD, Tershakovec AM, Musliner TA, Braunwald E, Califf RM; IMPROVE-IT Investigators.. N Engl J Med. 2015 Jun 18;372(25):2387-97

5. The Myocardial Infarction Genetics Consortium Investigators Inactivating mutations in NPC1L1 and protection from coronary heart disease. N Engl J Med 2014; 371:2072-2082

6. Schmidt AF, Pearce LS, Wilkins JT, Overington JP, Hingorani AD, Casas JP PCSK9 monoclonal antibodies for the primary and secondary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2017 Apr 28;4:CD011748. doi: 10.1002/14651858.CD011748.pub2

7. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006;354:1264–72

8. Marso SP, Daniels GH, Brown-Frandsen K et al. for the LEADER Steering Committee on behalf of the LEADER Trial Investigators Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes N Engl J Med 2016; 375:311-322

9. Scott RA, Freitag DF, Li L, et al. Genomic approach to therapeutic target validation identifies a glucose-lowering *GLP1R* variant protective for coronary heart disease Sci Transl Med. 2016 Jun 1; 8(341): 341ra76. doi: 10.1126/scitranslmed.aad3744

10. Darapladib for Preventing Ischemic Events in Stable Coronary Heart Disease. N Engl J Med. 2014;370: 1702–1711. doi:10.1056/NEJMoa1315878

11. O'Donoghue ML, Braunwald E, White HD, et al. Effect of darapladib on major coronary events after an acute coronary syndrome: The SOLID-TIMI-52 randomized clinical trial. JAMA. 2014;312: 1006–1015. doi:10.1001/jama.2014.11061

12. Casas JP, Ninio E, Panayiotou A, Palmen J, Cooper JA, Ricketts SL, et al. PLA2G7 Genotype, Lipoprotein-Associated Phospholipase A2 Activity, and Coronary Heart Disease Risk in 10 494 Cases and 15 624 Controls of European Ancestry. Circulation. 2010;121: 2284–2293. doi:10.1161/CIRCULATIONAHA.109.923383

13. Millwood IY, Bennett DA, Walters RG, Clarke R, Waterworth D, Johnson T, Chen Y, Yang L, Guo Y, Bian Z, Hacker A, Yeo A, Parish S, Hill MR, Chissoe S, Peto R, Cardon L, **Collins R**, Li L, **Chen Z**; China Kadoorie Biobank Collaborative Group. Lipoprotein-Associated Phospholipase A2 Loss-of-Function Variant and Risk of Vascular Diseases in 90,000 Chinese Adults. J Am Coll Cardiol. 2016 Jan 19;67(2):230-1

14. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. The Lancet. 2012;379: 1214–1224. doi:10.1016/S0140-6736(12)60110-X

15. Casas JP, Shah T, Hingorani AD, Danesh J, Pepys MB. C-reactive protein and coronary heart disease: a critical review. J Intern Med. 2008;264: 295–314. doi:10.1111/j.1365-2796.2008.02015.x

16. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ. 2011;342: d548–d548. doi:10.1136/bmj.d548

17. Nicholls SJ, Kastelein JP, Schwartz GG, et al. Varespladib and cardiovascular events in patients with an acute coronary syndrome: The VISTA-16 randomized clinical trial. JAMA. 2014;311: 252–262. doi:10.1001/jama.2013.282836

18. Holmes MV, Simon T, Exeter HJ, Folkersen L, Asselbergs FW, Guardiola M, et al. Secretory phospholipase A(2)-IIA and cardiovascular disease: a mendelian randomization study. J Am Coll Cardiol. 2013;62: 1966–76. doi:10.1016/j.jacc.2013.06.044

19. Martin RIR, Pogoryelova O, Koref MS, Bourke JP, Teare MD, Keavney BD. Atrial fibrillation associated with ivabradine treatment: meta-analysis of randomised controlled trials. Heart. 2014 Oct 1; 100(19): 1506–1510.

20. Ellinor PT, Lunetta KL, Albert CM, et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. Nat Genet 2012;44:670–5

21. den Hoed M, Eijgelsheim M, Esko T, et al. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. Nat Genet 2013;45:621–31

22. van Oosten BW, et al. Increased MRI activity and immune activation in two multiple sclerosis patients treated with the monoclonal anti-tumor necrosis factor antibody cA2. Neurology. 1996;47:1531–1534.

23. The Lenercept Multiple Sclerosis Study Group. The University of British Columbia MS/MRI Analysis Group TNF neutralization in MS: results of a randomized, placebo-controlled multicenter study. Neurology. 1999;53:457–465

24. Gregory AP, Dendrou CA., Attfield KE et al. TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis *Nature* 2012; 488: 508–511