Mendelian Randomization and Estimation of Treatment Efficacy for Chronic Diseases

C.M. Schooling*, G. Freeman, and B.J. Cowling

* Correspondence to C.M. Schooling, CUNY School of Public Health at Hunter College, 2180 Third Avenue, New York, NY 10035 (e-mail: mschooli@hunter.cuny.edu).

Initially submitted January 30, 2012; accepted for publication August 8, 2012.

Prevention and treatment of common noncommunicable chronic diseases have been revolutionized by the development of therapies. Recently, several randomized controlled trials (RCTs) designed to assess the efficacy of new therapies targeted at well-established risk factors for noncommunicable chronic diseases have reported lower benefits than expected. Subsequent observational analysis of the same trial data has not clarified these unexpected findings. Mendelian randomization (MR) provides an approach for estimating causal effects from observational or trial data and thus provides information complementary to that from an RCT. An RCT assesses the efficacy of a therapy but does not usually confirm the underlying mechanistic pathway. In contrast, an MR study does not assess the efficacy of a therapy but rather assesses causal effects on an underlying mechanistic pathway. We suggest that incorporating an MR study into an RCT at the design stage would improve etiologic understanding of current therapies and enhance the search for therapies for the significant amount of noncommunicable chronic diseases that resists current treatments.

causal effects; efficacy; Mendelian randomization; randomized controlled trial; study design

Abbreviations: DAG, directed acyclic graph; MR, Mendelian randomization; NCD, noncommunicable chronic disease; RCT, randomized controlled trial.

Therapies identified in the last 40 years have revolutionized the prevention and treatment of many noncommunicable chronic diseases (NCDs), such as ischemic heart disease. Notwithstanding these successes, a significant amount of NCDs remains difficult to treat, and new therapies for those diseases are continually being sought. Recently, there have been several large, rigorously designed, well-executed randomized controlled trials (RCTs) of key NCD therapies that singly or in meta-analyses have produced unexpected or equivocal results (1–6). These RCTs have used therapies that have successfully modulated well-established and reliable risk factors for NCDs, such as by raising high-density lipoprotein cholesterol levels (1, 6), lowering low-density lipoprotein cholesterol levels (2, 6), or implementing intensive glucose, lipid, or blood pressure control in people with diabetes (3–5). However, these therapies did not deliver the expected benefits on cardiovascular or all-cause mortality rates (1–6). Such RCT results have created uncertainty for patients and practitioners (7–9). In some RCTs, subsequent observational analysis of the data has successfully identified potential explanations for the unexpected results, such as off-target effects of torcetrapib on aldosterone, cortisol, or blood pressure (10), nullifying the beneficial lipid modulation (2). However, subsequent observational analyses of data from RCTs has not always been so revealing (11–14). Other approaches to elucidating these RCT results have included suggesting further RCTs of a larger size, in different conditions, or with a pragmatic design (3, 9, 15), although these may take many years to produce results and meanwhile expose patients to known harms.

Elucidating these unexpected RCT findings is vital to the prevention of, the development of effective therapies for, and the understanding of the etiology of NCDs. Here we consider whether formally integrating a complementary randomized study design into RCTs would facilitate elucidation of unexpected results. Both RCTs and Mendelian randomization...
(MR) studies (16) use random allocation of exposure by investigator or genetic lottery, respectively, to assess causal effects, and these studies are often seen as similar (17–19). MR studies may be described in terms of a design analogous to an RCT, which provides a method of assessing causal effects in observational data (17–19). MR studies have typically been used in the context of building causal arguments from observational data in which interventions were not randomized to participants, such as when assessing the causal role of C-reactive protein in coronary heart disease (20) or of moderate alcohol use in cognitive function (21). However, RCTs and MR studies answer different but complementary questions. In the present study, we delineate the differences between the questions answered by RCTs and those answered by MR studies and explain how MR could be incorporated into RCTs to maximize the information provided.

**RELATIONS ASSESSED IN RCTs COMPARED WITH THOSE IN MR STUDIES**

Figure 1 shows the different relations assessed in an ideal RCT and an ideal MR study using an instrumental variable analysis with genetic instruments that only influence the risk of disease through a biomarker. An RCT to evaluate treatment of an NCD typically tests a therapy designed to modulate a biomarker of a specific risk factor to prevent disease, and results are potentially assessed from death rates, cause-specific death rates, a composite endpoint, some measure of morbidity, or an intermediate disease marker. An RCT can be used to assess the efficacy of the therapy but cannot determine whether the therapy acts via modulation of the target biomarker or via other mechanisms, such as another factor, or acts directly (Figure 1A). In contrast, an MR study uses genetically determined differences in exposure to test whether a biomarker affects disease (Figure 1B), and thus it assesses the role of a specific biomarker on the postulated causal pathway (16, 18), assuming the biomarker is expressed and active. However, an MR study cannot demonstrate whether a therapy affecting a biomarker will change the risk of disease because that therapy might also increase or decrease the risk of disease through other factors.

**INFORMATION ABOUT CAUSAL EFFECTS PROVIDED BY RCTs AND MR STUDIES**

To illustrate the different information about causal effects provided by RCTs and MR studies, Figure 2 shows the different underlying causal effects as directed acyclic graphs (DAGs) (22). For illustrative purposes, these are simple examples; however, the same principles would apply in more complex situations in which, for example, a therapy acts on several biomarkers. In these diagrams, a solid arrow indicates a causal effect on the indicated pathway, whereas the absence of a solid arrow indicates that there may be an association (which for simplicity we have not shown) but it is not due to a causal effect.

Figure 2A is a DAG illustrating a case in which a therapy acts on a biomarker that causes disease. Given the multifactorial nature of most common chronic diseases, it also shows another factor, unrelated to the therapy or biomarker that may also have causal effects on the disease. Nevertheless, in this situation the therapy would modulate the biomarker in the expected direction, and an RCT would show the therapy to be efficacious. An MR study would confirm that the biomarker had a causal effect.

Figure 2B is a DAG illustrating a case in which a therapy acts on a biomarker that does not cause the disease and also acts on another factor that has causal effects on the disease. In this situation, the therapy would modulate the biomarker in the expected direction, and an RCT would show the therapy to be efficacious. Whether the effect of the therapy on the disease matched that predicted from its observed effects on the biomarker would depend on the correlation between the other factor and the biomarker. However, an MR study would indicate that the biomarker was not causal. A typical example of this situation would be different classes of drugs that target the same biomarker but have different effects due to some but not others serendipitously modulating the other factor.

Figure 2C is a DAG showing a case in which a therapy does not act on the biomarker and the biomarker does not cause the disease but the therapy does act on another factor, which has causal effects on the disease and the biomarker. In this situation, the therapy would again modulate the biomarker in the expected direction and an RCT would show the therapy to be efficacious. However, whether the effect of the therapy on the disease matched that predicted from its observed effects on the biomarker would depend on the causal effects of the other factor on the biomarker. However, an MR study would indicate that the biomarker was not causal.

Figure 2D shows a DAG in which a therapy acts on a biomarker that does not cause the disease and does not affect the other factor. In this situation, the therapy would

---

Figure 1. Relations tested (dashed line) in a randomized controlled trial (A) compared with a in Mendelian randomization study (B), where solid lines indicate potential causal effects.
modulate the biomarker in the expected direction, but an RCT would show the therapy to not be efficacious. An MR study would also indicate that the biomarker was not causal.

Figure 2E shows a DAG for a case in which a therapy does not act on a biomarker but acts on another factor that has causal effects on the disease. In this situation, an RCT would show the therapy to be unexpectedly efficacious, but the biomarker would be unaffected. An MR study would also indicate that the biomarker was not causal.

**HOW MR STUDIES COULD COMPLEMENT RCTs**

Figure 2 shows that different causal relations (Figure 2A compared with Figure 2B, 2C, and 2E) can lead to the same findings in an RCT, but these different underlying biological mechanisms could be clarified by use of an MR study in the same sample. If we simply relied on using therapies that had already been identified as effective in RCTs, this distinction would not matter, although we might not be sure whether therapies would be as effective in other populations with different distributions of risk factors. However, it is preferable to define the biological pathways through which a therapy produces its effects. In addition, new therapies are needed and are continually being developed. In RCTs of these therapies, adding a “dormant” MR component that would only be activated and implemented by a particular set of findings would provide an inexpensive and timely means of clarifying the mechanisms behind any unexpected RCT results.

A dormant MR component of an RCT would only need to consist of suitably stored biological samples obtained at

---

### Table: Directed Acyclic Graphs

<table>
<thead>
<tr>
<th>Directed Acyclic Graph</th>
<th>RCT shows therapy to be efficacious</th>
<th>MR indicates biomarker has causal effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>A)</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>B)</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>C)</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>D)</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>E)</td>
<td>YES</td>
<td>NO</td>
</tr>
</tbody>
</table>

**Figure 2.** Information provided by randomized controlled trials (RCTs) and Mendelian randomization (MR) studies with respect to different underlying causal effects (solid lines), as represented by directed acyclic graphs.
which can be statistically complex (25). MR studies typically use instrumental variable analysis, causal effects of that biomarker on a disease is simple of using genetic variation in a biomarker to deduce the implications for drugs in the same class. A dormant MR study could also provide invaluable information about successful trials. An MR study could be used to verify that predicted using external knowledge. A smaller than expected effect that required investigation. If the biomarker had no causal effects, it would imply the wrong target, with corresponding implications for drugs in the same class. A dormant MR study could also provide invaluable information about successful trials. An MR study could be used to verify that the therapy was causal and had an effect size match- ing different questions about causal effects, as shown in Figure 1B. A larger than expected effect in MR might suggest the presence of an additional off-target detrimental effect that cancelled out some of the benefit of the on-target effect.

Finally, given the disease outcomes and biomaterials available, a dormant MR component of an RCT could also be used to examine other potential biological pathways. This would be particularly valuable in situations in which the therapy being examined appeared to have an unexpectedly beneficial effect on another disease outcome (Figure 2E) and it was unclear whether such a finding was the chance result of multiple testing or a real discovery. For example, one study (23) recently found vitamin E to have a therapeutic effect on prostate cancer, whereas another (24) refuted those findings. One set of biological samples obtained at recruitment could be reused many times to examine potential other factors simply by assessing that factor and its genetic determinants in the stored samples. In fact, just as genetic data from major epidemiologic studies are increasingly open to the research community, so perhaps could data from major RCTs be.

MR studies have the potential to add to RCTs by addressing different questions about causal effects, as shown in Figures 1 and 2, and could also provide information even if there was low compliance with the RCT; however, as with all study designs, MR studies have limitations. Moreover, there are potentially issues specific to implementing an MR study within an RCT.

LIMITATIONS OF MR STUDIES

MR studies may be seen as complex. However, the idea of using genetic variation in a biomarker to deduce the causal effects of that biomarker on a disease is simple (Figure 1B) and has been in practice for over 20 years (16). MR studies typically use instrumental variable analysis, which can be statistically complex (25–27). However, the methods are similar to the evaluation of mediation and statistically identical to those used to adjust for treatment contamination in RCTs (28). Nevertheless, because of the indirect estimation of causal effects in MR studies (Figure 1B), larger sample sizes are needed, and power analysis can require simulation rather than analytic solutions. However, sample sizes required for various scenarios have been determined (29). Moreover, for a linear outcome with a normally distributed exposure, the MR sample size can be estimated as the sample size for exposure on disease (from a 1-sample t test) divided by the squared correlation between the gene and the exposure (30). For example, the sample size needed for a linear regression to detect an effect size of 0.25 for exposure on disease (i.e., an effect size of 0.25 for the regression β coefficient) with 80% power and a 5% α is 129, but the equivalent MR analysis with squared correlation for gene on exposure of 0.03 would require a sample size of 129 divided by 0.03, or 4300.

MR studies also have specific requirements (18). First, an MR study requires a genetic instrument that predicts the biomarker, for which genetic discovery is continually providing new materials (31). However, genetic instruments susceptible to epigenetic modification might be less suitable for MR studies because their association with the biomarker could be weaker or vary with factors that drive gene expression; alternatively, epigenetic changes might induce an association of a genetic instrument with disease in a way other than through the biomarker. Second, MR studies can be confounded by a factor related to both the gene and the disease, such as unmeasured population stratification. Because of this, investigators conducting MR studies need to use homogeneous populations or to identify and include population strata in the analyses. Third, the genetic instrument should only be associated with the disease through the biomarker; that is, it should not have pleiotropic effects. Pleiotropy could be detected from effects of the genetic instrument on the disease in subgroups in which the genetic instrument does not affect the biomarker or from different MR estimates for the effects of the biomarker on the disease when using different genetic instruments for the same biomarker (32). Moreover, MR estimates should also be expected to be coherent with other evidence from a variety of sources, including historical trends, geo-ethnic differences, RCT results, and even animal experiments. Finally, the effects of genetic instruments do tend to be rather small, which can be addressed by using larger sample sizes or multiple genetic variants as instruments (33).

POTENTIAL ISSUES WITH INCLUDING MR STUDIES IN RCTs

Currently, there may also be practical barriers to embedding complementary dormant MR studies in RCTs. First, an MR study, unlike other observational analyses of RCT data, cannot be easily implemented after the RCT is completed because an MR study requires not only stored specimens but also consent from the participants for human genetic testing. Inclusion of a dormant MR component into an RCT needs to occur at the inception of the RCT design, including consideration of potential issues with an MR study, such as population stratification. Even an MR study that seems...
infeasible at RCT inception because of a lack of genetic instruments may be possible by the time the trial is finished. Second, MR studies need large sample sizes (29). However, the sample sizes required are similar to those in many of the recent major RCTs (4, 6, 11, 14). Moreover, sample size could be increased by consortia of collaborators pooling RCTs. Third, genotyping all of the participants in an RCT could be costly. A nested case-control design for the MR study in which all trial participants with the disease were sampled in addition to a random sample of those without would reduce the costs of genotyping. Fourth, reusing RCT data to conduct an MR study would always be possible for the RCT placebo arm. Various issues could arise with also using the RCT treatment arm. For example, the therapy could remove the association of the genetic variant with biomarker by reducing the biomarker to undetectable levels, in which case there would be little association of any genetic variant with that biomarker and an MR study would not be possible for that biomarker in the RCT treatment arm. Another example would be a therapy that reversed the association of the genetic variant with the biomarker without affecting disease; in this case, the standard MR estimate could be biased. Pragmatically, such possibilities could usually be detected if there were quantitatively different MR estimates from each RCT arm, and investigators could identify the best way, if any, to make use of the data from the RCT treatment arm. Fifth, reusing RCT data to conduct an MR study would not necessarily distinguish between the causal relations shown in Figure 2B and 2C. However, explicit consideration of the underlying causal relations and an explicit search for the other factor in subsequent MR studies might do so. Sixth, RCTs may include a very selected population. RCT participants may be selected because of genetic vulnerability to the disease under study and thus may over-represent a particular genotype. However, such genetic vulnerability should not alter the specific associations of genetic variants with a phenotype or with disease. Seventh, results from MR studies may be seen as difficult to interpret because there are no standard reporting guidelines and checklists, which could easily be addressed by amending the current guidelines for RCT protocols to include MR studies and by developing guidelines for the reporting of MR studies.

Finally, there are several situations in which including an MR study in an RCT could be more complex but perhaps also more revealing. One example would be a multifactorial RCT at a community level that used cluster randomization and was designed to reduce chronic diseases by targeting interventions at several risk factors, such as overweight and depression, as well as to assess disease-related outcomes. The sample size might be large. Randomization by investigators and genetic lottery should be independent so that the specific randomization scheme would not affect the MR study. An MR study could help identify whether any changes in disease were due to changes in the targeted risk factors or to other effects of the intervention. On the other hand, such a setting might make it more challenging to acquire the relevant biomaterials and might include a heterogeneous population, the characteristics of which would need to be assessed and recorded to avoid population stratification. A second example would be an RCT from a setting with a double burden of infectious and noncommunicable diseases. Although infectious diseases, or perhaps the absence of infectious diseases, may relate to chronic diseases, infectious diseases should not relate to genetic determinants of chronic disease biomarkers unless there are pleiotropic genetic effects. However, informative differences are also possible. In developed settings, key biomarkers, such as lipids, blood pressure, or fasting glucose, may not cover the full biologically possible range; an MR gives causal effects for the observed range. However, in other settings, the range of exposures may be different, giving causal effects in a different part of the biologically possible range, which would help clarify whether causal effects are monotonic or have more complex curvilinear effects. On the other hand, there could be an issue with combining RCTs from very different populations because of population stratification, which could be addressed by identifying and incorporating population strata in the analysis. Finally, a third example concerns personalized therapy. If a genetic variant and an effective treatment both influenced the biomarker, then the genetic variant would be an excellent candidate for effect-measure modification, in which both the treatment and the genetic variant potentially operated through the same biologic pathway. If the genetic variant were, in fact, an effect modifier, it could be used clinically to identify a subcategory of patients who would particularly benefit (or not) under the studied treatment, with correspondingly more targeted and effective treatment regimes for NCDs.

CONCLUSION

We have described and justified an inexpensive adjunct to RCTs that could be implemented at little additional cost and activated as needed to improve causal inference. Such an approach is in keeping with the current transition to greater reliance on experimental or quasieperimental evidence. Designing RCTs to include MR studies could also speed up the search for new prevention and treatment strategies, particularly for NCDs, and improve understanding of etiologic pathways, with corresponding benefits including facilitating the building of the causal structures needed to test hypotheses with epidemiologic dataand facilitating the development of even more effective treatments.

ACKNOWLEDGMENTS

Author affiliations: School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China (C.M. Schooling, G. Freeman, B.J. Cowling); and School of Public Health at Hunter College, The City University of New York, New York (C.M. Schooling).

Conflict of interest: B.J. Cowling has received research funding from MedImmune, Inc., and consults for Crucell MV. The authors report no other potential conflicts of interest.
