This is a nice, crisp, thoughtful review of an important topic. Addressing the comments below would hopefully make it even better.

**We thank the reviewer for the positive and helpful comments.**

In the introduction where MR is compared to RCT, I think this would be a good time to mention that there is one known confounder (population stratification) that makes the MR not a “random” process (people from different ancestries may be more or less likely to inherit a particularly trait-increasing allele, and ancestry will be confounded with lots of other things). In the Box, the proposed correction for population stratification is probably not adequate as “homogeneous” populations can still have stratification – the “and/or” should probably be “and” – homogeneous populations analyzed separately AND principal components or another method for detecting and correcting for stratification.

**Stratification has been mentioned specifically in the RCT section, and 'and/or' has been modified to read 'and'**

Unlike stratification, which is fairly easy to detect and correct for, I think pleiotropy may be harder and deserves a little more discussion as a limitation, especially when using polygenic risk scores. You already discuss it, but mostly in the context of where it’s not worrisome -- type II pleiotropy -- and I think it’s important enough to expand on further. If there are two measured traits, both of which have predictors, then the bidirectional MR can be helpful, but if one trait is unmeasured or has no instruments, then MR could still lead to incorrect conclusions about causality where the real causal factor is the unmeasured trait. Obviously, the more direct the effect of the instrument is on the trait (eg CRP variants and CRP levels), the less chance there is that pleiotropy can intervene. But in polygenic scores, the likelihood increases that some or many of the associated variants will have pleiotropic effects. So, I think emphasizing that the best case scenario is a well-powered clearly directly (in addition to reliably) linked set of instruments, and that more complexities emerge as one moves away from that scenario.

**A concluding sentence has been added to the Box on pleiotropy that emphasises the scenario in which MR can be interpreted with most clarity.**

Another issue that is somewhat related to pleiotropy is whether the “right” trait has been measured. I’m not sure there is anything concrete I have to suggest here, but let’s say that one is studying serum triglycerides or CRP, and MR shows no causality. It turns out that the actual causal traits are triglyeride flux through the liver (which some of the TG-increasing variants might increase and others decrease and others not affect at all) or rises in CRP in the setting of infection (where perhaps the CRP gene variants might not have an effect but other more indirect, pleiotropic variants might show an effect). Is there a way to get a clue that something like this might be going on, without measuring the actual causal trait – for example, more association with outcome of the individual instruments than expected by chance, or pathway analysis to group the instruments and re-do the MR analysis? Again, not sure there is enough here even to speculate, but at least a sentence or two on having measured the “right” trait seems worthwhile.

**This has now been discussed as "complexity of association" along side the section on multiphenotype MR**

For the maternal influences cited in the Table, does one need to account for the correlation between maternal and child genotype?

**This has now been detailed in the table.**

For those attempting MR (especially the more complex versions like two-sample MR), are there specific software tools you can point them to in Box 1 or elsewhere, beyond citing references? Are there critical missteps in implementation to avoid in applying these tools that you can point out in this Box or elsewhere and that you haven’t already mentioned?

**We have now included a list of published software that is available for the use of various implementations of IV analysis in Box 1**

I couldn’t tell – do you think there is a major flaw in the HDL/triglyceride studies because they used regression? If so, can you be a little more expansive about what you would propose?

**This has been clarified as being an important first step that still requires methodological development**

Can you expand a little bit on the last point about Mendelian diseases? Would you also look for effects in carriers (eg GBA and Parkinson’s)?

**This has been clarified**

Can the author reference some of the failures of observational epidemiology cited in the introduction?

**These have been added to the first paragraph.**

Page 3 line 21, reference 5 needs to be superscripted, or all of them unsuperscripted (which may be the right decision for references following an abbreviation)

**Amended to superscript**