The Future of Mendelian Randomization Studies: MR Data Challenge December 2021

Instructions

Challenge Aims

The Data Challenge aims to highlight some of the methodological challenges of inferring causal effects from real-world data using Mendelian Randomization (MR) and provide a concrete anchor for discussions of these complexities among conference attendees. In particular, we encourage participants to consider issues of selecting potential instruments, clearly defining causal estimands, and timing. To complete the data challenge, you will need the following prompt and the datasets provided.

Schedule and Checkpoints

The Data Challenge will be released prior to our virtual meeting. We encourage you to form your teams (recommended N between 2 and 6 members) before the start of the meeting. Email Joy Shi (joyshi@hsph.harvard.edu) with your team members so we can keep track of teams. You can also email if you want help identifying potential teammates.

You are welcome to work on the Data Challenge at your own pace prior to and during the meeting. Because of our many time zones, we suspect you will work with your team during times that are conducive to your own schedules. Within the meeting's virtual environment, a board will be formed for cross-team discussion. We will also have short live meetings during the week for further cross-team discussion. These meetings will be attended by Elizabeth Diemer and Joy Shi, who led the creation of this Data Challenge and can also answer any practical questions about the datasets.

Please submit your final answers by 20.00hr Thursday December 16 to Joy Shi (joyshi@hsph.harvard.edu). We will collate your answers for a culminating discussion on Friday December 17. If possible, please ensure at least one member of each team will be present for the live Friday discussion.

Background

Researchers have recently identified a new blood biomarker, "Marker A", that occurs naturally within the body and whose concentration varies over time. Marker A is associated with many health outcomes related to insulin resistance, including diabetes and polycystic ovarian syndrome. Maternal levels of Marker A during pregnancy are further associated with offspring macrosomia (i.e., high birthweight), and has been demonstrated to cross the placenta barrier in animal studies. However, it is currently unclear whether Marker A causally affects these outcomes, and there are concerns that studies using non-MR methods are biased as a result of unmeasured confounding. Complicating the intergenerational findings further, Marker A is also associated with both childlessness and decreased fertility, though the reasons for these associations are unknown.

A recent GWAS has identified several single nucleotide polymorphisms (SNPs) associated with Marker A. We would like you to use these findings, along with the provided background knowledge, and below-described data, to estimate the causal effect of Marker A in utero on offspring macrosomia risk and the effect of exposure to Marker A across the life-course on diabetes risk using MR-based methods (i.e., instrumental variable estimation with genetic variants as proposed instruments for Marker A).

Definitions

Throughout the document, we refer to the exposure of Marker A as A (a continuous variable that can take on values ranging from 0 to 50), the outcome of offspring macrosomia (birthweight > 4000g, also referred to as high birthweight) as B, and the outcome of maternal diabetes as Y. Counterfactuals are indicated with superscripts (e.g., $Y^{a=1}$ is an indicator for maternal diabetes had Marker A been set to 1), and subscripts are used to index time (e.g., Y_k is an indicator of maternal diabetes at time k). Overbars denote treatment history up to the indexed time; that is, $\bar{A}_t = (A_0, A_1, \dots, A_t)$ denotes the values of Marker A at each time point from time 0 to time t. Underbars are used to denote future values of the treatment from the indexed time t until the end of follow-up at time K; that is, $\underline{A}_t = (A_t, A_{t+1}, \dots, A_K)$. In general, we will use k to index the time at which the outcome occurred, and t to index the time at which the exposure was measured.

Available Data

50,000 women were enrolled in a longitudinal biobank study at the age of 25 and prospectively followed for 30 years. Blood samples were collected at baseline and used for genotyping. Participants also provided blood samples at 5-year intervals over the duration of follow-up. Given the recent discovery and interest in Marker A, these blood samples have been assayed for Marker A levels. Throughout follow-up, incidence of various health outcomes, including diabetes, were collected via self-report and subsequently confirmed by medical records.

A substudy nested within this cohort was developed to examine offspring health outcomes. All women who became pregnant during the course of the study were recruited into this additional study, which linked maternal data to offspring outcome data, and collected paternal data when fathers consented to blood collection.

To estimate the causal effect of Marker A in utero on offspring macrosomia risk and the effect of exposure to Marker A across the life-course on diabetes, we have provided the following datasets:

1. ***MR_data_challenge_p2.csv***: This dataset contains data on Marker A levels at baseline and during pregnancy, and offspring outcomes for participants included in the pregnancy substudy.
2. ***MR_data_challenge_p3.csv***: This dataset contains longitudinal measurements on Marker A levels and diabetes status for the full cohort.

Part 1: Identifying your candidate instruments

The most recent and largest genome-wide association study (GWAS) identified 11 SNPs that were significantly associated with Marker A (based on a p-value threshold of 5×10^{-8}). These 11 SNPS are found across 4 different chromosomes:

- Chromosome 3: two *cis*-acting SNPs in a non-coding region which is suspected to affect mRNA expression levels of a nearby gene; this gene encodes a protein which is believed to be involved with the metabolism of Marker A
- Chromosome 4: a functional variant (or one that is in LD with a functional variant) in a gene that encodes a protein which catalyzes the conversion of Marker Pre-A to Marker A
- Chromosome 9: five SNPs in non-coding regions with unknown function
- Chromosome 17: three *cis*-regulatory variants for a gene involved in regulating the follicle-stimulating hormone

Using the information above, the dataset provided (which includes individual-level data on these SNPs, the exposures and the outcomes) and any falsification tools or strategies of your choice, identify which SNPs you would like to incorporate into your analysis in Part 2 and Part 3. Note that

each part aims to estimate a different effect and thus you may choose the same or a separate set of SNPs for each analysis.

a. List the SNPs incorporated into your analysis in Part 2.
b. List the SNPs incorporated into your analysis in Part 3.

Part 2: In utero exposure to "A"

We are interested in estimating the effect of exposure to maternal Marker A in utero on offspring high birthweight (macrosomia). Note that in this question *A* refers to *maternal* levels of Marker A, and the outcome *B* refers to *offspring* outcome. This differentiation may be especially important if offspring Marker A affects macrosomia, given that offspring genetic variants are partly determined by maternal genetics. For simplicity of notation, we do not index *A* over time, as *A* is only assessed once during pregnancy in the dataset to be used in Part 2.

a. Estimate the average causal effect of increasing maternal Marker A from 50 units to 60 units during pregnancy on offspring high birthweight, i.e., $\mathrm{E}[B^{a=60}]/\mathrm{E}[B^{a=50}]$. In presenting your results, articulate all assumptions made that underlie your chosen approach. Results should include a 95% confidence interval, as appropriate.
b. The effect we have asked you to estimate may not be what you would consider the most meaningful or interesting estimand to estimate or causal null hypothesis to test. If you have other causal estimands or tests of interest, describe and justify them. Use your preferred method to address any additional questions.
c. During the workshop, we also want to engage in a rich discussion of what substantive and methodologic difficulties arose during the course of the data challenge. Please list the 5 biggest difficulties your team identified during this portion of the challenge and be prepared to discuss how these issues impacted your analysis.

Part 3: Long-term exposure to "A" in women

We also want to estimate sustained effects of Marker A on the cumulative risk of diabetes. Marker A is lipophilic, and therefore believed to accumulate in adipose tissues over time. However, the critical period of exposure is suspected to be around 35-40 years of age.

a. Estimate the "lifetime" average causal effect of increasing maternal Marker A from 50 units to 60 units on the cumulative risk of diabetes by the end of follow-up, i.e., $\mathrm{E}\left[Y_K^{\bar{a}_K=\overline{60}}\right] / \mathrm{E}\left[Y_K^{\bar{a}_K=\overline{50}}\right]$. In presenting your results, articulate all assumptions made that underlie your chosen approach. Results should include a 95% confidence interval, as appropriate.
b. Estimate the average controlled direct effect of increasing maternal Marker A from 50 units to 60 units during ages 35 to 40 on the cumulative risk of diabetes by the end of follow-up, i.e., $\mathrm{E}\left[Y_k^{\bar{a}_{t<35}=50, a_{35\leq t\leq 40}=60, \underline{a}_{t>40}=50}\right] / \mathrm{E}\left[Y_K^{\bar{a}_K=\overline{50}}\right]$. In presenting your results, articulate all assumptions made that underlie your chosen approach. Results should include a 95% confidence interval, as appropriate.
c. The effects we have asked you to estimate may not be what you would consider the most meaningful or interesting estimand to estimate or causal null hypothesis to test. If you have other causal estimands or tests of interest, describe and justify them. Use your preferred method to address any additional questions.
d. During the workshop, we also want to engage in a rich discussion of what substantive and methodologic difficulties arose during the course of the data challenge. Please list the 5

biggest difficulties your team identified during this portion of the challenge and be prepared to discuss how these issues impacted your analysis.