Almost Exact Mendelian Randomization

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

April 20, 2023 @ European Causal Inference Meeting, Oslo, Norway

(Based on joint work with Matt Tudball and George Davey Smith: arXiv:2208.14035.)

The randomization principle in causal inference

We should use randomization in

1. The **design** of an **experiment**. (Nearly univerally adopted.)
2. The **analysis** of an **experiment**. (Repeatedly forgotten and brought back.)

We should mimic randomization in

3. The **design** of an **observational study**. (Repeatedly forgotten and brought back.)
4. The **analysis** of an **observational study**. (Never very popular.)

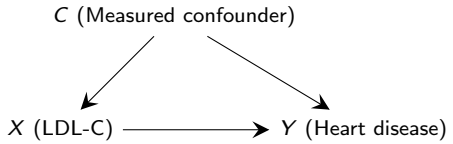
Main idea

Apply 3 & 4 to Mendelian randomization (MR)—using random genetic inheritance for causal inference.

Outline

No unmeasured confounders/ignorability/exchangeability

- In observational studies, it is typically assumed that all confounders X are measured, so that the study mimics a randomized experiment.



- Modern theory for causal graphical models interprets this as $A \perp\!\!\!\perp Y(a) \mid X$, but the role of randomization is obscure.
- For this reason, natural experiments such as MR are usually thought to be more credible.

Pre-history of MR

- Wright (1923), in a defence of his method of path coefficients, argues that the validity of this method "rests on the validity of the premises, i.e., on the evidence for Mendelian heridity", and the "universality" of Mendelian laws justifies ascribing a causal interpretation to his findings.[1]
- Fisher must have also known this by heart. Below are quotes from his 1951 Bateson lecture.

  And here I may mention a connection between our two subjects which seems not to be altogether accidental, namely that the **"factorial" method of experimentation** *. . . derives its structure, and its name, from the simultaneous inheritance of* **Mendelian factors**.

  Genetics is indeed in a peculiarly favoured condition in that Providence has shielded the geneticist from many of the difficulties of a reliably controlled comparison. **The different genotypes possible from the same mating have been beautifully randomised by the meiotic process.**

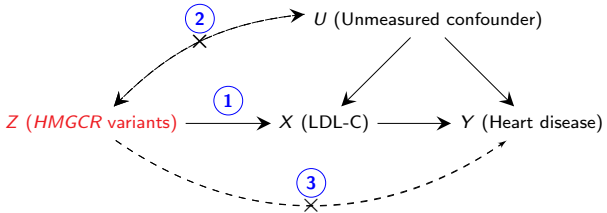
- Independent proposals appeared in 1970s-90s before the idea was brought to the front stage by Davey Smith and Ebrahim (2003).[2]

[1] Sewall Wright (1923). "The Theory of Path Coefficients: A Reply to Niles's Criticism". In: *Genetics* 8.3, pp. 239–255. DOI: 10.1093/genetics/8.3.239.

[2] George Davey Smith and Shah Ebrahim (2003). "'Mendelian Randomization': Can Genetic Epidemiology Contribute To Understanding Environmental Determinants of Disease?" In: *International Journal of Epidemiology* 32.1, pp. 1–22. DOI: 10.1093/ije/dyg070.

Modern interpretation of MR

- The most popular view is that genetic variants are used as **instrumental variables**.[3]



- This is the basis of most existing methodological and applied work, which often take advantage of the wealth of GWAS (summary) data.[4]
- But the role of randomization is still not entirely clear.

[3] Vanessa Didelez and Nuala Sheehan (Aug. 2007). "Mendelian Randomization as an Instrumental Variable Approach to Causal Inference". In: *Statistical Methods in Medical Research* 16.4, pp. 309–330. ISSN: 0962-2802. DOI: 10.1177/0962280206077743; Duncan C Thomas and David V Conti (Feb. 2004). "Commentary: The Concept of 'Mendelian Randomization'". In: *International Journal of Epidemiology* 33.1, pp. 21–25. ISSN: 0300-5771. DOI: 10.1093/ije/dyh048.

[4] For a recent review, see Eleanor Sanderson et al. (Feb. 2022). "Mendelian Randomization". In: *Nature Reviews Methods Primers* 2.1, pp. 1–21. ISSN: 2662-8449. DOI: 10.1038/s43586-021-00092-5.

Genetic inheritance as a natural experiment

International Journal of Epidemiology 2003;**32**:1–22
DOI: 10.1093/ije/dyg070

30TH THOMAS FRANCIS JR MEMORIAL LECTURE

'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*

George Davey Smith and Shah Ebrahim

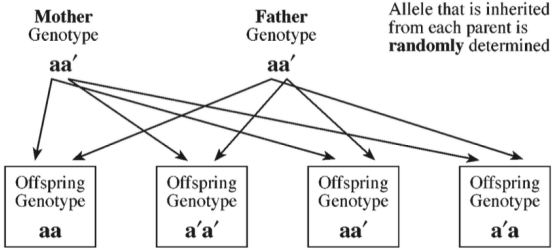


Figure 3 Mendelian randomization in parent–offspring design

Offspring should have an equal chance of receiving either of the alleles that the parents have at any particular locus

The basis of Mendelian randomization is most clearly seen in parent–offspring designs... A shift from this 50/50 ratio indicates an association between disease or phenotypic characteristic and the alleles at this locus (Figure 3)... Thus the Mendelian randomization in genetic association studies is approximate, rather than absolute.

Genetic trio studies

- Not surprisingly, Mendelian randomization has been used for mapping causal genetic variants.

Data: Genotypes and phenotypes of mother, father, and offspring.

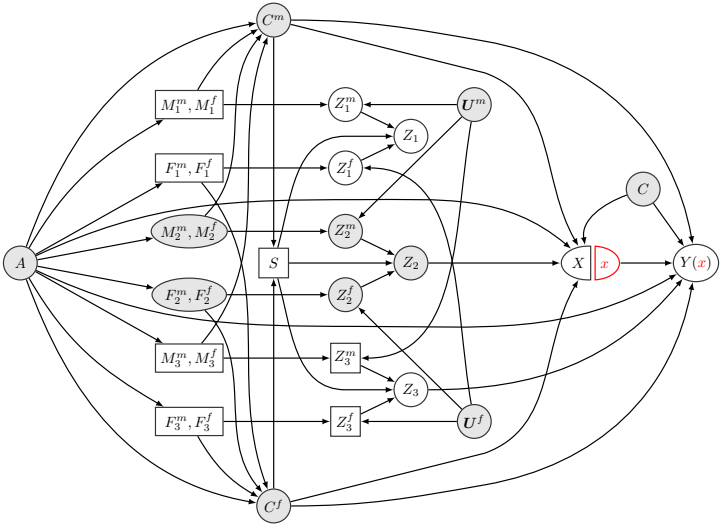
- $M/F/Z$: mother/father/offspring.
- Superscript f/m: Haplotypes inherited from father/mother.
- So $M_j^f \in \{0,1\}$ is mother's haplotype at locus j inherited from her father.
- No superscript means genotypes: $Z_j = Z_j^f + Z_j^m \in \{0,1,2\}$.
- Y: A phenotype of the offspring

- The **transmission disequilibrium test (TDT)** by Spielman, McGinnis, and Ewens (1993)[5] tests the conditional independence $\boldsymbol{Z}_j^{mf} \perp\!\!\!\perp Y \mid \boldsymbol{M}_j^{mf}, \boldsymbol{F}_j^{mf}$.
- Bates et al. (2020)[6] use existing meiosis models to obtain $\boldsymbol{Z}^{mf} \mid \boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f$.

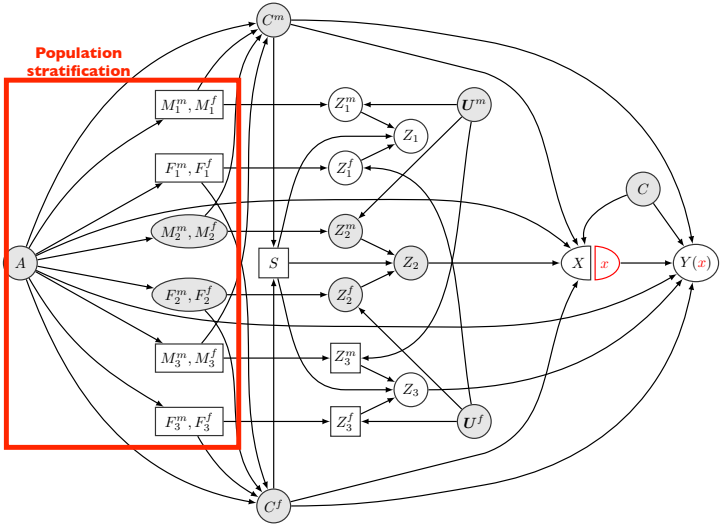
[5] R S Spielman, R E McGinnis, and W J Ewens (Mar. 1993). "Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-Dependent Diabetes Mellitus (IDDM).". In: *American Journal of Human Genetics* 52.3, pp. 506–516. ISSN: 0002-9297.

[6] Stephen Bates et al. (Sept. 2020). "Causal Inference in Genetic Trio Studies". In: *Proceedings of the National Academy of Sciences* 117.39, pp. 24117–24126. DOI: 10.1073/pnas.2007743117.

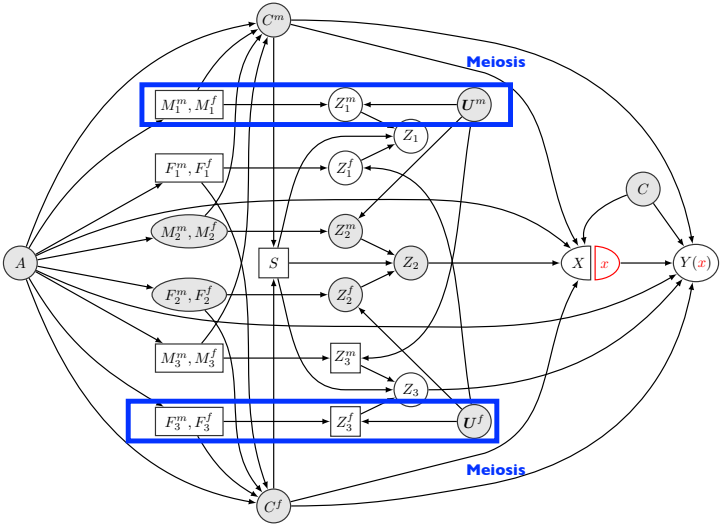
Illustration of within-family Mendelian randomization



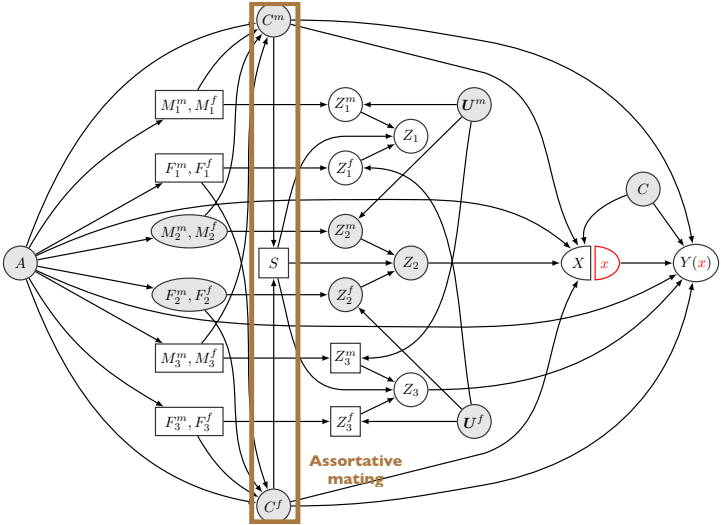
Graphical diagram for within-family Mendelian randomization



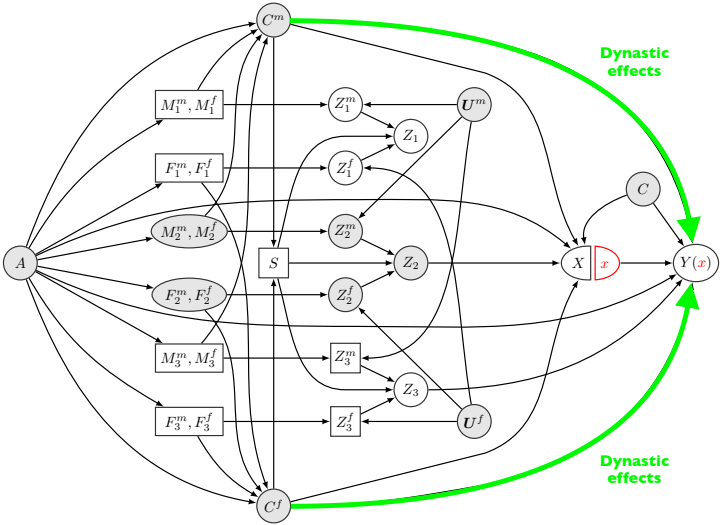
Graphical diagram for within-family Mendelian randomization



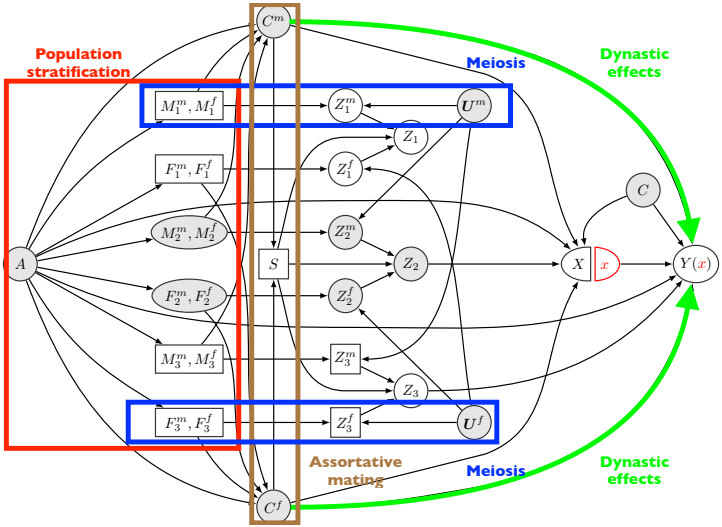
Graphical diagram for within-family Mendelian randomization



Graphical diagram for within-family Mendelian randomization



Graphical diagram for within-family Mendelian randomization



When is Z_j a valid IV?

Instrumental variable (IV)

Z_j is a valid IV given $\boldsymbol{V}$ (for estimating the causal effect of X on Y) if

1. Relevance: $Z_j \not\perp\!\!\!\perp X \mid \boldsymbol{V}$;
2. Exogeneity: $Z_j \perp\!\!\!\perp Y(x) \mid \boldsymbol{V}$ for all x;
3. Exclusion restriction: $Y(z_j, x) = Y(x)$ for all $z_j \in \{0, 1, 2\}$ and x.

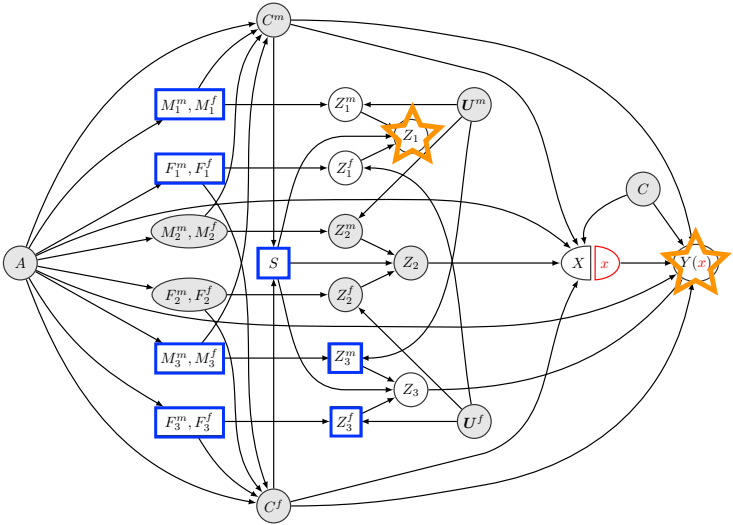
How should we choose $\boldsymbol{V}$? In the running example,

$$Z_1^m \perp\!\!\!\perp Y(x) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{V}_3^m, S = 1), \tag{1}$$

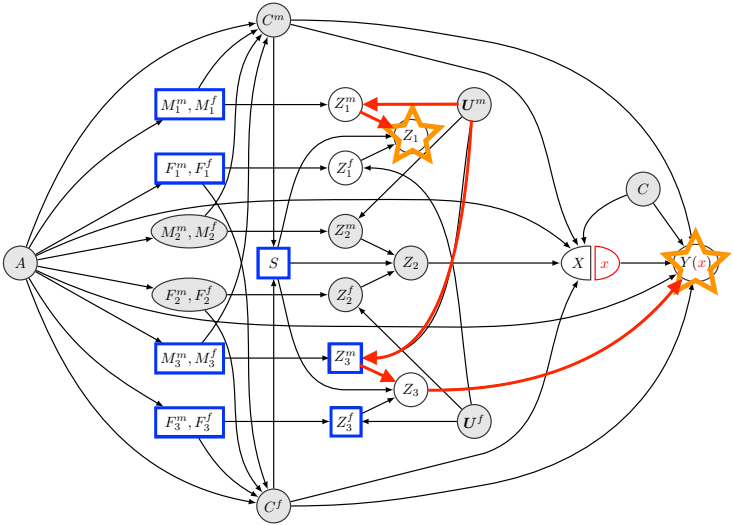
$$Z_1 \perp\!\!\!\perp Y(x) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{F}_1^{mf}, \boldsymbol{V}_3, S = 1), \tag{2}$$

where $\boldsymbol{V}_3^m = (\boldsymbol{M}_3^{mf}, Z_3^m)$ and $\boldsymbol{V}_3 = (\boldsymbol{M}_3^{mf}, \boldsymbol{F}_3^{mf}, Z_3)$.

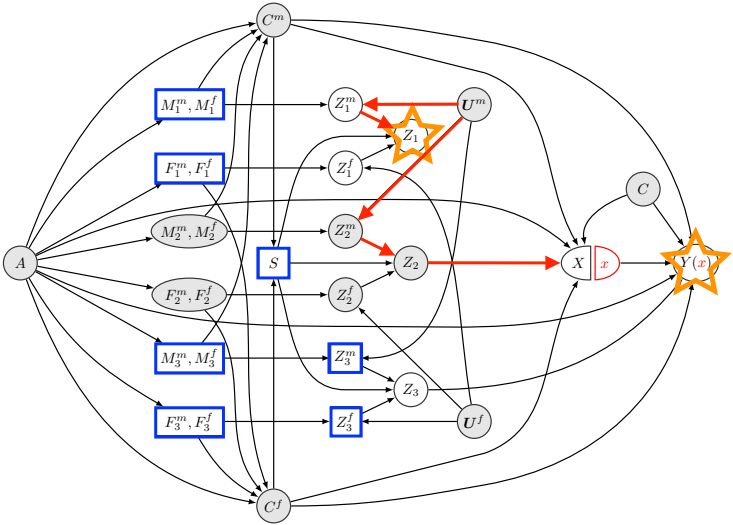
Illustration



Illustration



Illustration



General results

- $\mathcal{J}_x$ includes all causal variants for X with no direct effect on Y;
- $\mathcal{J}_y$ includes all variants with a direct causal effect on Y (not mediated by X).

Theorem

Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ *is a full chromosome. Then* Z_j^m *is a valid instrument conditional on* $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m = (\boldsymbol{M}_{\mathcal{B}}^{mf}, \boldsymbol{Z}_{\mathcal{B}}^m))$ *for some* $\mathcal{B} \subset \{1, \ldots, p\}$ *if the following conditions are satisfied:*

1. $Z_j^m \not\perp\!\!\!\perp \boldsymbol{Z}_{\mathcal{J}_x}^m \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m, S = 1)$;
2. $Z_j^m \perp\!\!\!\perp \boldsymbol{Z}_{\mathcal{J}_y}^m \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m, S = 1)$.

Intrinsic tradeoff: choosing a smaller $\mathcal{B}$ makes

- condition 1 more likely (increased power);
- condition 2 less likely (decreased validity).

Simplification via Markovian structure

Haldane (1919)'s meiosis model

- Given mother's haplotypes $\boldsymbol{M}_j^{mf}$, selection indicator U_j^m, and conception $S = 1$,

$$Z_j^m = \begin{cases} M_j^{(U_j^m)}, & \text{with probability } 1-\epsilon, \\ 1 - M_j^{(U_j^m)}, & \text{with probability } \epsilon. \end{cases}$$

- $\boldsymbol{U}^m$ is a Markov process (basically a Poisson process).

Theorem

Let b_1 and b_2 ($b_1 < j < b_2$) be two heterozygous loci in the mother's genome, i.e., $M_{b_1}^f \neq M_{b_1}^m$ and $M_{b_2}^f \neq M_{b_2}^m$. Under Haldane's model with $\epsilon = 0$, Z_j^m is a valid IV given $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\{b_1,b_2\}}^m)$ if

$$\{b_1+1, \ldots, b_2-1\} \cap \mathcal{J}_x \neq \emptyset \quad \textit{and} \quad \{b_1+1, \ldots, b_2-1\} \cap \mathcal{J}_y = \emptyset.$$

Almost exact inference

- Consider any sharp null hypothesis $H_0 : Y(x) = Y(0) + \beta x$ for some β.
- Basic idea: $Z_j^m \perp\!\!\!\perp (Y - X\beta) \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m)$ under H_0.
- Randomization test p-value can be obtained from the known distribution $Z_j^m \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m)$ from a meiosis model.[7]
- This test is "almost exact".

See our paper for

- a test statistic based on using inverse probability weight as a "clever covariate";
- how to combine "evidence factors" from multiple instruments.

[7] This extends the unconditional randomiation test using an instrumental variable in Hyunseung Kang, Laura Peck, and Luke Keele (2018). "Inference for Instrumental Variables: A Randomization Inference Approach". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4, pp. 1231–1254. ISSN: 1467-985X. DOI: 10.1111/rssa.12353; Paul R. Rosenbaum (1996). "Identification of Causal Effects Using Instrumental Variables: Comment". In: *Journal of the American Statistical Association* 91.434, pp. 465–468. ISSN: 0162-1459. DOI: 10.2307/2291633. JSTOR: 2291633.

Dataset

- 6,222 mother-child duos from Avon Longitudinal Study of Parents and Children (ALSPAC).
- Negative control example: effect of child's BMI at age 7 on mother's pre-pregnancy BMI (spurious correlation due to dynastic effect).
- Positive control example: effect of child's BMI at age 7 on a simulated, noisy version of itself.
- 11 candidate instruments selected from a GWAS study for childhood BMI.
- Condition on all variants outside of a 500 kilobase window around each instrument.

Negative control example

Instrument (rsID)	Chromosome	Proximal gene	P-value
rs11676272	2	*ADCY3*	0.45
rs7138803	12	*BCDIN3D*	0.55
rs939584	2	*TMEM18*	0.39
rs17817449	16	*FTO*	0.06
rs12042908	1	*TNNI3K*	0.35
rs543874	1	*SEC16B*	0.07
rs56133711	11	*BDNF*	0.59
rs571312, rs76227980	18	*MC4R*	0.48
rs12641981	4	*GNPDA2*	0.62
rs1094647	1	*SLC45A3*	0.19
Fisher's combination test			0.21
Two-stage least squares			0.02

Positive control example

Instrument (rsID)	Chromosome	Gene	P-value for noise R^2 of 10%	20%	50%
rs11676272	2	*ADCY3*	0.01	0.01	0.01
rs7138803	12	*BCDIN3D*	0.01	0.01	0.01
rs939584	2	*TMEM18*	0.98	0.95	0.88
rs17817449	16	*FTO*	0.33	0.35	0.44
rs12042908	1	*TNNI3K*	0.77	0.79	0.85
rs543874	1	*SEC16B*	0.48	0.64	0.92
rs56133711	11	*BDNF*	0.12	0.14	0.25
rs571312, rs76227980	18	*MC4R*	0.31	0.39	0.63
rs12641981	4	*GNPDA2*	0.49	0.56	0.76
rs1094647	1	*SLC45A3*	0.23	0.25	0.35
Fisher's combination test			0.03	0.05	0.16
Two-stage least squares			$< 10^{-20}$		4.5×10^{-11}

Summary

- Main ideas are contained in the title: **Almost Exact Mendelian Randomization**.

Advantages

- Many conceptual advantages.
- Robustness to misspecified phenotype models and weak instruments.
- Elimination of bias arising from population structure, assortative mating, dynastic effects, and horizontal pleiotropy.

Limitations

- Relatively low power;
- Possibly incorrect model for meiosis (e.g. due to transmission ratio distortion).