THE UNIVERSITY OF TEXAS
MD Anderson ~~Cancer~~ Center
UTHealth Houston
Graduate School of Biomedical Sciences

Novel statistical methods for mediation analysis with high-dimensional omics mediators

Seminar @ AMGEN

Zhichao (Zachary) Xu
Nov 15, 2024

Agenda

01 Background

02 Projects

03 Summary

04 Q & A

Background

Education

UT M.D. Anderson Cancer Center
Ph.D. candidate in Biostatistics (2021 - 2025)

Yale University
M.S. in Biostatistics (2019 - 2021)

University of International Business and Economics
B.S. in Statistics (2015 - 2019)

Experience

Yale University
Research Assistant (2019 - 2021)

UT M.D. Anderson Cancer Center
Research Assistant (2021 - present)

Merck
Summer Intern (2024 Summer)

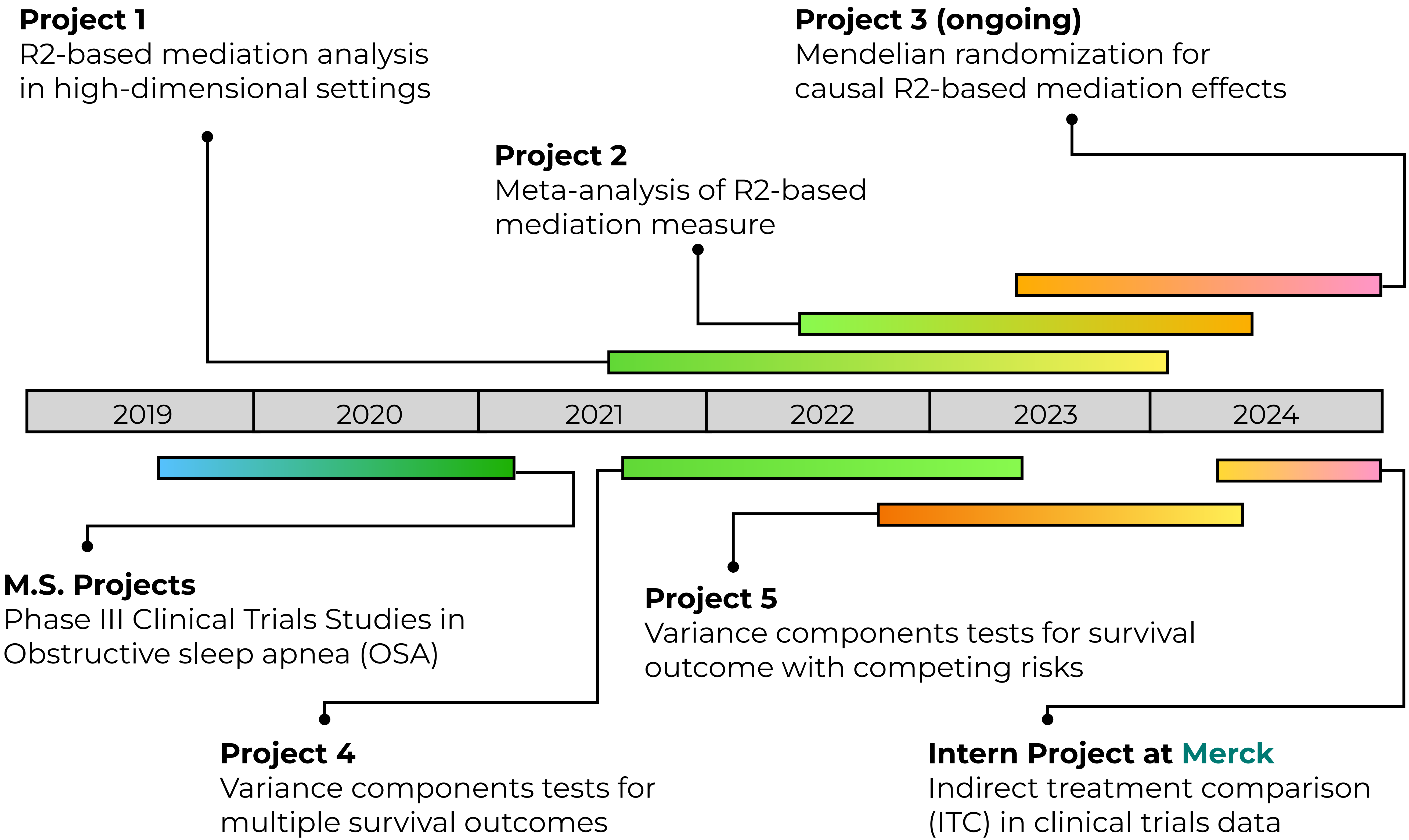
Methodology Publications

1. **Xu, Z.**, Li, C., Chi, S., Yang, T., & Wei, P. (2024). Speeding up interval estimation for R 2-based mediation effect of high-dimensional mediators via cross-fitting. *Biostatistics*, kxae037.
2. **Xu, Z.**, & Wei, P. (2024). A novel statistical framework for meta-analysis of total mediation effect with high-dimensional omics mediators in large-scale genomic consortia. *PLOS Genetics*.
3. **Xu, Z.**, Choi, J., & Sun, R. (2024). Set-Based Tests for Genetic Association Studies with Interval-Censored Competing Risks Outcomes. *Statistics in Biosciences,* 1-18.
4. Choi, J., **Xu, Z.**, & Sun, R. (2024). Variance-components tests for genetic association with multiple interval-censored outcomes. *Statistics in Medicine*, 43(13), 2560-2574.
5. Li, H., Zhu, B., **Xu, Z.**, Adams, T., Kaminski, N., & Zhao, H. (2021). A Markov random field model for network-based differential expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 22, 1-16.

Contributed Talks

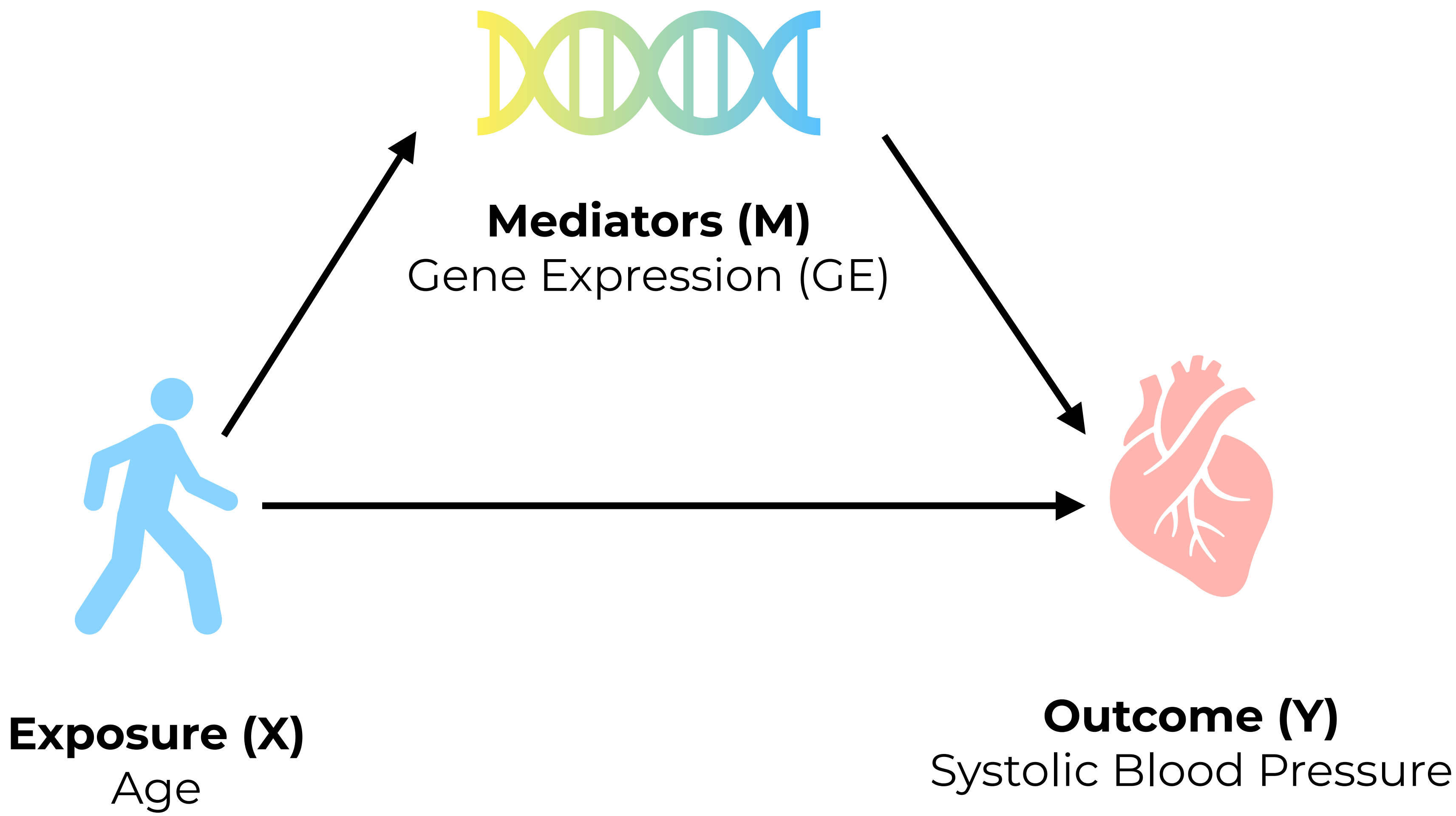
1. American Society of Human Genetics (ASHG) Annual Meeting. Washington, DC. November 2023.
2. Eastern North American Region (ENAR) Spring Meeting. Baltimore, Maryland. March 2024.
3. Joint Statistical Meetings (JSM). Portland, Oregon. August 2024.

Projects Overview



Introduction

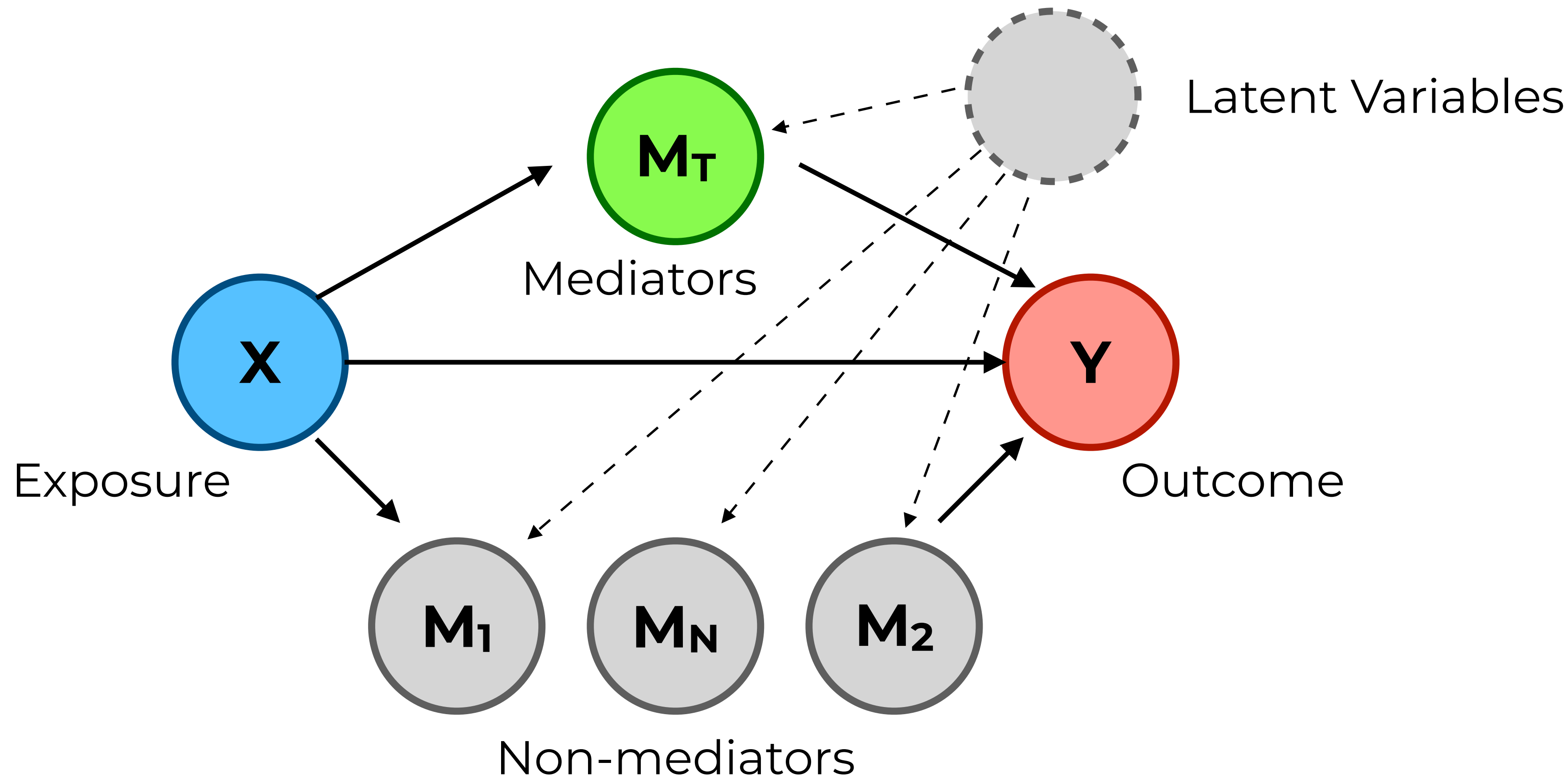
Mediation analysis



- **How** does age affect systolic blood pressure through gene expression?
- **How important** are mediators in this pathway?
- **How to measure** this importance?

Introduction

Mediators

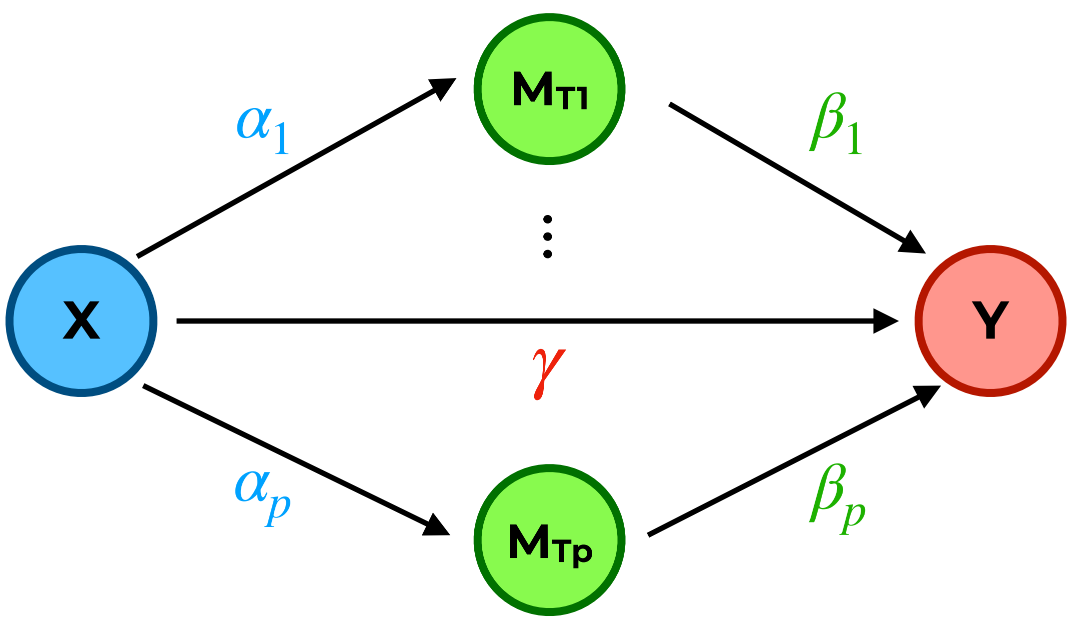


- M_T True mediators: associated with X and Y
- M_1 Type 1 non-mediators: only associated with X
- M_2 Type 2 non-mediators: only associated with Y
- M_N Noise: not associated with either X or Y
- Latent variables introduce correlations

1. Rm, B. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*, *51*, 1173-1182.

Introduction

Total mediation effect measure



$$M_j = \alpha_j X + \xi$$

$$Y = \gamma X + \sum \beta_j M_j + \epsilon$$

1. Product measure

$$\sum_{j=1}^{p} \alpha_j \beta_j = \alpha_1\beta_1 + \ldots + \alpha_p\beta_p$$

2. Ratio measure

$$\sum_{j=1}^{p} \alpha_j \beta_j / \gamma = (\alpha_1\beta_1 + \ldots + \alpha_p\beta_p)/\gamma$$

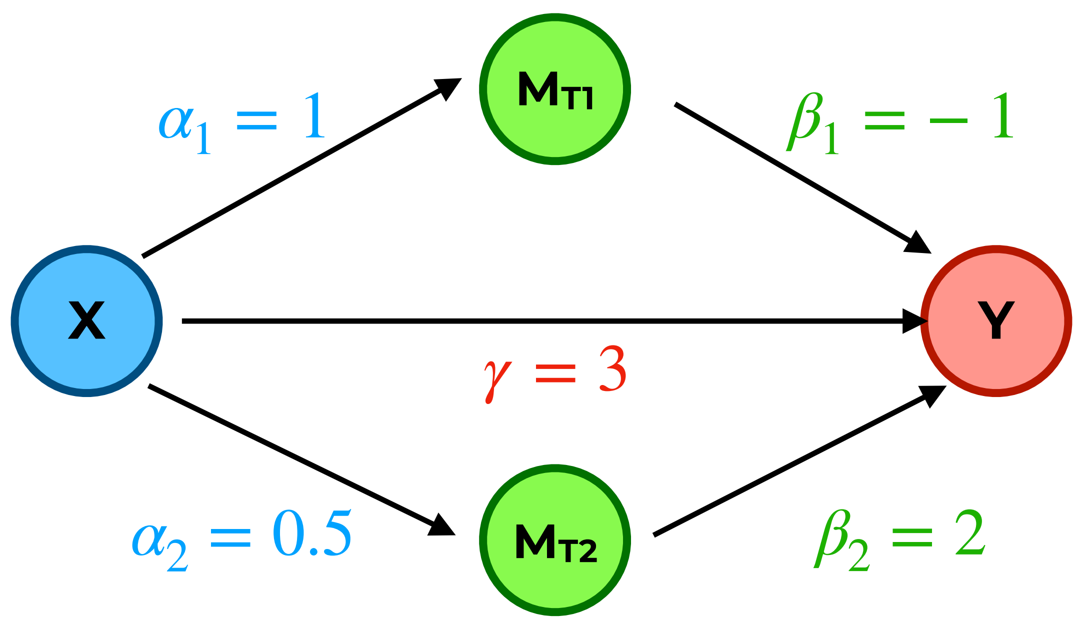
3. Proportion measure

$$\sum_{j=1}^{p} \alpha_j \beta_j / \left(\sum_{j=1}^{p} \alpha_j \beta_j + \gamma\right) =$$
$$(\alpha_1\beta_1 + \ldots + \alpha_p\beta_p)/(\alpha_1\beta_1 + \ldots + \alpha_p\beta_p + \gamma)$$

1. MacKinnon, D. (2012). Introduction to statistical mediation analysis. Routledge.

Methods

Total mediation effect measure



1. Product measure

$\sum_{j=1}^{p} \alpha_j \beta_j = \alpha_1\beta_1 + \alpha_2\beta_2$

= 1 x (-1) + 0.5 x 2

= 0

2. Ratio measure

$\sum_{j=1}^{p} \alpha_j \beta_j / \gamma = (\alpha_1\beta_1 + \alpha_2\beta_2)/\gamma$

= (1 x (-1) + 0.5 x 2) / 3

= 0

3. Proportion measure

$\sum_{j=1}^{p} \alpha_j \beta_j / (\sum_{j=1}^{p} \alpha_j \beta_j + \gamma)$

$= (\alpha_1\beta_1 + \alpha_2\beta_2)/(\alpha_1\beta_1 + \alpha_2\beta_2 + \gamma)$

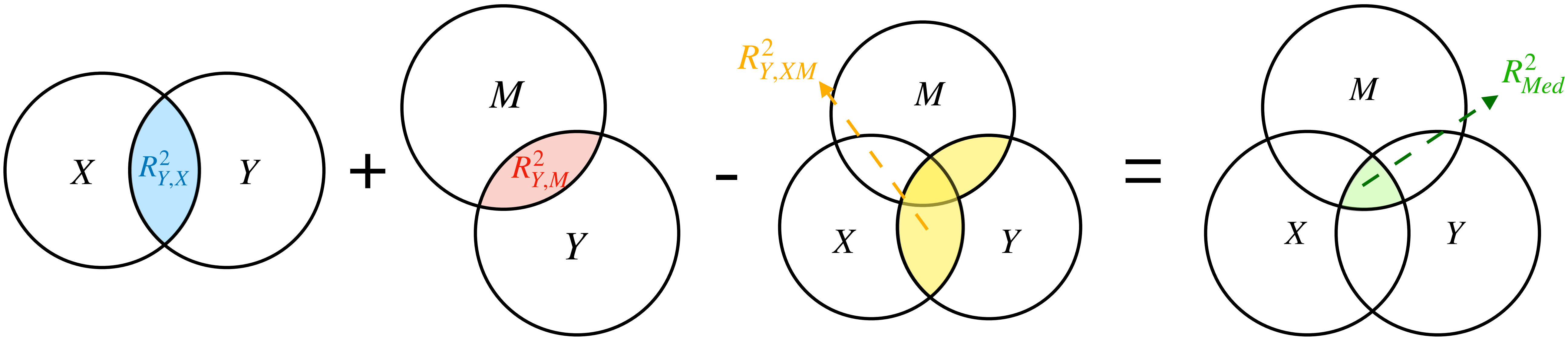
= (1 x (-1) + 0.5 x 2) / (1 x (-1) + 0.5 x 2 + 3)

= 0 / 3

=0

Methods

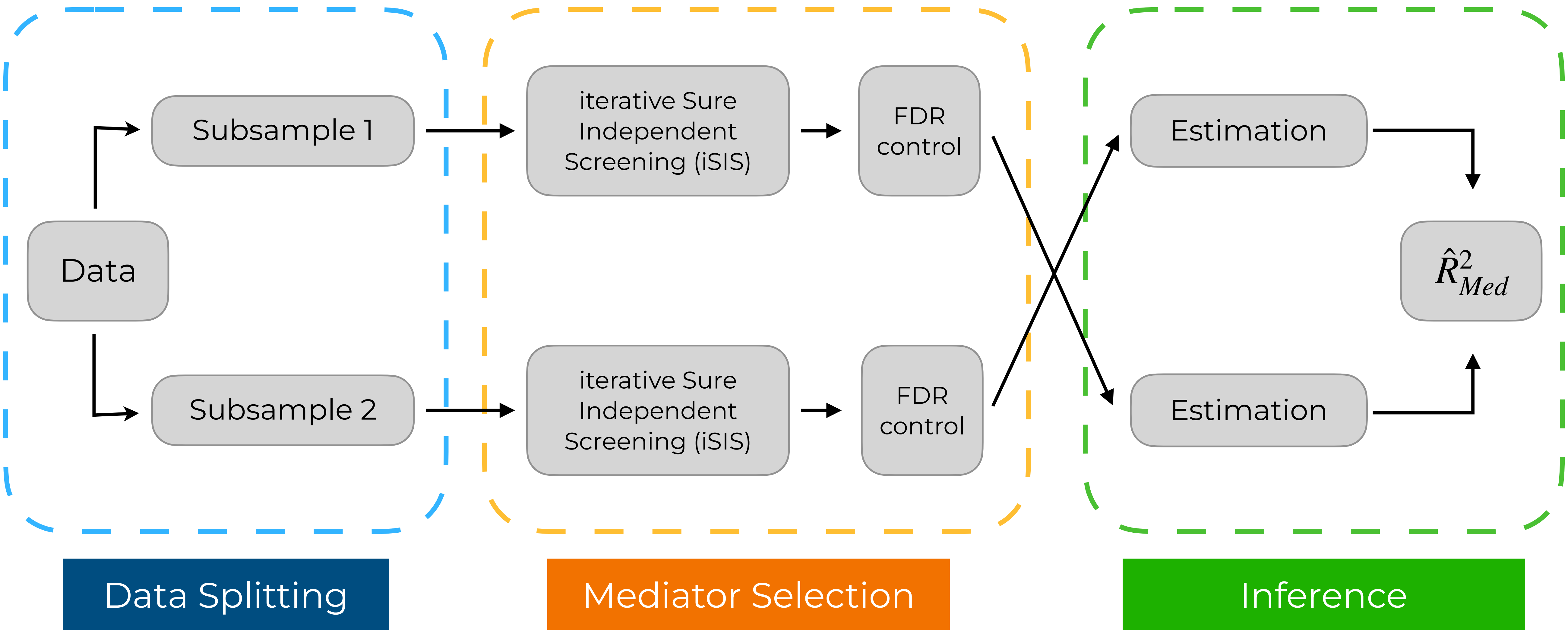
R2-based measure



- $R^2_{Med} = R^2_{Y,X} + R^2_{Y,M} - R^2_{Y,XM}$
- Variance of Y explained by X through M

Methods

Cross-fitted R2 measure (CFR2M)



1. Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(5), 849-911.
2. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Results

Simulations

Settings of selected scenarios (A1)-(A3)

	# of M_T	# of M_1	# of M_2	# of M_N
A1	15	0	0	1485
A2	150	1350	0	0
A3	150	0	1350	0

- Coverage Probability
- Bias
- Selection Accuracy
- Empirical Standard Error
- Computational Time

Results of selected scenarios (A1)-(A3)

	Sample Size	Coverage (%)	Bias ($\times 10^{-2}$)	True Positive (%)	False Positive (%)
A1	750	92.0	0.739	94.5	2.1
	1500	93.5	0.658	92.9	1.8
	3000	93.5	0.133	96.7	0.8
A2	750	93.5	0.269	31.0	1.1
	1500	95.0	0.198	50.5	2.6
	3000	95.0	0.168	76.2	6.5
A3	750	96.0	0.029	13.0	2.5
	1500	95.0	-0.255	38.6	2.2
	3000	97.0	0.113	72.4	0.1

Results

Applications

- Data: Framingham Heart Study (FHS)
  - Exposure: age
  - Outcome: systolic blood pressure
  - Mediators: gene expression (d = 17,873)
  - Sample Size: N = 4,542

Mediation effect sizes using CFR2M, B-Mixed and HDMT with FHS data

	R2M	R2-YX	Selected genes	CPU Time (hrs)
CFR2M	0.126 [0.109, 0.144]	0.201	166 / 194	4.67
B-Mixed	0.120 [0.081, 0.147]	0.200	200	1899.75
HDMT	0.042 [0.034, 0.051]	0.201	7 / 11	1367.50

1. https://www.framinghamheartstudy.org/
2. Yang, T., Niu, J., Chen, H., & Wei, P. (2021). Estimation of total mediation effect for high-dimensional omics mediators. *BMC bioinformatics*, *22*, 1-17.
3. Dai, J. Y., Stanford, J. L., & LeBlanc, M. (2022). A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, *117*(537), 198-213.

Summary

- **Method**
  - Propose a novel R2-based mediation measure
  - Derive the asymptotic distribution
  - Relax the assumption of oracle property (i.e. asymptotically exact variable selection)
- **Application**
  - Implement the cross-fitting and sample-splitting estimation procedure
  - Achieve speed improvement of over 400 times compared to resampling-based methods
- **Output**
  - **Xu, Z.**, Li, C., Chi, S., Yang, T., & Wei, P. (2024). Speeding up interval estimation for R2-based mediation effect of high-dimensional mediators via cross-fitting. *Biostatistics*, kxae037.
  - R package CFR2M at GitHub

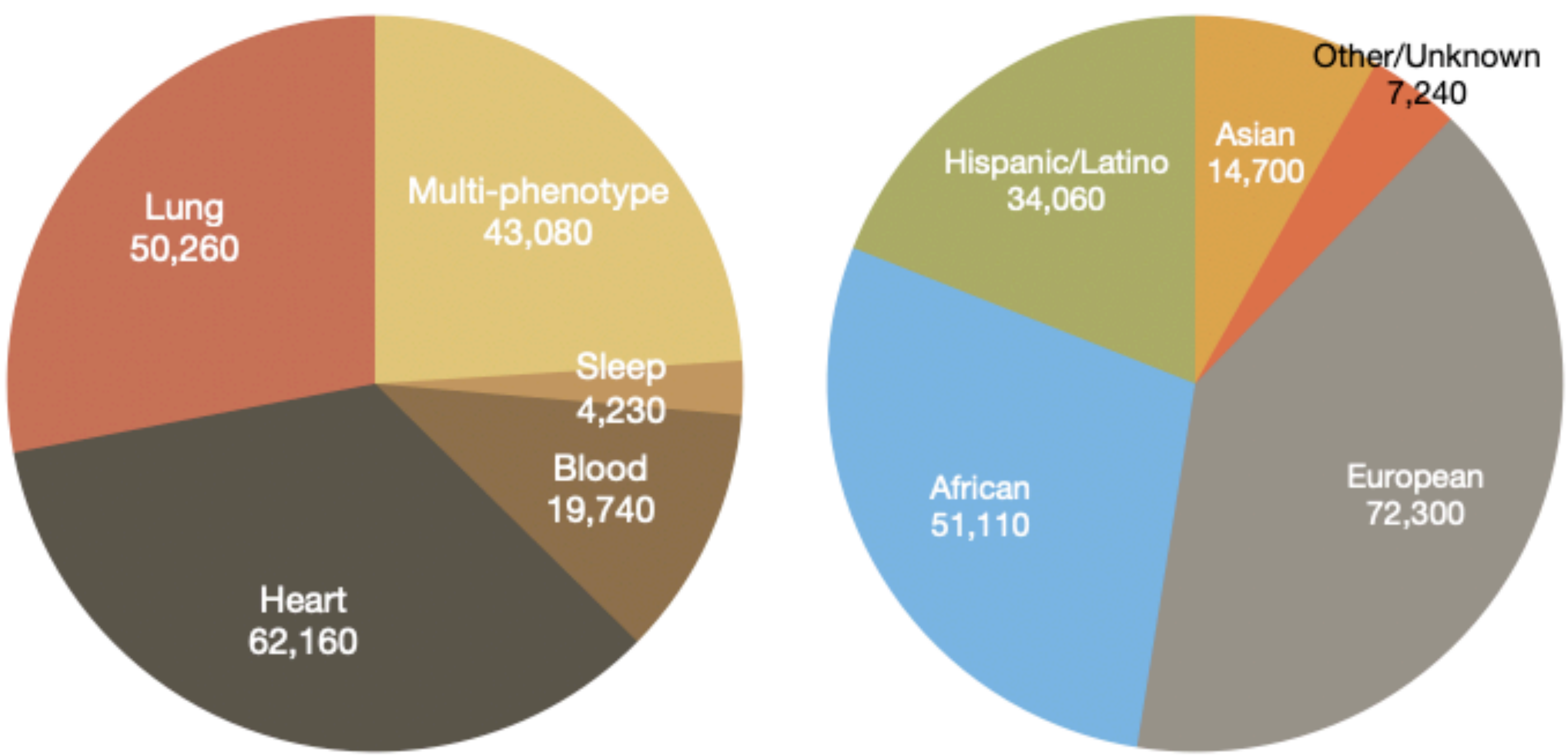
Introduction

TOPMed program

Trans-Omics for Precision Medicine (TOPMed):

- ~180k participants from >85 different studies
- Multi-omics: RNA-seq, DNA methylation, etc.
- Multiple ethnic groups

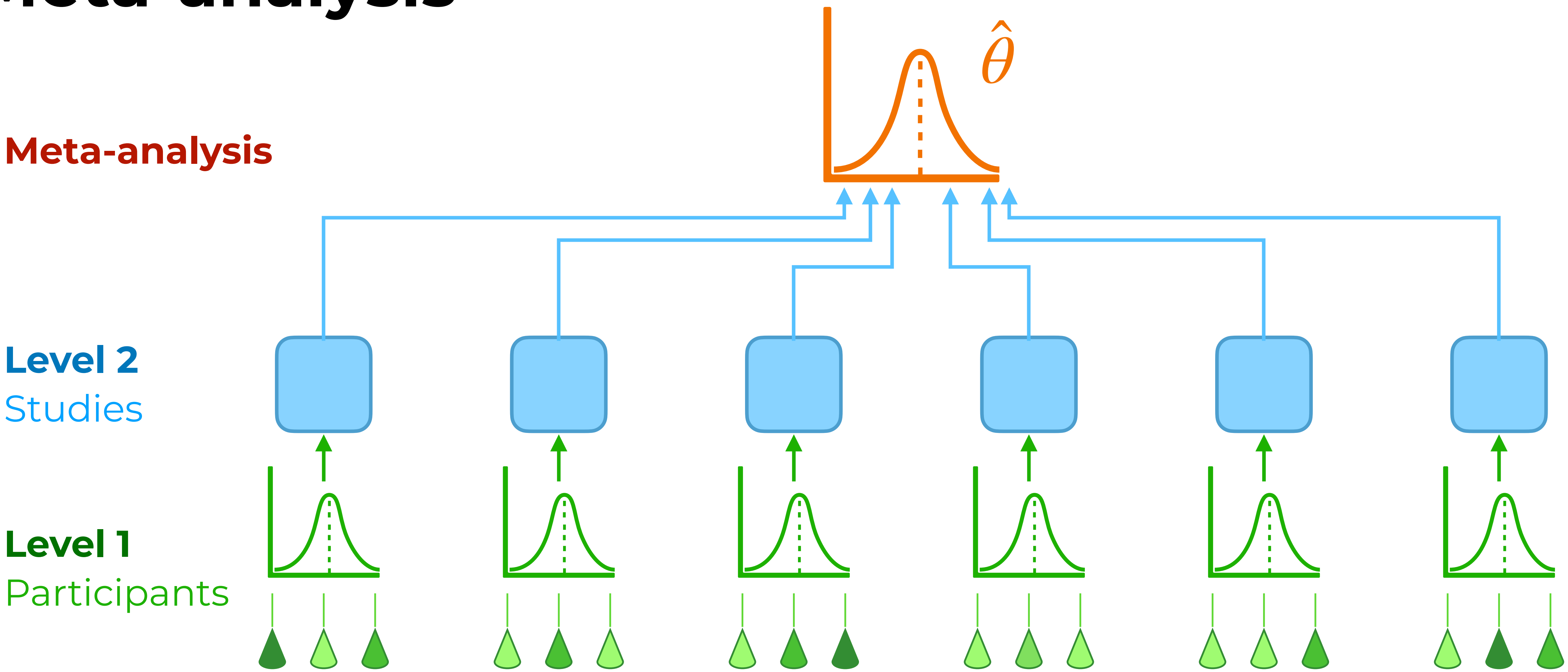
Phenotype focus Phases 1-7 (left) and participant diversity (right) in TOPMed program



1. https://topmed.nhlbi.nih.gov/

Introduction

Meta-analysis

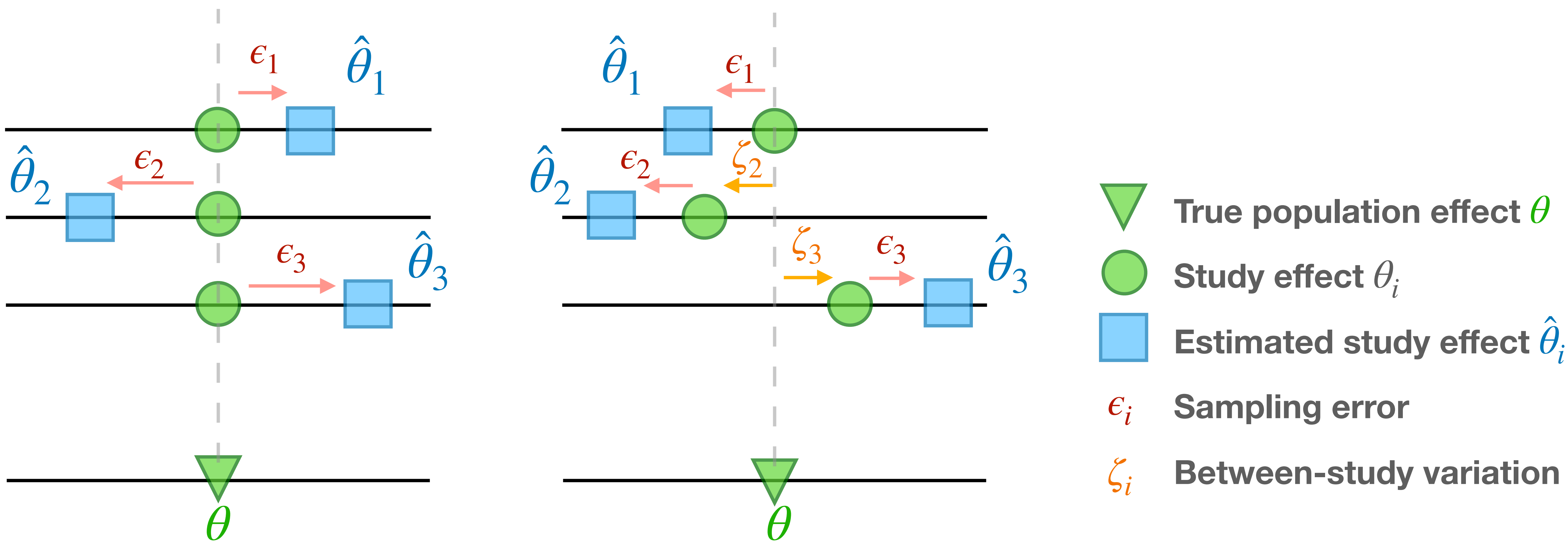


- Analysis of analyses
- Heterogeneity
- Different weights

1. Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Introduction

Fixed / Random-effects models



Fixed-effects model

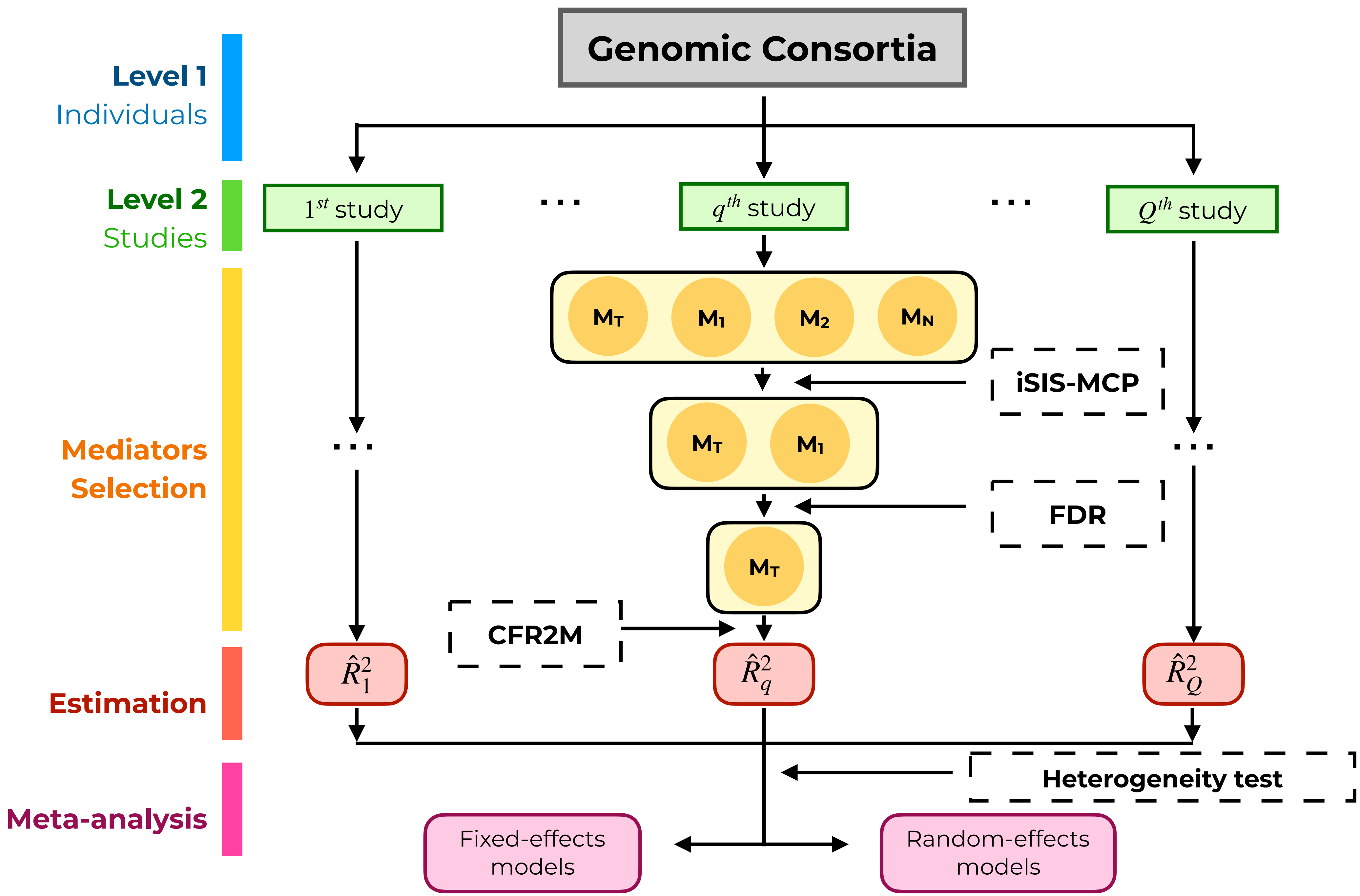
1. Assume common effect size
2. Precise estimates
3. Underestimate variability

Random-effects model

1. Assume effect sizes vary
2. Conservative estimates
3. Reflect additional uncertainty

Methods

Meta-analysis of R2-based measure



Methods

Meta-analysis estimators

Fixed-effects Inverse-variance estimator: $\hat{R}^2_{IW} = \frac{\sum_Q \hat{R}^2_q / Var(\hat{R}^2_q)}{\sum_Q 1/Var(\hat{R}^2_q)}$, $Var(\hat{R}^2_{IW}) = \frac{1}{\sum_Q 1/Var(\hat{R}^2_q)}$.

Random-effects model denotes that

$\hat{R}^2_{RE} = (\sum_Q \frac{1}{S_q + \hat{\tau}^2})^{-1} \times \sum_Q (\frac{\hat{R}^2_q}{S_q + \hat{\tau}^2})$, where S_q is the estimated variance of $\hat{R}^2_q$ and $\hat{\tau}^2$ is an estimate of the between-study variance.

DerSimonian and Laird (DL) estimator: $\hat{\tau}^2_{DL} = \max \left\{ \frac{\sum_{q=1}^{Q} S_q^{-1}(\hat{R}^2_q - \hat{R}^2_{IW}) - (Q-1)}{\sum_{q=1}^{Q} S_q^{-1} - \sum_{q=1}^{Q} S_q^{-2} / \sum_{q=1}^{Q} S_q^{-1}}, 0 \right\}$.

Median-unbiased Paule-Mandel (MPM) estimator: $\hat{\tau}^2_{MPM}$ is given by the value of τ^2 such that

$T_{Gen} = \sum_Q \frac{1}{S_q + \hat{\tau}^2} (\hat{R}^2_q - \hat{R}^2_{IW})^2 = \chi^2_{Q-1,0.5}$ (median of a chi-square distribution with Q-1 degrees of freedom).

Results

Simulations: fixed-effects models

Settings of selected scenarios (B1)-(B2)

	# of M_T	# of M_1	# of M_2	# of M_N
B1	5	0	0	1495
B2	150	150	150	1050

- Coverage Probability
- Bias
- Asymptotic Standard Error (SE)
- Empirical Standard deviation (SD)

Results of fixed-effects model in selected scenarios (B1) and (B2)

	Sample Sizes	Coverage (%)	Bias ($\times 10^{-2}$)	SE ($\times 10^{-2}$)	SD ($\times 10^{-2}$)
B1	3000	93.5	-0.167	1.213	1.237
	1000 / 2000	94.0	-0.098	1.210	1.235
	750 / 750 / 1500	93.5	-0.029	1.207	1.226
	750 / 750 / 750 / 750	94.0	0.037	1.204	1.212
	600 / 600 / 600 / 600 / 600	93.5	0.098	1.201	1.233
B2	3000	93.5	-0.009	1.403	1.476
	1000 / 2000	93.5	0.019	1.402	1.480
	750 / 750 / 1500	93.0	0.046	1.402	1.498
	750 / 750 / 750 / 750	92.0	0.059	1.401	1.511
	600 / 600 / 600 / 600 / 600	93.0	0.115	1.400	1.510

Results

Simulations: random-effects models

Settings of selected scenarios (B1)-(B2)

	# of M_T	# of M_1	# of M_2	# of M_N
B1	5	0	0	1495
B2	150	150	150	1050

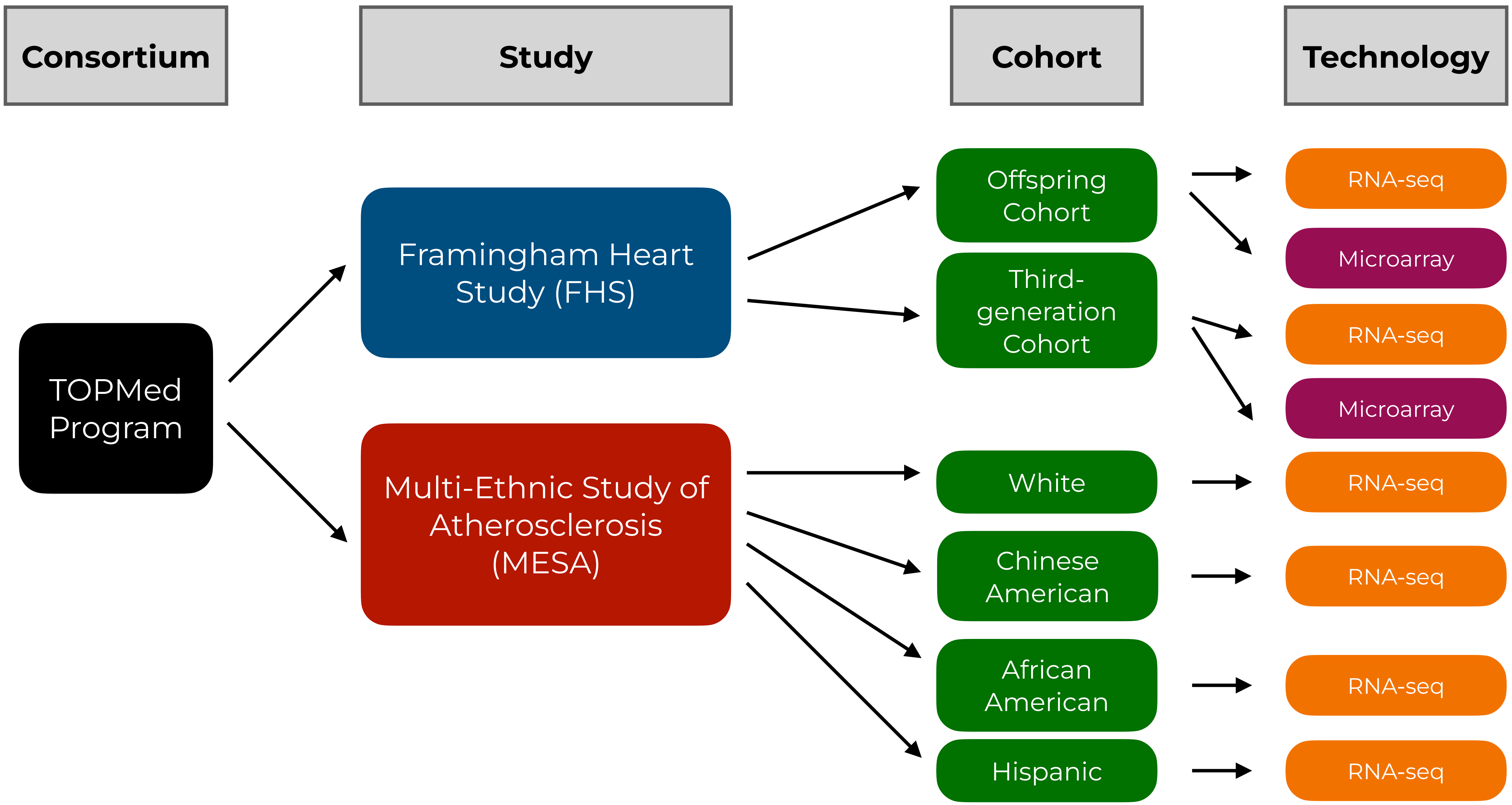
- Coverage Probability
- Bias
- Two estimators (DL vs. MPM)

Results of fixed-effects model in selected scenarios (B1) and (B2)

			DerSimonian and Laird (DL)		Median-unbiased Paule-Mandel (MPM)	
	# of studies	Sample Size	Coverage (%)	Bias (x10^{-2})	Coverage (%)	Bias (x10^{-2})
B1	5	2400	84.5	-3.406	89.0	3.421
	8	1500	87.5	-3.389	89.0	-3.408
	10	1200	90.5	-3.567	89.0	-3.588
	16	750	92.5	-3.923	92.0	-4.919
	20	600	91.0	-4.132	88.5	-4.154
B2	5	2400	85.0	-1.781	90.5	-1.771
	8	1500	88.5	-2.082	92.5	-2.076
	10	1200	91.5	-2.103	94.0	-2.097
	16	750	92.5	-1.923	94.0	-1.919
	20	600	95.0	-1.682	92.0	-1.681

Results

Applications



1. Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., ... & Stilp, A. M. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290-299.
2. Olson, J. L., Bild, D. E., Kronmal, R. A., & Burke, G. L. (2016). Legacy of MESA. *Global heart*, *11*(3), 269-274.

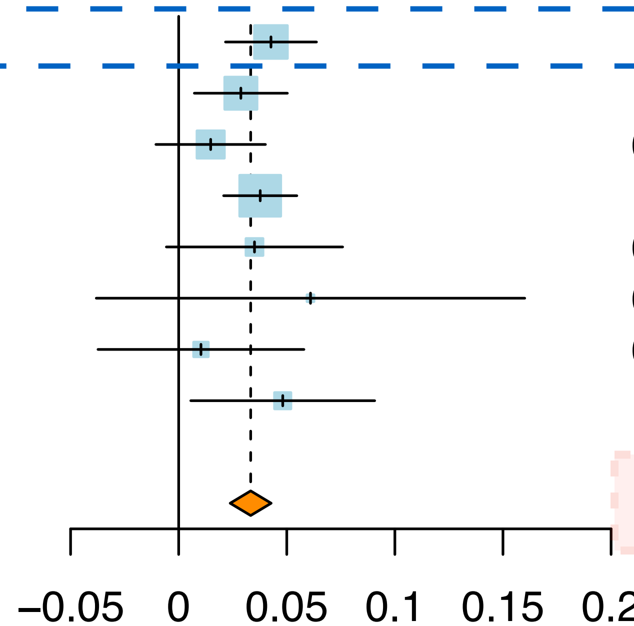
Results

Applications

- **Outcome**: systolic blood pressure
- **Exposure**: age
- **Mediators**: gene expression

Cohort	N	Tech	SE	# of transcripts	R2 [95% CI]	Weight
FHS Offspring	1299	Microarray	0.011	17873	0.043 [0.022; 0.064]	20.2%
FHS Generation 3	1118	Microarray	0.011	17873	0.029 [0.007; 0.050]	19.4%
FHS Offspring	687	RNA–seq	0.013	43232	0.015 [–0.011; 0.040]	14.0%
FHS Generation 3	1635	RNA–seq	0.009	47396	0.038 [0.021; 0.055]	31.3%
MESA White	415	RNA–seq	0.021	34129	0.035 [–0.006; 0.076]	5.4%
MESA CA	104	RNA–seq	0.051	30336	0.061 [–0.038; 0.160]	0.9%
MESA AA	321	RNA–seq	0.024	33888	0.010 [–0.037; 0.058]	4.0%
MESA Hispanic	285	RNA–seq	0.022	33347	0.048 [0.006; 0.091]	5.0%
Fixed–effects model					**0.033 [0.024; 0.043]**	**100.0%**

Heterogeneity: $I^2 = 0\%$, $p = 0.67$

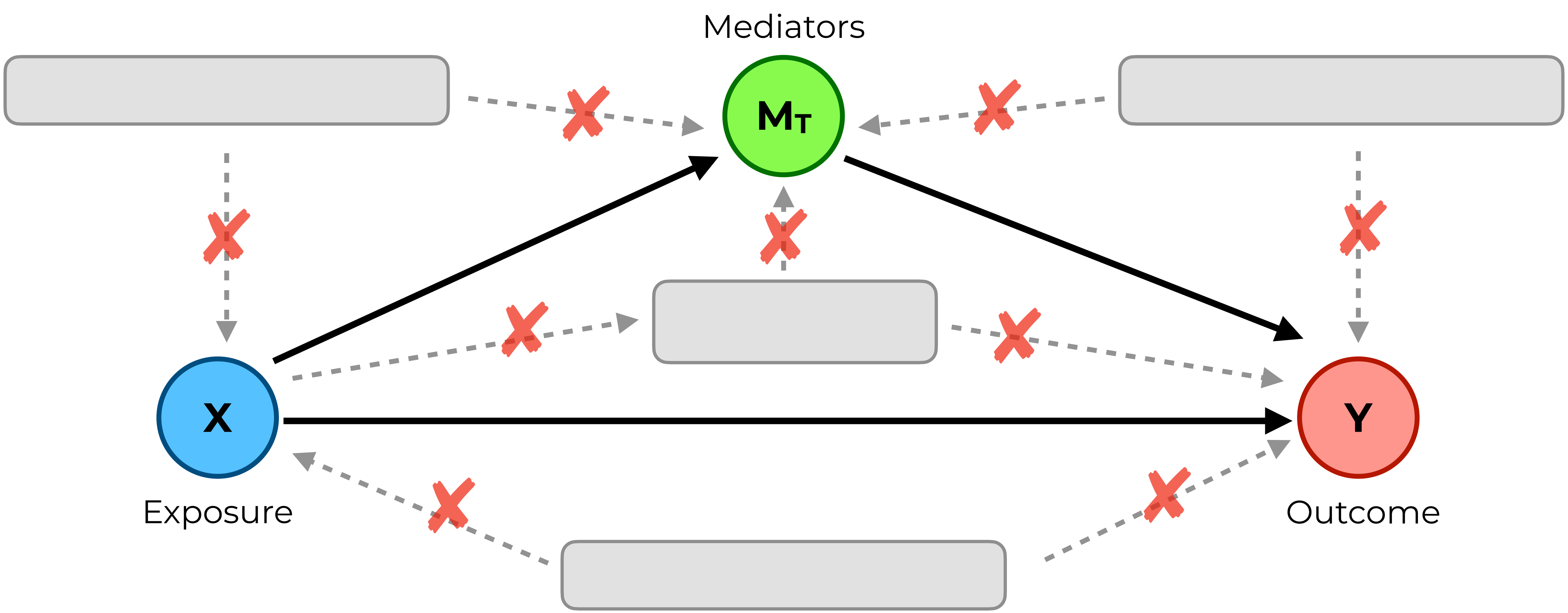


Summary

- **Method**
  - Propose a novel meta-analysis framework for R2-based mediation measure
  - Require only summary statistics and allow between-study heterogeneity
- **Application**
  - Adjust for different covariates in separate studies
  - Verify the meta-analysis with a minimum sample size of around 300
- **Output**
  - **Xu, Z.** & Wei, P. (2024). A novel statistical framework for meta-analysis of total mediation effect with high-dimensional omics mediators in large-scale genomic consortia. *PLOS Genetics*.
  - R package MetaR2M at GitHub
  - Three contributed talks at conferences (ENAR 2024, JSM 2024 & ASHG 2023)

Introduction

Mediation assumptions

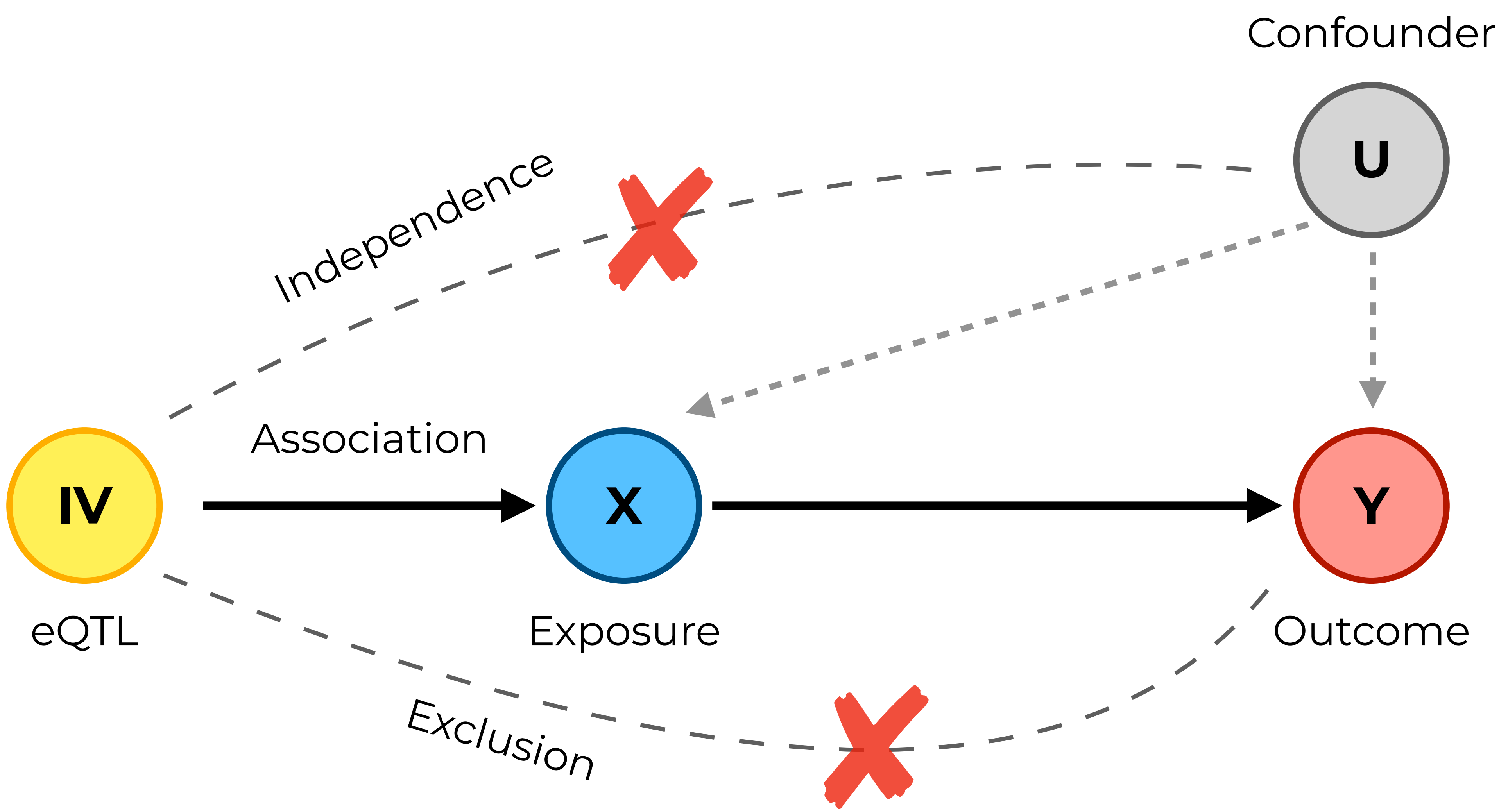


- No unmeasured confounding between X, M, and Y
- No exposure-caused confounders of the mediator and outcome
- No exposure-mediator interaction

1. VanderWeele, T. J. (2016). Mediation analysis: a practitioner's guide. *Annual review of public health*, *37*(1), 17-32.
2. MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol.*, *58*(1), 593-614.

Introduction

Mendelian Randomization (MR)

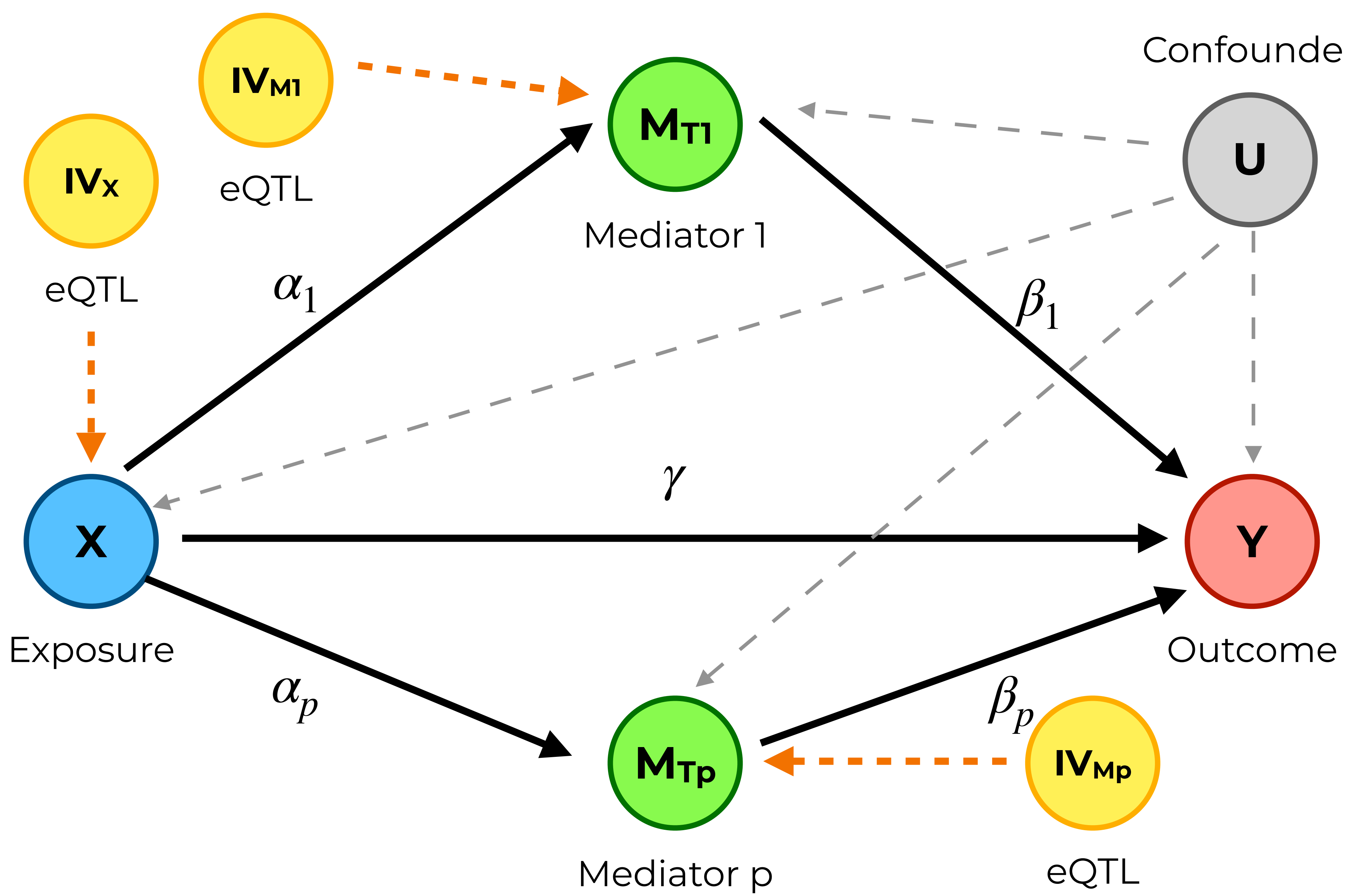


- **Relevance**: the genetic variant must be associated with the exposure
- **Independence**: the variant should be independent of confounders
- **Exclusion restriction**: the variant affects outcome only through exposure

1. Sanderson, E. (2021). Multivariable Mendelian randomization and mediation. *Cold Spring Harbor perspectives in medicine*, *11*(2), a038984.
2. Burgess, S., Thompson, D. J., Rees, J. M., Day, F. R., Perry, J. R., & Ong, K. K. (2017). Dissecting causal pathways using Mendelian randomization with summarized genetic data: application to age at menarche and risk of breast cancer. *Genetics*, *207*(2), 481-487.

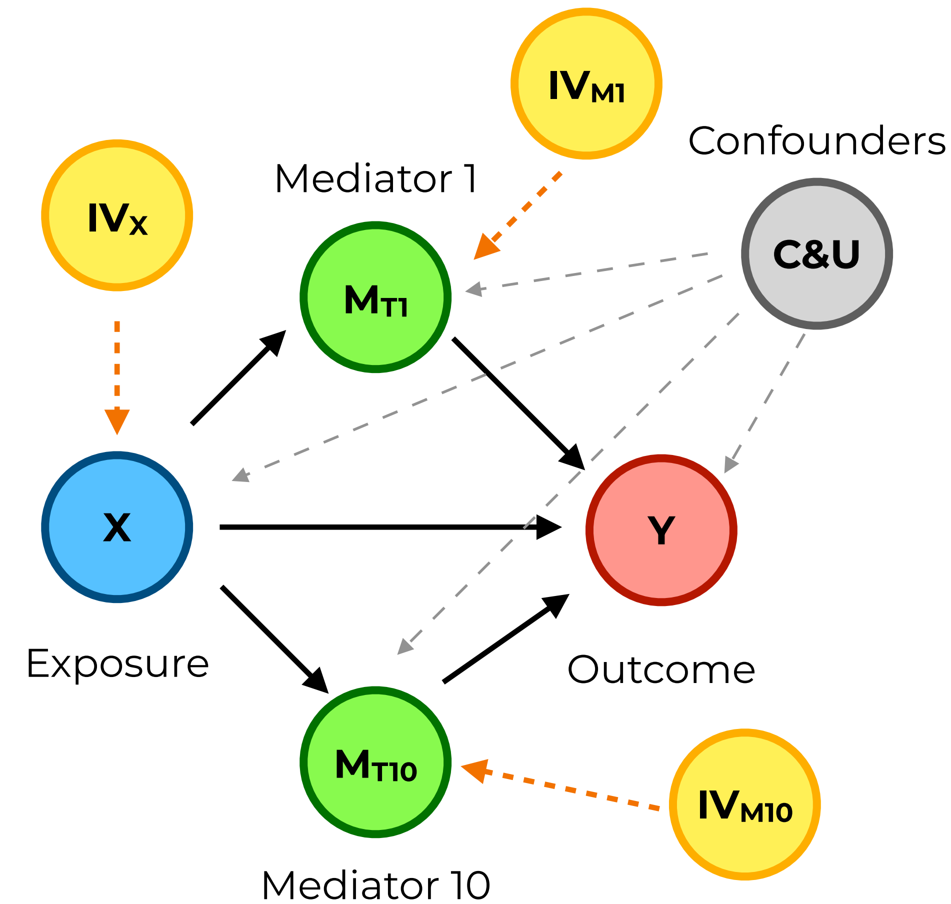
Methods

MR in R2-based mediation analysis



Results

Simulations



- **Model 1:** linear regression
- **Model 2:** linear regression adjusting measured confounder
- **Model 3:** IV regression adjusting measured confounder

Bias for R2-based mediation effects using three models

	R2-YX	R2-YMX	R2-YM	R2-Med
Model 1: Y ~ X + M	-0.0112	-0.4303	-0.4488	-0.0296
Model 2: Y ~ X + M + C	-0.0140	-0.4442	-0.4601	-0.0299
Model 3: Y ~ X + M + C | IVx + IVm	-0.0016	0.0071	0.0060	-0.0027

1. Fox, J., Kleiber, C., & Zeileis, A. (2021). Ivreg: instrumental-variables regression by '2SLS','2SM', or '2SMM', with diagnostics. *R package version 0.6-0.*

Summary

- **Method**
  - Propose a novel R2-based mediation measure to infer the causal mechanisms
  - Integrate multivariable MR into the estimation procedure
- **Application**
  - Adjust for measured and unmeasured confounders
  - Mitigate the reverse causation and measurement error
- **Output**
  - **Xu, Z.** & Wei, P. (2024). Inferring causal R2-based mediation effect with high-dimensional omics mediators vis Mendelian randomization. *Manuscript in preparation.*

Summary

R2-based mediation analysis

1. **Propose the R2-based mediation effects with its asymptotic distribution**
   - Xu, Z., Li, C., Chi, S., Yang, T., & Wei, P. (2024). Speeding up interval estimation for R2-based mediation effect of high-dimensional mediators via cross-fitting. *Biostatistics*, kxae037.
2. **Propose a novel meta-analysis framework with efficient application**
   - Xu, Z. & Wei, P. (2024). A novel statistical framework for meta-analysis of total mediation effect with high-dimensional omics mediators in large-scale genomic consortia. *PLOS Genetics*.
3. **Propose the methods to infer the causal mechanisms**
   - Xu, Z. & Wei, P. (2024). Inferring causal R2-based mediation effect with high-dimensional omics mediators vis Mendelian randomization. *Manuscript in preparation*.

- **Key references:**

  - Yang, T., Niu, J., Chen, H., & Wei, P. (2021). Estimation of total mediation effect for high-dimensional omics mediators. *BMC bioinformatics*, *22*, 1-17.
  - Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348-1360.
  - Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(5), 849-911.

Summary

Other works & publications

- **Sequence kernel association tests (SKAT) with survival outcomes**
  - Choi, J., **Xu, Z.**, & Sun, R. (2024). Variance-components tests for genetic association with multiple interval-censored outcomes. *Statistics in Medicine*, *43*(13), 2560-2574.
  - **Xu, Z.**, Choi, J., & Sun, R. (2024). Set-Based Tests for Genetic Association Studies with Interval-Censored Competing Risks Outcomes. *Statistics in Biosciences*, 1-18.
- **Indirect treatment comparison (ITC) in clinical trials data**
  - **Xu, Z.**, Mukina L., Mt-Isa S., Baumartner R. (2024). The Impact of Effect Modification on Indirect Treatment Comparisons with Time-to-Event outcomes in Health Technology Assessment. *Manuscript in preparation.*
- **Bayesian differential expression (DE) analysis of single cell RNA-seq data**
  - Li, H., Zhu, B., **Xu, Z.**, Adams, T., Kaminski, N., & Zhao, H. (2021). A Markov random field model for network-based differential expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, *22*, 1-16.
- **Sleep for Stroke Management and Recovery Trial (Sleep SMART) Phase III trials**
  - Adekolu, O., **Xu, Z.**, Chu, J. H., Kushida, C., Yaggi, H., Knauert, M., & Zinchuk, A. (2021). 441 Influence Of Chronotype On CPAP Adherence. *Sleep*, *44*(Supplement_2), A174-A175.
  - Walker, A., Baldassarri, S., Chu, J. H., Deng, A., **Xu, Z.**, ... & Zinchuk, A. (2023). 0480 Psychoactive Substance Use and Sleep Characteristics Among Individuals with Untreated Obstructive Sleep Apnea. *Sleep*, *46*(Supplement_1), A213-A214.
  - Knauert, M. P., Adekolu, O., **Xu, Z.**, Deng, A., ... & Zinchuk, A. (2023). Morning chronotype is associated with improved adherence to continuous positive airway pressure among individuals with obstructive sleep apnea. *Annals of the American Thoracic Society*, *20*(8), 1182-1191.
  - Baldassarri, S. R., Chu, J. H., Deng, A., **Xu, Z.**, Blohowiak, ... & Zinchuk, A. (2023). Nicotine, alcohol, and caffeine use among individuals with untreated obstructive sleep apnea. *Sleep and Breathing*, *27*(6), 2479-2490.

Acknowledgements

Dr. Peng Wei

Professor

The University of Texas MD Anderson Cancer Center

Dr. Ryan Sun

Assistant Professor

The University of Texas MD Anderson Cancer Center

Dr. Hongyu Zhao

Ira V. Hiscock Professor

Yale University

Collaborators

Dr. Chunlin Li, Iowa State University

Dr. Tianzhong Yang, University of Minnesota

Dr. Sunyi Chi, Amazon

Dr. Henry Yaggi, Yale University

Dr. Andrey Zinchuk, Yale University

Dr. Jen-hwa Chu, Biogen

Dr. Jingfei Ma, UTMDACC

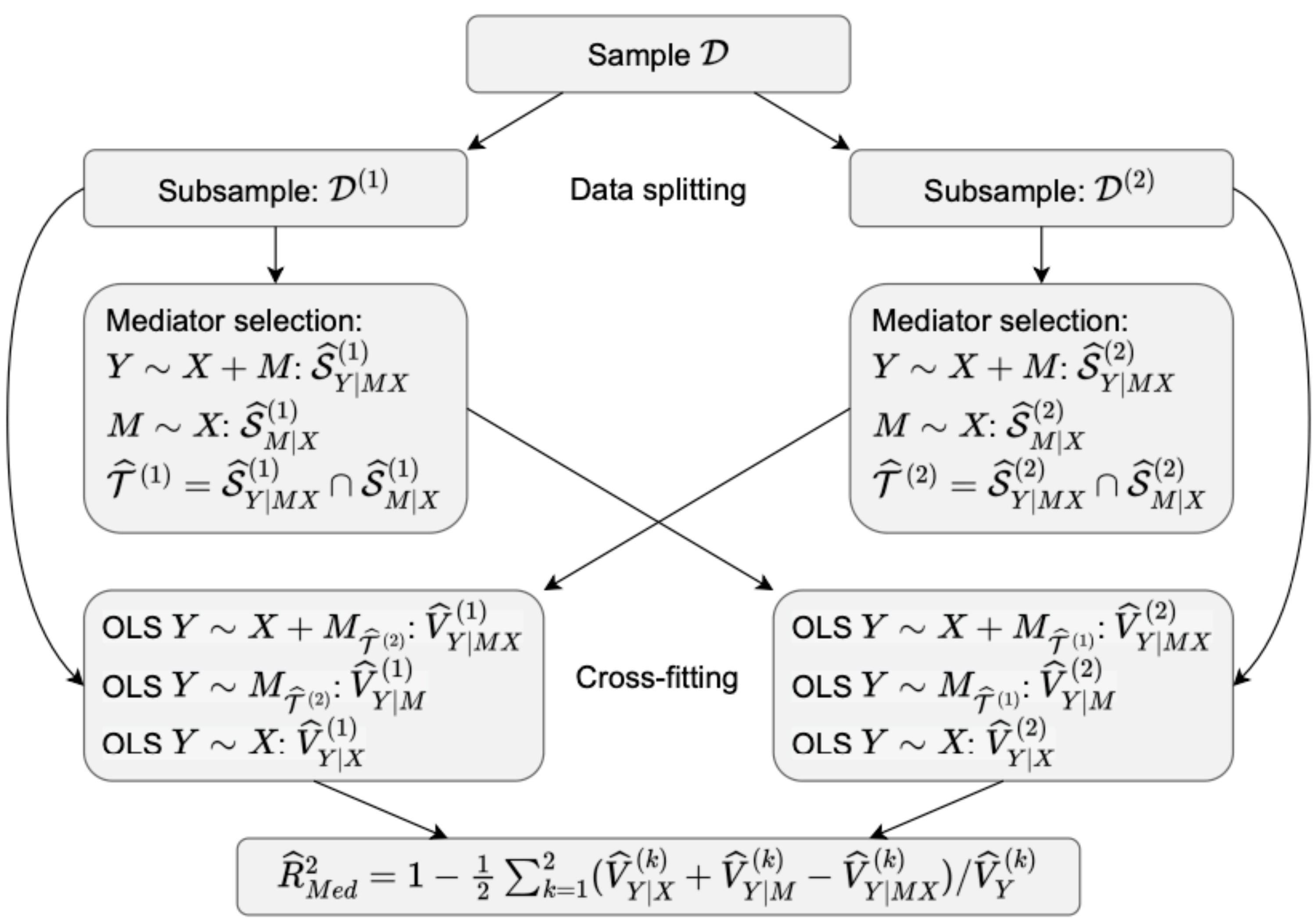
Dr. Gaiane M. Rauch, UTMDACC

Thanks & Questions :)

APPENDIX

Interval estimation for R^2-based mediation measure
Method

Figure 5: Cross-fitted estimation of R2



Interval estimation for R^2-based mediation measure
Simulation setting

- Use iterative Sure Independence Screening[1] (iSIS) with Minimax Concave Penalty[2] (MCP) to exclude M_{I_2}
- Compare the proposed CF-OLS method with the previous B-Mixed method[3]
- Compute the coverage probability, bias, width of confidence interval, mean squared error (MSE), empirical standard deviation, selection accuracy, and computational efficiency

1. J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5):849–911, 2008.
2. C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics, 38 (2):894–942, 2010.
3. T. Yang, J. Niu, H. Chen, and P. Wei. Estimation of total mediation effect for high-dimensional omics mediators. BMC bioinformatics, 22(1):1–17, 2021.

Table 1: Details of simulation scenarios (A1)-(A6)

	# of M_T	**# of M_{I1}**	**# of M_{I2}**	**# of M_{I3}**
A1	15	0	0	1485
A2	150	0	0	1350
A3	150	1350	0	0
A4	150	0	1350	0

Interval estimation for R^2-based mediation measure

Simulation results

- Comparable and satisfactory coverage probability (CP)
- Lower empirical standard deviation of replicated estimations and MSE
- Better computational efficiency

Coverage Probability

Computational Time (sd)

Table 2: Simulation results using the CF-OLS and B-Mixed for independent mediators

		CF-OLS									B-Mixed							
Scenario (R^2_{Med})	N	CP %	Width ($\times 10^{-2}$)	SE (10^{-2})	Bias (10^{-2})	SD (10^{-2})	MSE (10^{-4})	TP	FP	Time	CP %	Width ($\times 10^{-2}$)	Bias (10^{-2})	SD (10^{-2})	MSE (10^{-4})	TP	FP	Time
A1	750	92.0	3.664	1.870	0.739	1.940	4.292	0.945	0.021	0.12 (0.00)	98.5	5.159	0.149	2.646	6.990	0.940	0.020	44.96 (2.27)
(0.065)	1500	93.5	2.601	1.327	0.658	1.316	2.155	0.929	0.018	3.44 (0.04)	95.0	3.615	0.236	2.084	4.377	0.923	0.015	85.09 (4.44)
	3000	93.5	1.844	0.941	0.133	0.994	1.001	0.967	0.008	4.80 (0.07)	93.0	2.591	0.138	1.491	2.230	0.968	0.008	153.49 (8.12)
A2	750	94.5	5.383	2.747	-0.032	2.736	7.450	0.403	0.001	1.98 (0.04)	95.0	7.702	-0.263	3.908	15.266	0.402	0.001	51.23 (2.83)
(0.418)	1500	92.0	3.787	1.932	0.334	1.956	3.920	0.694	0.003	5.30 (0.11)	94.0	5.353	0.355	2.647	7.097	0.696	0.003	88.22 (4.54)
	3000	94.5	2.691	1.373	-0.131	1.390	1.940	0.943	0.003	6.78 (0.04)	94.0	3.777	-0.103	1.953	3.807	0.943	0.002	149.68 (6.28)
A3	750	93.5	3.494	1.782	0.269	1.790	3.259	0.310	0.011	2.13 (0.04)	92.5	5.054	0.365	2.762	7.725	0.311	0.011	38.51 (1.56)
(0.064)	1500	95.0	2.431	1.240	0.198	1.259	1.617	0.505	0.026	5.10 (0.05)	94.0	3.390	-0.008	1.820	3.297	0.506	0.026	74.06 (2.69)
	3000	95.0	1.707	0.871	0.168	0.817	0.692	0.762	0.065	8.62 (0.10)	96.0	2.391	0.015	1.118	1.245	0.763	0.065	147.08 (4.46)
A4	750	96.0	5.445	2.778	0.029	2.769	7.630	0.130	0.025	1.47 (0.03)	93.5	7.781	-0.227	4.088	16.680	0.131	0.026	41.79 (1.54)
(0.390)	1500	95.0	3.845	1.962	-0.255	1.956	3.873	0.386	0.022	4.95 (0.08)	96.5	5.430	-0.456	2.479	6.321	0.382	0.022	72.28 (2.57)
	3000	97.0	2.720	1.388	0.113	1.303	1.702	0.724	0.001	6.78 (0.12)	95.0	3.831	-0.011	1.839	3.367	0.723	0.002	125.16 (3.89)

Interval estimation for R^2-based mediation measure
Application to Framingham Heart Study (FHS)

- Metrics:
  - R^2_{Med}
  - Shared Over Simple (SOS) = R^2_{Med} / $R^2_{Y,X}$
  - $R^2_{Y,X}$
  - product measure (ab)
  - proportion measure ($prop$)
  - total effect measure

Table 3: Mediation effect size and 95% confidence interval in the FHS data

Outcome	Exp	Method	R^2_{Med}	SOS	$R^2_{Y,X}$	ab	prop	total	p̂
Systolic BP (N=4542)	Age	CF-OLS	0.030 (0.021, 0.039)	0.262 (0.196, 0.329)	0.113	-4.628/-4.953	-7.103/-6.866	0.651/0.721	77/91
		B-Mixed	0.038 (0.013, 0.053)	0.333 (0.118, 0.458)	0.113 (0.091, 0.139)	-4.946 (-5.693, -4.347)	-7.030 (-7.635, -6.641)	0.704 (0.626, 0.786)	95 (56, 152)

Meta-analysis of R^2-based mediation effect
Simulation setting and results

- Comparable and satisfactory coverage probability (CP)
- Fix total sample size at 3000, equally divided into Q studies
- Individual level data ($Q = 1$) vs. summary statistics ($Q \neq 1$)

Table 4: Simulation results using fixed-effects meta-analysis

Setting N = 3000	# of studies	Estimate	Bias x10-2	MSE x10-2	Coverage Proability %
A 50-0-0-1950	1 (Pooled)	0.494	-0.023	0.018	94.40
	2	0.495	0.052	0.018	94.30
	3	0.496	0.146	0.018	93.80
	4	0.497	0.223	0.019	93.10
B 50-200-0-1750	1 (Pooled)	0.369	0.022	0.019	95.90
	2	0.370	0.105	0.019	95.50
	3	0.371	0.180	0.019	95.50
	4	0.372	0.268	0.020	94.90
C 50-200-0-1750	1 (Pooled)	0.204	0.037	0.018	94.80
	2	0.205	0.065	0.018	94.40
	3	0.205	0.076	0.018	94.90
	4	0.205	0.073	0.018	94.90
D 50-200-200-1550	1 (Pooled)	0.081	0.141	0.010	94.90
	2	0.081	0.111	0.010	94.70
	3	0.080	0.074	0.010	94.00
	4	0.080	0.031	0.010	93.90

Project 1: interval estimation for R^2-based mediation measure
Method

- Assumption 1 (Sure screening property): The mediator selection satisfies the property $P(\widehat{\mathcal{T}}^{(k)} \supseteq \mathcal{T}) \to 1$ as $n \to \infty$ for $k = 1,2$.
- Assumption 2: Suppose $|\alpha_j| \lesssim \sqrt{\log(p)/n}$ and $|\beta_j| \lesssim \sqrt{\log(p)/n}$ for $j \notin \mathcal{T}$.
- Assumption 3: Suppose $\max\{|\Sigma_{kj}| : k \in \mathcal{T}, j \in \mathcal{T}^c\} \lesssim \sqrt{\log(p)/n}$ and $c_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_2$, where Σ is the covariance of ξ.

Project 1: interval estimation for R^2-based mediation measure

B-Mixed

Modelling and estimation

In order to obtain stable estimation under high-dimensional settings, we use the mixed-effect model for improved statistical efficiency, as shown later in the numerical examples. Specifically, we assume that the coefficients for the mediators in models (2) and (6) are random effects. In model (2), b_j is assumed to follow a normal distribution $b_j \sim N(0, \tau_1)$ for $j = 1, 2, \ldots, p$ and $e_2 \sim N(0, \phi_1)$, thus

$$R^2_{Y,MX} = 1 - \phi_1. \tag{7}$$

$R^2_{Y,MX}$ can be interpreted as one minus the variance that is unexplained by the independent variable and mediators. Similarly, in model (6), we assume $d_j \sim N(0, \tau_2)$ for $j = 1, 2, \ldots, p$ and $e_4 \sim N(0, \phi_2)$, such that $R^2_{Y,M} = 1 - \phi_2$.

We estimate τ_1,τ_2, ϕ_1 and ϕ_2 by the restricted maximum likelihood method, which is consistent under mild conditions [34]. Note that we avoid the direct use of the estimation of a total of $2p$ coefficients $(\beta_1, \ldots, \beta_p, d_1, \ldots, d_p)$; instead, we use two parameters (ϕ_1 and ϕ_2) to calculate $R^2_{Mediated}$. The estimation of latter is robust to the misspecification of the distribution of the random effects; it has been supported by multiple theoretical studies and real-data analysis [35–37]. Finally, $\hat{r}^2_{Y,X} = \sum_{i=1}^{n} \hat{y}_i^2/(n-2)$, where $\hat{y}_i$ is the fitted value estimated by MLE in model (1).

When $p << n$, it is also feasible to estimate the three R^2 components by MLE in the fixed-effect models (also proposed in Lachowicz 2018 [38]), and we evaluate its performance in the simulation study for comparison.

Project 1: interval estimation for R^2-based mediation measure

Method

- Suppose Assumptions 1-3 are met. If $|\mathcal{T}| + |\mathcal{I}_1| + |\mathcal{I}_2| \leq s$, $\max\{|\widehat{\mathcal{T}}^{(1)}|, |\widehat{\mathcal{T}}^{(2)}|\} \leq s$, and $s\log(p)/\sqrt{n} = o(1)$, then we will have:
- $\sqrt{n}(\widehat{R}^2_{Med} - R^2_{Med})/\sqrt{u^T A u} \xrightarrow{d} N(0,1)$
- $u = (1/V_Y, -1/V_Y, -1/V_Y, (V_{Y|X} + V_{Y|M} - V_{Y|MX})/V_Y^2)$
- A is the (constant) covariance matrix of $(\varepsilon^2, \eta^2, \zeta^2, Y^2)$

Project 1: interval estimation for R^2-based mediation measure

Method

Proof of Theorem 1

Before proceeding, we first introduce some notations used in the proof. Recalling Equation (1) in the main text, we have

$$\boldsymbol{M} = \boldsymbol{\alpha} X + \boldsymbol{\xi}, \qquad Y = \gamma X + \boldsymbol{\beta}_{\mathcal{T}}^\top \boldsymbol{M}_{\mathcal{T}} + \boldsymbol{\beta}_{\mathcal{I}_1}^\top \boldsymbol{M}_{\mathcal{I}_1} + \varepsilon.$$

Let $\mathcal{A}$ denote a generic subset $\mathcal{T} \subseteq \mathcal{A} \subseteq \{1, \ldots, p\}$ such that $\mathcal{A} \subseteq \mathcal{T} \cup \mathcal{I}_1$ and $\mathcal{A} \subseteq \mathcal{T} \cup \mathcal{I}_2$. Define

$$\begin{aligned}
\eta &= \varepsilon + \boldsymbol{\beta}^\top \{\boldsymbol{M} - \mathrm{E}(\boldsymbol{M} \mid X)\}, \\
\omega_{\mathcal{A}} &= \varepsilon + \boldsymbol{\beta}_{\mathcal{I}_1}^\top \{\boldsymbol{M}_{\mathcal{I}_1} - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid X, \boldsymbol{M}_{\mathcal{A}})\}, \\
\zeta_{\mathcal{A}} &= \gamma\big\{X - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{A}})\big\} + \varepsilon + \boldsymbol{\beta}_{\mathcal{I}_1}^\top \{\boldsymbol{M}_{\mathcal{I}_1} - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid \boldsymbol{M}_{\mathcal{A}})\}.
\end{aligned}$$

Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_n)$, $\boldsymbol{\zeta}_{\mathcal{A}} = (\zeta_{\mathcal{A},1}, \ldots, \zeta_{\mathcal{A},n})$, $\widehat{\boldsymbol{\zeta}}_{\mathcal{A}} = (\widehat{\zeta}_{\mathcal{A},1}, \ldots, \widehat{\zeta}_{\mathcal{A},n})$, $\boldsymbol{\omega}_{\mathcal{A}} = (\omega_{\mathcal{A},1}, \ldots, \omega_{\mathcal{A},n})$, and $\widehat{\boldsymbol{\omega}}_{\mathcal{A}} = (\widehat{\omega}_{\mathcal{A},1}, \ldots, \widehat{\omega}_{\mathcal{A},n})$, where $(\eta_i, \zeta_{\mathcal{A},i}, \omega_{\mathcal{A},i})$ are independent and identically distributed copies of $(\eta, \zeta_{\mathcal{A}}, \omega_{\mathcal{A}})$, and $\widehat{\eta}_i$, $\widehat{\zeta}_{\mathcal{A},i}$, and $\widehat{\omega}_{\mathcal{A},i}$ are the residuals of OLS regressions of Y over X, over $\boldsymbol{M}_{\mathcal{A}}$, and over $(X, \boldsymbol{M}_{\mathcal{A}})$. Further, we denote $\boldsymbol{y} = (Y_1, \ldots, Y_n)$ and define

$$\begin{aligned}
\rho_{\mathcal{A}}^2 &= 1 - \Big(\mathrm{E}\,\eta^2 + \mathrm{E}\,\zeta_{\mathcal{A}}^2 - \mathrm{E}\,\omega_{\mathcal{A}}^2\Big)/\,\mathrm{E}\,Y^2, \\
\widetilde{\rho}_{\mathcal{A}}^2 &= 1 - \Big(\frac{\boldsymbol{\eta}^\top \boldsymbol{\eta}}{n} + \frac{\boldsymbol{\zeta}_{\mathcal{A}}^\top \boldsymbol{\zeta}_{\mathcal{A}}}{n} - \frac{\boldsymbol{\omega}_{\mathcal{A}}^\top \boldsymbol{\omega}_{\mathcal{A}}}{n}\Big)/\Big(\frac{\boldsymbol{y}^\top \boldsymbol{y}}{n}\Big), \\
\widehat{\rho}_{\mathcal{A}}^2 &= 1 - \Big(\frac{\widehat{\boldsymbol{\eta}}^\top \widehat{\boldsymbol{\eta}}}{n} + \frac{\widehat{\boldsymbol{\zeta}}_{\mathcal{A}}^\top \widehat{\boldsymbol{\zeta}}_{\mathcal{A}}}{n} - \frac{\widehat{\boldsymbol{\omega}}_{\mathcal{A}}^\top \widehat{\boldsymbol{\omega}}_{\mathcal{A}}}{n}\Big)/\Big(\frac{\boldsymbol{y}^\top \boldsymbol{y}}{n}\Big),
\end{aligned}$$

As a result, we have $R_{Med}^2 = \rho_{\mathcal{A}=\mathcal{T}}^2$. We simplify the notation of $\rho_{\mathcal{A}=\mathcal{T}}^2$ as $\rho_{\mathcal{T}}^2$.

Now, we outline the strategy for establishing the asymptotic distribution of $\widehat{R}_{Med}^2$. Firstly, we control the difference $|\widehat{\rho}_{\mathcal{A}}^2 - \widetilde{\rho}_{\mathcal{A}}^2|$ uniformly in $\mathcal{A}$. Secondly, we upper-bound $|\widetilde{\rho}_{\mathcal{A}}^2 - \widetilde{\rho}_{\mathcal{T}}^2|$ uniformly in $\mathcal{A}$. Thirdly, we derive the asymptotics for the oracle estimator $\widetilde{\rho}_{\mathcal{T}}^2$. Finally, we establish the large-sample properties of $\widehat{R}_{Med}^2$ estimated by our proposed algorithm.

Project 1: interval estimation for R^2-based mediation measure

Method

Bounding the difference between $\widehat{\rho}^2_{\mathcal{A}}$ and $\widetilde{\rho}^2_{\mathcal{A}}$. Note that

$$\widehat{\boldsymbol{\eta}}^\top \widehat{\boldsymbol{\eta}} = \boldsymbol{\eta}^\top \boldsymbol{\eta} - \boldsymbol{\eta}^\top \boldsymbol{P}_X \boldsymbol{\eta},$$
$$\widehat{\boldsymbol{\zeta}}_{\mathcal{A}}^\top \widehat{\boldsymbol{\zeta}}_{\mathcal{A}} = \boldsymbol{\zeta}_{\mathcal{A}}^\top \boldsymbol{\zeta}_{\mathcal{A}} - \boldsymbol{\zeta}_{\mathcal{A}}^\top \boldsymbol{P}_{\boldsymbol{M}_{\mathcal{A}}} \boldsymbol{\zeta}_{\mathcal{A}},$$
$$\widehat{\boldsymbol{\omega}}_{\mathcal{A}}^\top \widehat{\boldsymbol{\omega}}_{\mathcal{A}} = \boldsymbol{\omega}_{\mathcal{A}}^\top \boldsymbol{\omega}_{\mathcal{A}} - \boldsymbol{\omega}_{\mathcal{A}}^\top \boldsymbol{P}_{X,\boldsymbol{M}_{\mathcal{A}}} \boldsymbol{\omega}_{\mathcal{A}},$$

where $\boldsymbol{P}_X$, $\boldsymbol{P}_{\boldsymbol{M}_{\mathcal{A}}}$, and $\boldsymbol{P}_{X,\boldsymbol{M}_{\mathcal{A}}}$ are the projection matrices onto the column spaces of X, $\boldsymbol{M}_{\mathcal{A}}$, and $(X, \boldsymbol{M}_{\mathcal{A}})$, respectively. Thus, we have

$$|\widehat{\rho}^2_{\mathcal{A}} - \widetilde{\rho}^2_{\mathcal{A}}| \lesssim \frac{1}{n} \left(\boldsymbol{\eta}^\top \boldsymbol{P}_X \boldsymbol{\eta} + \boldsymbol{\zeta}_{\mathcal{A}}^\top \boldsymbol{P}_{\boldsymbol{M}_{\mathcal{A}}} \boldsymbol{\zeta}_{\mathcal{A}} + \boldsymbol{\omega}_{\mathcal{A}}^\top \boldsymbol{P}_{X,\boldsymbol{M}_{\mathcal{A}}} \boldsymbol{\omega}_{\mathcal{A}} \right).$$

By Theorem 2.1 of Hsu et al. (2012), there exists a constant $C > 0$ such that we have

$$\mathbb{P}\Big(\boldsymbol{\eta}^\top \boldsymbol{P}_X \boldsymbol{\eta} \geq C(1 + 2\sqrt{t} + 2t)\Big) \leq \exp(-t),$$
$$\mathbb{P}\Big(\boldsymbol{\zeta}_{\mathcal{A}}^\top \boldsymbol{P}_{\boldsymbol{M}_{\mathcal{A}}} \boldsymbol{\zeta}_{\mathcal{A}} \geq C(|\mathcal{A}| + \sqrt{|\mathcal{A}|t} + 2t)\Big) \leq \exp(-t),$$
$$\mathbb{P}\Big(\boldsymbol{\omega}_{\mathcal{A}}^\top \boldsymbol{P}_{X,\boldsymbol{M}_{\mathcal{A}}} \boldsymbol{\omega}_{\mathcal{A}} \geq C(|\mathcal{A}| + \sqrt{|\mathcal{A}|t} + 2t)\Big) \leq \exp(-t),$$

for any $\mathcal{A}$ with $|\mathcal{A}| \leq s$. Consequently, if $t = s\log(p)$ and if $s\log(p) \ll n$, we have

$$\sqrt{n} \sup_{\mathcal{A}} |\widehat{\rho}^2_{\mathcal{A}} - \widetilde{\rho}^2_{\mathcal{A}}| \lesssim s\log(p)/\sqrt{n}.$$

This provides a uniform bound of $|\widehat{\rho}^2_{\mathcal{A}} - \widetilde{\rho}^2_{\mathcal{A}}|$ for any $|\mathcal{A}| \leq s$.

Project 1: interval estimation for R^2-based mediation measure

Method

Bounding the difference between $\widetilde{\rho}^2_{\mathcal{A}}$ and $\widetilde{\rho}^2_{\mathcal{T}}$. Note that

$$\begin{aligned}\sqrt{n}|\widetilde{\rho}^2_{\mathcal{A}} - \widetilde{\rho}^2_{\mathcal{T}}| &\lesssim \frac{1}{\sqrt{n}} \left|(\boldsymbol{\zeta}_{\mathcal{A}}^\top \boldsymbol{\zeta}_{\mathcal{A}} - \boldsymbol{\omega}_{\mathcal{A}}^\top \boldsymbol{\omega}_{\mathcal{A}}) - (\boldsymbol{\zeta}_{\mathcal{T}}^\top \boldsymbol{\zeta}_{\mathcal{T}} - \boldsymbol{\omega}_{\mathcal{T}}^\top \boldsymbol{\omega}_{\mathcal{T}})\right| \\ &\le \frac{1}{\sqrt{n}} \left|\boldsymbol{\zeta}_{\mathcal{A}}^\top \boldsymbol{\zeta}_{\mathcal{A}} - \boldsymbol{\zeta}_{\mathcal{T}}^\top \boldsymbol{\zeta}_{\mathcal{T}}\right| + \frac{1}{\sqrt{n}} \left|\boldsymbol{\omega}_{\mathcal{A}}^\top \boldsymbol{\omega}_{\mathcal{A}} - \boldsymbol{\omega}_{\mathcal{T}}^\top \boldsymbol{\omega}_{\mathcal{T}}\right|.\end{aligned}$$

To establish its upper bound, denoting

$$D_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\zeta^2_{\mathcal{A},i} - \zeta^2_{\mathcal{T},i}), \qquad D_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\omega^2_{\mathcal{A},i} - \omega^2_{\mathcal{T},i}),$$

we would like to show $\mathrm{E}(D_1) = o(1)$, $\mathrm{Var}(D_1) = o(1)$, $\mathrm{E}(D_2) = o(1)$, and $\mathrm{Var}(D_2) = o(1)$. To this end, consider an independent observation, $(\zeta_{\mathcal{A}}, \zeta_{\mathcal{T}}, \omega_{\mathcal{A}}, \omega_{\mathcal{T}})$. We have

$$\begin{aligned}\mathrm{E}(D_1) &= \sqrt{n}\,\mathrm{E}(\zeta^2_{\mathcal{A}} - \zeta^2_{\mathcal{T}}), \quad \mathrm{Var}(D_1) = \mathrm{Var}(\zeta^2_{\mathcal{A}} - \zeta^2_{\mathcal{T}}), \\ \mathrm{E}(D_2) &= \sqrt{n}\,\mathrm{E}(\omega^2_{\mathcal{A}} - \omega^2_{\mathcal{T}}), \quad \mathrm{Var}(D_2) = \mathrm{Var}(\omega^2_{\mathcal{A}} - \omega^2_{\mathcal{T}}).\end{aligned}$$

For $\mathrm{E}(D_1)$, we have

$$\mathrm{E}(D_1) = \sqrt{n}\gamma^2(v_1 - v_2) + \sqrt{n}\boldsymbol{\beta}_{\mathcal{I}_1}^\top (\boldsymbol{O}_1 - \boldsymbol{O}_2)\boldsymbol{\beta}_{\mathcal{I}_1},$$

where

$$\begin{aligned}v_1 &= \mathrm{E}\left\{X - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{A}})\right\}^2, \\ v_2 &= \mathrm{E}\left\{X - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{T}})\right\}^2, \\ \boldsymbol{O}_1 &= \mathrm{E}\left\{X - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{A}})\right\}\left\{X - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{A}})\right\}^\top, \\ \boldsymbol{O}_2 &= \mathrm{E}\left\{X - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{T}})\right\}\left\{X - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{T}})\right\}^\top.\end{aligned}$$

Note that

$$\sqrt{n}(v_1 - v_2) = \frac{\sigma_X^2 \boldsymbol{\alpha}_{\mathcal{A}}^\top (\boldsymbol{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}} - \widetilde{\boldsymbol{\Sigma}}^{-1}_{\mathcal{A}\mathcal{A}})\boldsymbol{\alpha}_{\mathcal{A}}}{(1 + \boldsymbol{\alpha}_{\mathcal{T}}^\top \boldsymbol{\Sigma}^{-1}_{\mathcal{T}\mathcal{T}} \boldsymbol{\alpha}_{\mathcal{T}})(1 + \boldsymbol{\alpha}_{\mathcal{A}}^\top \boldsymbol{\Sigma}^{-1}_{\mathcal{A}\mathcal{A}} \boldsymbol{\alpha}_{\mathcal{A}})} \lesssim \|\boldsymbol{\alpha}_{\mathcal{B}}\|_2^2 \lesssim s\log(p)/\sqrt{n} = o(1),$$

where $\widetilde{\boldsymbol{\Sigma}}^{-1}_{\mathcal{A}\mathcal{A}}$ is the Moore-Penrose inverse of block diagonal matrix $\mathrm{Diag}(\boldsymbol{\Sigma}_{\mathcal{T}\mathcal{T}}, \boldsymbol{0})$, $\boldsymbol{\alpha}_{\mathcal{A}} = (\boldsymbol{\alpha}_{\mathcal{T}}, \boldsymbol{\alpha}_{\mathcal{B}})$, and $\mathcal{B} = \mathcal{A} \setminus \mathcal{T}$. Also, note that

$$\sqrt{n}\boldsymbol{\beta}_{\mathcal{I}_1}^\top (\boldsymbol{O}_1 - \boldsymbol{O}_2)\boldsymbol{\beta}_{\mathcal{I}_1} \lesssim \|\boldsymbol{\beta}_{\mathcal{I}_1}\|_2^2 \lesssim s\log(p)/\sqrt{n} = o(1).$$

Thus, $\mathrm{E}(D_1) = o(1)$.

Project 1: interval estimation for R^2-based mediation measure

Method

For $\mathrm{Var}(D_1)$, we have

$$\mathrm{Var}(D_1) \le \mathrm{E}\{(\zeta_{\mathcal{A}} + \zeta_{\mathcal{T}})^2(\zeta_{\mathcal{A}} - \zeta_{\mathcal{T}})^2\} \lesssim \mathrm{E}(\zeta_{\mathcal{A}} - \zeta_{\mathcal{T}})^2.$$

Note that

$$|\zeta_{\mathcal{A}} - \zeta_{\mathcal{T}}| \le \gamma |\mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{A}}) - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{T}})| + |\boldsymbol{\beta}_{\mathcal{I}_1}^{\top}\{\mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid \boldsymbol{M}_{\mathcal{A}}) - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid \boldsymbol{M}_{\mathcal{T}})\}|,$$

where

$$\mathrm{E}\,|\,\mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{A}}) - \mathrm{E}(X \mid \boldsymbol{M}_{\mathcal{T}})|^2 \le \|\boldsymbol{\alpha}_{\mathcal{I}_2}\|_2^2 \lesssim s\log(p)/n = o(1),$$
$$\mathrm{E}\,|\boldsymbol{\beta}_{\mathcal{I}_1}^{\top}\{\mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid \boldsymbol{M}_{\mathcal{A}}) - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid \boldsymbol{M}_{\mathcal{T}})\}|^2 \lesssim \|\boldsymbol{\beta}_{\mathcal{I}_1}\|_2^2 \lesssim s\log(p)/n = o(1).$$

Therefore, $\mathrm{Var}(D_1) = o(1)$.

Similarly, for $\mathrm{E}(D_2)$ we have

$$\mathrm{E}(D_2) = \sqrt{n}\boldsymbol{\beta}_{\mathcal{I}_1}^{\top}(\boldsymbol{Q}_1 - \boldsymbol{Q}_2)\boldsymbol{\beta}_{\mathcal{I}_1} \lesssim \|\boldsymbol{\beta}_{\mathcal{I}_1}\|_2^2 \lesssim s\log(p)/\sqrt{n} = o(1),$$

where

$$\boldsymbol{Q}_1 = \mathrm{E}\left\{\boldsymbol{M}_{\mathcal{I}_1} - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid X, \boldsymbol{M}_{\mathcal{A}})\right\}\left\{\boldsymbol{M}_{\mathcal{I}_1} - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid X, \boldsymbol{M}_{\mathcal{A}})\right\}^{\top},$$
$$\boldsymbol{Q}_2 = \mathrm{E}\left\{\boldsymbol{M}_{\mathcal{I}_1} - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid X, \boldsymbol{M}_{\mathcal{A}})\right\}\left\{\boldsymbol{M}_{\mathcal{I}_1} - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid X, \boldsymbol{M}_{\mathcal{A}})\right\}^{\top}.$$

Also, we have

$$\mathrm{Var}(D_2) \le \mathrm{E}\{(\omega_{\mathcal{A}} + \omega_{\mathcal{T}})^2(\omega_{\mathcal{A}} - \omega_{\mathcal{T}})^2\} \lesssim \mathrm{E}(\omega_{\mathcal{A}} - \omega_{\mathcal{T}})^2.$$

Further, note that

$$|\omega_{\mathcal{A}} - \omega_{\mathcal{T}}| \lesssim |\boldsymbol{\beta}_{\mathcal{I}_1}^{\top}\{\mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid X, \boldsymbol{M}_{\mathcal{A}}) - \mathrm{E}(\boldsymbol{M}_{\mathcal{I}_1} \mid X, \boldsymbol{M}_{\mathcal{T}})\}| \lesssim \|\boldsymbol{\beta}_{\mathcal{I}_1}\|_2 \lesssim \sqrt{s\log(p)/n}.$$

Thus, $\mathrm{Var}(D_2) \lesssim s\log(p)/n = o(1)$.

As a result, we have

$$\sqrt{n}|\tilde{\rho}_{\mathcal{A}}^2 - \tilde{\rho}_{\mathcal{T}}^2| \lesssim D_1 + D_2 = o_p(1).$$

This bound holds uniformly for $|\mathcal{A}| \le s$.

Project 1: interval estimation for R^2-based mediation measure

Method

Analysis of the oracle estimator. Now, we turn to the asymptotic distribution of $\sqrt{n}(\widetilde{\rho}^2_{\mathcal{T}} - \rho^2_{\mathcal{T}})$. Note that $\boldsymbol{\varepsilon} = \boldsymbol{\omega}_{\mathcal{T}}$ and $\boldsymbol{\zeta} = \boldsymbol{\zeta}_{\mathcal{T}}$. By the central limit theorem,

$$\frac{1}{\sqrt{n}}\begin{pmatrix} \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} - V_{Y|MX} \\ \boldsymbol{\eta}^\top\boldsymbol{\eta} - V_{Y|X} \\ \boldsymbol{\zeta}^\top\boldsymbol{\zeta} - V_{Y|M} \\ \boldsymbol{y}^\top\boldsymbol{y} - V_Y \end{pmatrix} \xrightarrow{d} N\left(\mathbf{0}, \underbrace{\begin{pmatrix} \mathrm{Var}(\varepsilon^2) & \mathrm{Cov}(\varepsilon^2,\eta^2) & \mathrm{Cov}(\varepsilon^2,\zeta^2) & \mathrm{Cov}(\varepsilon^2,Y^2) \\ \mathrm{Cov}(\varepsilon^2,\eta^2) & \mathrm{Var}(\eta^2) & \mathrm{Cov}(\eta^2,\zeta^2) & \mathrm{Cov}(\eta^2,Y^2) \\ \mathrm{Cov}(\varepsilon^2,\zeta^2) & \mathrm{Cov}(\eta^2,\zeta^2) & \mathrm{Var}(\zeta^2) & \mathrm{Cov}(\zeta^2,Y^2) \\ \mathrm{Cov}(\varepsilon^2,Y^2) & \mathrm{Cov}(\eta^2,Y^2) & \mathrm{Cov}(\zeta^2,Y^2) & \mathrm{Var}(Y^2) \end{pmatrix}}_{=\boldsymbol{A}}\right).$$

Consequently,

$$\sqrt{n}(\widetilde{\rho}^2_{\mathcal{T}} - \rho^2_{\mathcal{T}})/\sqrt{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} \xrightarrow{d} N(0,1),$$

where $\boldsymbol{u} = (1/V_Y, -1/V_Y, -1/V_Y, (V_{Y|X} + V_{Y|M} - V_{Y|MX})/V_Y^2)$.

Asymptotic distribution of $\sqrt{n}(\widehat{R}^2_{Med} - R^2_{Med})$. We characterize $\widehat{\rho}^2_{\widehat{\mathcal{T}}^{(2)}}$ and $\widehat{\rho}^2_{\widehat{\mathcal{T}}^{(1)}}$ as $1 - (\widehat{V}^{(1)}_{Y|X} + \widehat{V}^{(1)}_{Y|M} - \widehat{V}^{(1)}_{Y|MX})/\widehat{V}^{(1)}_Y$ and $1 - (\widehat{V}^{(2)}_{Y|X} + \widehat{V}^{(2)}_{Y|M} - \widehat{V}^{(2)}_{Y|MX})/\widehat{V}^{(2)}_Y$, respectively. Following the above analysis, with n being replaced by $n/2$, we have $\widehat{\rho}^2_{(1)} = 1 - (\widehat{V}^{(1)}_{Y|X} + \widehat{V}^{(1)}_{Y|M} - \widehat{V}^{(1)}_{Y|MX})/\widehat{V}^{(1)}_Y$ and $\widehat{\rho}^2_{(2)} = 1 - (\widehat{V}^{(2)}_{Y|X} + \widehat{V}^{(2)}_{Y|M} - \widehat{V}^{(2)}_{Y|MX})/\widehat{V}^{(2)}_Y$, both are asymptotically independent and normal in that

$$\sqrt{n/2}(\widehat{\rho}^2_{(1)} - R^2_{Med})/\sqrt{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} \xrightarrow{d} N(0,1),$$
$$\sqrt{n/2}(\widehat{\rho}^2_{(2)} - R^2_{Med})/\sqrt{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} \xrightarrow{d} N(0,1).$$

Consequently,

$$\sqrt{n}(\widehat{R}^2_{Med} - R^2_{Med})/\sqrt{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} = \sqrt{n/2}((\widehat{\rho}^2_{(1)} + \widehat{\rho}^2_{(2)})/2 - R^2_{Med})/\sqrt{\boldsymbol{u}^\top\boldsymbol{A}\boldsymbol{u}} \xrightarrow{d} N(0,1),$$

which completes the proof.

Asymptotic distribution of SOS. The shared over simple effect (SOS) is defined as

$$\mathrm{SOS} = \frac{R^2_{Mediated}}{R^2_{Y,X}} = \frac{V_Y - V_{Y|X} - V_{Y|M} + V_{Y|MX}}{V_Y - V_{Y|X}} = 1 - \frac{V_{Y|M} - V_{Y|MX}}{V_Y - V_{Y|X}}.$$

Let the estimate for SOS be $\widehat{\mathrm{SOS}} = \frac{1}{2}\sum_{k=1}^2 (1 - (\widehat{V}^{(k)}_{Y|M} - \widehat{V}^{(k)}_{Y|MX})/(\widehat{V}^{(k)}_Y - \widehat{V}^{(k)}_{Y|X}))$. Similarly, we can show that

$$\sqrt{n}(\widehat{\mathrm{SOS}}^2 - \mathrm{SOS})/\sqrt{\boldsymbol{v}^\top\boldsymbol{A}\boldsymbol{v}} \xrightarrow{d} N(0,1),$$

where $\boldsymbol{v} = (1/(V_Y - V_{Y|X}), -(V_{Y|M} - V_{Y|MX})/(V_Y - V_{Y|X})^2, -1/(V_Y - V_{Y|X}), (V_{Y|M} - V_{Y|MX})/(V_Y - V_{Y|X})^2)$.

Project 1: interval estimation for R^2-based mediation measure
Method: relax normality assumption

Relaxing normality assumption. We have assumed the variables are jointly normal to avoid technicality and to improve the presentation. This assumption, however, is unnecessary. In fact, any conditional expectation $\mathrm{E}(\cdot \mid \star)$ in our derivation can be replaced with the "best linear approximation" operator $\mathbb{L}(\cdot \mid \star)$, defined as follows. Given random variables U and $\boldsymbol{W}$, let $\mathbb{L}(U \mid \boldsymbol{W})$ be the best linear approximation of U using $\boldsymbol{W}$, namely $\mathbb{L}(U \mid \boldsymbol{W}) = \widetilde{\boldsymbol{\theta}}^\top \boldsymbol{W}$ where

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ \mathrm{E}(U - \boldsymbol{\theta}^\top \boldsymbol{W})^2.$$

For random variables U, U', and $\boldsymbol{W}$, we have that (a) $\mathbb{L}(U + U' \mid \boldsymbol{W}) = \mathbb{L}(U \mid \boldsymbol{W}) + \mathbb{L}(U' \mid \boldsymbol{W})$, (b) $\mathbb{L}(cU \mid \boldsymbol{W}) = c\mathbb{L}(U \mid \boldsymbol{W})$ for $c \in \mathbb{R}$, (c) $\mathbb{L}(U \mid \boldsymbol{W}) = 0$ if $\mathrm{Cov}(U, \boldsymbol{W}) = \mathbf{0}$, (d) $\mathbb{L}(U \mid \boldsymbol{W}) = U$ if $U \in \mathrm{Span}(\boldsymbol{W})$, and (e) $\mathbb{L}(U \mid \boldsymbol{W}) = \mathbb{L}(U \mid \mathbf{A}\boldsymbol{W})$ for invertible $\mathbf{A}$. Thus, $\mathbb{L}(\cdot \mid \star)$ mimics $\mathbb{E}(\cdot \mid \star)$. As a result, if the data are sub-Gaussian, the proof of Theorem 1 continues to hold with $\mathrm{E}(\cdot \mid \star)$ being replaced by $\mathbb{L}(\cdot \mid \star)$.

Project 1: interval estimation for R^2-based mediation measure

Method: true value of R2

1 Previous R2

Suppose we have the following set of regression models:

$$\begin{aligned} Y &= cX + e_1, \\ Y &= rX + \sum_{j=1}^{p} M_j b_j + e_2, \\ M_j &= a_j X + \xi_j, \end{aligned}$$

where p is the number of mediators, $e_2 \sim N(0, \phi_1)$, and $\xi = [\xi_1, \ldots, \xi_p] \sim N(0, \mathbf{D}_{p\times p})$. Without loss of generality, we assume X and M have variances of 1, and Y is centered at 0. $R^2_{Mediated}$ is defined as

$$R^2_{Mediated} = R^2_{Y,M} + r^2_{Y,X} - R^2_{Y,MX}.$$

Then, we derive $r^2_{Y,X}$, $R^2_{Y,MX}$ and $R^2_{Y,M}$ separately. First, we have

$$\begin{aligned} r^2_{Y,X} &= \mathrm{Cor}^2(Y,X) = \frac{1}{\sigma_Y^2}\mathrm{Cov}^2(Y,X) \\ &= \frac{1}{\sigma_Y^2}\mathrm{Cov}^2\left(X, rX + \sum_{j=1}^{p} M_j b_j + e_2\right) \\ &= \frac{1}{\sigma_Y^2}\left[\mathrm{Cov}(rX, X) + \mathrm{Cov}\left(X, \sum_{j=1}^{p} M_j b_j\right) + \mathrm{Cov}(X, e_2)\right]^2 \\ &= \frac{1}{\sigma_Y^2}\left[r + \sum_{j=1}^{p} a_j b_j\right]^2 = \frac{(r + \mathbf{b}^T\mathbf{a})^2}{\sigma_Y^2} \end{aligned}$$

Project 1: interval estimation for R^2-based mediation measure

Method: true value of R2

It can also be shown that

$$
\begin{aligned}
R^2_{Y,MX} &= \frac{1}{\sigma_Y^2}\mathrm{Var}\left(rX + \sum_{j=1}^{p} M_j b_j\right) \\
&= \frac{1}{\sigma_Y^2}\left[\mathrm{Var}(rX) + \mathrm{Var}\left(\sum_{j=1}^{p} M_j b_j\right) + 2\mathrm{Cov}\left(rX, \sum_{j=1}^{p} M_j b_j\right)\right] \\
&= \frac{1}{\sigma_Y^2}\left[r^2 + \mathrm{Var}\left(\sum_{j=1}^{p} a_j b_j X + \sum_{j=1}^{p} b_j \xi_j\right) + 2\mathrm{Cov}\left(rX, \sum_{j=1}^{p} a_j b_j X + \sum_{j=1}^{p} b_j \xi_j\right)\right] \\
&= \frac{1}{\sigma_Y^2}\left[r^2 + \left(\sum_{j=1}^{p} a_j b_j\right)^2 + \mathbf{b}^T\mathbf{D}\mathbf{b} + 2\mathrm{Cov}\left(rX, \sum_{j=1}^{p} a_j b_j X\right)\right] \\
&= \frac{1}{\sigma_Y^2}\left[r^2 + \left(\sum_{j=1}^{p} a_j b_j\right)^2 + \mathbf{b}^T\mathbf{D}\mathbf{b} + 2r\sum_{j=1}^{p} a_j b_j\right] \\
&= \frac{(r + (\sum_{j=1}^{p} a_j b_j))^2 + \mathbf{b}^T\mathbf{D}\mathbf{b}}{\sigma_Y^2} = \frac{(r + \mathbf{b}^T\mathbf{a})^2 + \mathbf{b}^T\mathbf{D}\mathbf{b}}{\sigma_Y^2}
\end{aligned}
$$

Finally, we compute $R^2_{Y,M}$. Suppose we have $h = (r_{M_1Y}, \ldots r_{M_pY},)^T$, and V_{MM} as a $p \times p$ matrix with $\mathrm{Cor}(M_i, M_j)$ as the (i,j)th entry. Note that

$$\mathrm{Cor}(M_i, M_j) = \mathrm{Cov}(a_iX + \xi_i, a_jX + \xi_j) = a_ia_j + D_{ij}.$$

Project 1: interval estimation for R^2-based mediation measure
Method: true value of R2

It follows that $V_{MM} = (\mathrm{Cor}(M_i, M_j))_{i,j} = \mathbf{a}\mathbf{a}^T + \mathbf{D}$. On the other hand,

$$\begin{aligned} r_{M_iY} &= \frac{1}{\sqrt{\sigma_Y^2}} \mathrm{Cov}\left(a_i X + \xi_i, rX + \sum_{j=1}^{p} M_j b_j + e_2\right) \\ &= \frac{1}{\sqrt{\sigma_Y^2}} \mathrm{Cov}\left(a_i X + \xi_i, rX + \sum_{j=1}^{p} a_j b_j X + \sum_{j=1}^{p} b_j \xi_j + e_2\right) \\ &= \frac{1}{\sqrt{\sigma_Y^2}} \left(a_i r + a_i \sum_{j=1}^{p} a_j b_j + \sum_{j=1}^{p} b_j D_{ij}\right). \end{aligned}$$

Hence, we have

$$h = \frac{1}{\sigma_Y}\left[r\mathbf{a} + (\mathbf{b}^T\mathbf{a})\mathbf{a} + \mathbf{D}\mathbf{b}\right].$$

In view of the Sherman–Morrison formula, we have

$$V_{MM}^{-1} = \mathbf{D}^{-1} - \frac{\mathbf{D}^{-1}\mathbf{a}\mathbf{a}^T\mathbf{D}^{-1}}{1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}}.$$

Consequently,

$$R_{Y,M}^2 = h^T V_{MM}^{-1} h = \frac{(r + \mathbf{b}^T\mathbf{a})^2 - r^2/(1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}) + \mathbf{b}^T\mathbf{D}\mathbf{b}}{\sigma_Y^2}$$

Project 1: interval estimation for R^2-based mediation measure

Range of R2

1.2.3 Range of the $R^2_{Mediated}$

Proposition 1: In the consistent model, where a_jb_j and r are in the same direction ($a_jb_jr > 0$) for $j = 1, 2, \ldots, p$, $R^2_{Mediated} \in (0, 1)$.

Proof: To show $R^2_{Mediated} > 0$ is equivalent to showing that the nominator $(r + \mathbf{b}^T\mathbf{a})^2 - r^2/(1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}) > 0$.

Since D is a (semi)positive definite matrix, $\mathbf{a}^T\mathbf{D}^{-1}\mathbf{a} > 0$.

Thus, $(r + \mathbf{b}^T\mathbf{a})^2 - r^2/(1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}) > (\mathbf{b}^T\mathbf{a})^2 + 2r\mathbf{b}^T\mathbf{a} > 0$.

In addition, $(r + \mathbf{b}^T\mathbf{a})^2 - r^2/(1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}) < (r + \mathbf{b}^T\mathbf{a})^2 + \mathbf{b}^T\mathbf{D}\mathbf{b} + \tau$ for any $\mathbf{a}$, $\mathbf{b}$ and r.

When $\mathbf{b}^T\mathbf{a} \to \infty$, $R^2_{Mediated} \to 1$.

Therefore, $R^2_{Mediated} \in (0, 1)$ under the consistent model.

Project 1: interval estimation for R^2-based mediation measure

Range of R2

Proposition 3: When $|r/c| > \sqrt{1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}}$, $R^2_{Mediated}$ is negative.

Proof: When r is large and $c \approx 0$,

$$R^2_{Mediated} \approx -\frac{r^2/(1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a})}{\sigma_Y^2}.$$

Since $r^2 > 0$ and $1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a} > 0$, $R^2_{Mediated} < 0$.

The first step to establish mediation according to Baron and Kelly (9) is that the independent variable must affect the dependent variable, that is, the total effect c should be different from 0. If the effect is not significant, the analysis for mediation analysis stops. Coinciding with this step, we show that the negativity of $R^2_{Mediated}$ happens when this criterion does not hold.

More generally, it can be proven that $R^2_{Mediated} < 0$ when $|r/c| > \sqrt{1 + \mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}}$ using algebra. Under the high-dimensional setting, $\mathbf{a}^T\mathbf{D}^{-1}\mathbf{a}$ can be large and c is large enough to pass the first step of (9); therefore, the scenario where $R^2_{Mediated} < 0$ may not be very likely to happen.

Project 1: interval estimation for R^2-based mediation measure

M1 does not bias

Proposition 4: $\mathbf{M}^{(1)}$ in the mediation model does not change the point estimation of $R^2_{Mediated}$.

Proof: Without loss of generality, we assume $\hat{\mathbf{M}} = [\mathbf{M}, \mathbf{M}^{(1)}]$, then $\mathbf{a}^T = [a_1, a_2, \ldots, a_t, 0, \ldots, 0]$. By some linear algebra, it can be shown that $\mathbf{a}^T \mathbf{D}^{-1} \mathbf{a} = [a_1, a_2, \ldots, a_t] \mathbf{D_{tt}} [a_1, a_2, \ldots, a_t]^T$, where $\mathbf{D_{tt}}$ is $\mathbf{D}^{-1}$ with the first t columns and the first t rows.

In addition, $\mathbf{b}^T \mathbf{a} = [b_1, b_2, \ldots, b_t][a_1, a_2, \ldots, a_t]^T$. Therefore,

$$R^2_{Mediated}(\hat{\mathbf{M}}) = R^2_{Mediated}(\mathbf{M})$$

In the special case where all $\mathbf{M} = \emptyset$,

$$R^2_{Y,MX} = R^2_{Y,M} + r^2_{Y,X},$$

$$R^2_{Mediated} = R^2_{Y,M} + r^2_{Y,X} - (R^2_{Y,M} + r^2_{Y,X}) = 0.$$

Therefore, the inclusion of $\mathbf{M}^{(1)}$ does not lead to bias in the point estimate of $R^2_{Mediated}$. When such variables are included in the mediation model, both $R^2_{Y,M}$ and $R^2_{Y,XM}$ increase the same amount. As a result, their effects cancel out in $R^2_{Mediated}$. Since the noise variables also have $a_j = 0$, they do not bias the estimation either.

Project 1: interval estimation for R^2-based mediation measure

Simulation: correlated mediators

Table 2. Simulation results using the CF-OLS method for highly-correlated putative mediators in scenarios (A7)–(A12).

Scenario	N	Correlation Structure 1								Correlation Structure 2							
Scenario (R^2_{Med})	N	CP %	Width ($\times 10^{-2}$)	SE ($\times 10^{-2}$)	Bias ($\times 10^{-2}$)	SD ($\times 10^{-2}$)	MSE ($\times 10^{-2}$)	TP %	FP %	CP %	Width ($\times 10^{-2}$)	SE ($\times 10^{-2}$)	Bias ($\times 10^{-2}$)	SD ($\times 10^{-2}$)	MSE ($\times 10^{-2}$)	TP %	FP %
A7	750	**91.5**	5.082	2.593	1.317	2.738	0.092	\	1.3	**91.5**	4.942	2.522	1.313	2.697	0.090	\	1.4
(0)	1500	**93.0**	3.489	1.780	0.876	1.946	0.045	\	1.2	**93.5**	3.550	1.811	0.720	1.911	0.042	\	1.3
	3000	**95.0**	2.497	1.274	0.281	1.336	0.019	\	1.2	**98.0**	2.494	1.272	0.177	1.146	0.013	\	1.3
A8	750	**95.0**	5.667	2.891	0.455	2.732	0.076	100.0	0.0	**93.0**	5.162	2.634	0.037	2.811	0.079	100.0	0.0
(0.128)	1500	**93.5**	3.992	2.037	-0.163	2.165	0.047	100.0	0.0	**94.5**	3.666	1.870	-0.251	1.830	0.034	100.0	0.0
	3000	**94.5**	2.830	1.444	-0.059	1.484	0.022	100.0	0.0	**94.5**	2.600	1.327	-0.250	1.299	0.017	100.0	0.0
A9	750	**96.0**	4.074	2.079	-0.090	1.957	0.038	83.5	0.3	**95.0**	4.218	2.152	-0.164	2.100	0.044	79.0	0.5
(0.645)	1500	**96.0**	2.878	1.469	-0.147	1.445	0.021	86.0	1.6	**95.5**	2.991	1.526	-0.028	1.498	0.022	79.2	3.6
	3000	**93.0**	2.049	1.045	-0.404	1.075	0.013	86.9	0.3	**95.0**	2.120	1.082	-0.221	1.122	0.013	73.5	2.2
A10	750	**95.0**	5.439	2.775	0.089	2.960	0.087	86.4	2.9	**93.5**	5.462	2.787	0.304	2.849	0.082	83.5	2.6
(0.315)	1500	**95.0**	3.869	1.974	-0.196	1.886	0.036	94.5	4.9	**93.5**	3.871	1.975	0.507	1.995	0.042	67.0	4.8
	3000	**93.5**	2.749	1.403	-0.087	1.462	0.021	95.0	4.3	**95.5**	2.742	1.399	0.205	1.316	0.018	62.4	3.3
A11	750	**92.5**	2.015	1.028	0.579	1.131	0.016	96.4	1.6	**95.0**	1.784	0.910	0.459	0.887	0.010	94.3	1.2
(0.015)	1500	**94.5**	1.428	0.729	0.334	0.764	0.007	96.8	1.4	**95.0**	1.278	0.652	0.314	0.706	0.006	94.5	0.3
	3000	**94.5**	0.996	0.508	0.193	0.500	0.003	98.4	0.9	**95.0**	0.908	0.463	0.218	0.441	0.002	95.0	0.1
A12	750	**95.5**	1.057	0.539	0.533	0.613	0.007	73.4	3.0	**93.5**	1.167	0.596	0.542	0.576	0.006	65.2	2.7
(0.003)	1500	**93.0**	0.690	0.352	0.301	0.374	0.002	99.7	5.4	**93.5**	0.746	0.381	0.258	0.380	0.002	67.5	4.4
	3000	**97.5**	0.464	0.237	0.140	0.247	0.001	98.6	5.1	**96.5**	0.492	0.251	0.080	0.261	0.001	60.2	3.4

N refers to the sample size. **CP** refers to coverage probability based on 200 replications. **Width** refers to half the width of the 95% confidence interval. **SE** refers to the average asymptotic standard error. **SD** refers to the empirical standard deviation of replicated estimations. MSE refers to mean squared error. **TP** refers to the average true positive rate. **FP** refers to the average false positive rate. True value of R^2_{Med} is listed within the parentheses.

Aim 1: interval estimation for R^2-based mediation measure
Application to Framingham Heart Study (FHS)

- Canonical correlation analysis (CCA)[1]
  - Genes selected in different subsamples from CF-OLS
  - Genes selected by different variable selection methods
- Computational time
  - Proposed method: 4.67 hrs using a single core
  - Previous method: 75.99 hrs using 25 cores in parallel

1. H. Harold. Relations between two sets of variates. Biometrika, 28(3/4):321, 1936.

Aim 2: Meta-analysis of R^2-based mediation effect
Method

- Inverse-variance estimator: $w_q = 1/S_q$, $\hat{\theta_{IW}} = \sum_{q=1}^{Q} \hat{\theta}_q w_q / \sum_{q=1}^{Q} w_q$

- DerSimonian and Laird (DL) estimator[1]: $w_k^* = 1/(Var_k + \tau^2)$, and $\hat{\theta_{DL}} = \sum_{q=1}^{Q} \hat{\theta}_q w_q^* / \sum_{q=1}^{Q} w_q^*$,
$$\hat{\tau}^2 = \max\left\{ \frac{\sum_{q=1}^{Q} S_q^{-1}(\hat{\theta}_q^2 - \hat{\theta}_{IW}^2) - (Q-1)}{\sum_{q=1}^{Q} S_q^{-1} - \sum_{q=1}^{Q} S_q^{-2} / \sum_{q=1}^{Q} S_q^{-1}}, 0 \right\}$$, where $\widehat{S_q}$ denotes the estimated variance of $\widehat{\theta_q^2}$, Q denotes the number of the studies.

- Other estimators

1. DerSimonian, Rebecca, and Nan Laird. 1986. “Meta-Analysis in Clinical Trials.” *Controlled Clinical Trials* 7 (3): 177–88.