Mendelian Randomization Contextualized Model

Neha Talluri

September 8th - October 3rd 2025

1 Introduction

Classical statistical and machine learning models typically assume that a single global set of parameters is sufficient to describe the relationship between inputs and outputs across all individuals or conditions.

1.1 What is Contextualized Modeling?

Contextualized modeling is a machine learning idea that generates models tailored to a specific context rather than a single model for an entire dataset or population.

This approach is important in fields with high data heterogeneity, where a single model may miss localized effects and nuances inherent to different subpopulations or conditions. Instead of forcing all data into a uniform model, contextualized modeling adapts to the specific conditions under which a phenomenon is observed. Importantly, the context encoder does not operate in isolation for each individual. It is trained across all available data and it learns how different contextual variables influence the relationship between inputs and outputs. This means that information from one context can inform predictions in another context, allowing the model to generalize and borrow statistical strength across subpopulations.

The core of this approach relies on a context encoder. It is a reusable component that translates contextual information into parameters to create a sample-specific model whose behavior is dictated by the output of the context encoder, allowing for locally optimized predictions.

The context encoder is the component responsible for transforming observed contextual variables into model parameters. Given a sample's context C_i (age, sex, or any other variables that capture heterogeneity), the encoder produces parameter values that are specific to that sample. In practice, the encoder can be a simple linear function, a multilayer perceptron (MLP), or even a more complex architecture such as a transformer.

Formally, if the base model predicts an outcome Y_i from an input X_i, the context encoder modifies the model's parameters according to the context:

$$Y_i = f_{\theta(C_i)}(X_i)$$

where $\theta(C_i)$ are the parameters generated by the encoder given context C_i.

In other words, the encoder does not directly predict outcomes itself, but rather maps context into the coefficients that the prediction model will use. This allows each sample to effectively have its own "local" model, while still sharing a common global structure via the encoder [1].

1.2 What is Mendelian Randomization?

Mendelian Randomization (MR) is a method that uses genetic variants G to investigate the causal effect of an exposure X on an outcome Y.

The basic idea relies on the random allocation of alleles at conception, which provides a source of variation in the exposure that is less susceptible to other confounding factors.

A valid MR study relies on four key assumptions, often referred to as the instrumental variable (IV) assumptions:

- The relevance assumption: The genetic variant (instrumental variable) must be reliably associated with the exposure of interest.
- The independence assumption: The genetic variant must be independent of any confounding factors that influence the exposure-outcome relationship.
- The exclusion restriction assumption: The genetic variant must only affect the outcome through its association with the exposure. There should be no other pathways, known as pleiotropy, through which the genetic variant affects the outcome.
- The point-estimate-identifying condition: The causal effect of the exposure on the outcome is assumed to be homogeneous across individuals

Under these assumptions, MR estimates the causal effect β of X on Y by leveraging the variation in X explained by G. In practice, this is most commonly achieved using the two-stage least squares (2SLS) estimator:

$$\text{First stage: } X = \pi_0 + G\pi + v_X,$$
$$\text{Second stage: } Y = \alpha + \beta\hat{X} + u$$

where the first stage uses genetic instruments G to predict the exposure X, and the second stage regresses the outcome Y on the predicted exposure $\hat{X}$ [2].

Within a two-stage Mendelian Randomization framework, the context encoder can be inserted into either or both stages:

1. First stage (exposure model): Encodes context into parameters for the regression of exposure X on instruments G.
2. Second stage (outcome model): Encodes context into parameters for the regression of outcome Y on the predicted exposure $\hat{X}$.

By doing so, the model adapts to context-specific relationships at each stage, rather than assuming a single global mapping across all individuals.

2 Data

I simulated individual level genotype and phenotype (context) data to mirror the structure of a genome-wide association study (GWAS) cohort. The dataset consists of n individuals and L single nucleotide polymorphisms (SNPs).

2.1 Genotypes

For each SNP $j \in \{1, \dots, L\}$ I draw from a minor allele frequency (MAF)

$$m_j \sim \mathcal{U}(0.05, 0.5)$$

and simulate

$$G_{ij} \sim \text{Binomial}(2, m_j),$$

so $G_{ij} \in \{0, 1, 2\}$ counts minor-allele copies for individual i at SNP j.

2.2 Unobserved confounding

To cause X and Y to be spuriously correlated, I draw $U \sim \mathcal{N}(0, I_n)$, mirroring real-world GWAS where confounders are unobserved.

2.3 Data generating process

The exposure and outcome are generated as:

$$X = G\pi + U + \varepsilon_X, \qquad \varepsilon_X \sim \mathcal{N}(0, \sigma_X^2 I_n)$$

$$Y = \beta X + U + \varepsilon_Y, \qquad \varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2 I_n)$$

where instrument–exposure (strength of each instrument) effects are

$$\pi_j \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \texttt{pi\_scale})\,.$$

This formulation captures the structure required for a MR setting; the instruments G induce variation in the exposure X, the exposure in turn causally influences the outcome Y through the parameter β, and confounding is explicitly modeled through U. By simulating data in this way, I create a controlled environment where the true causal effect β is known, which is necessary for evaluating how well different models recover the correct effect under varying conditions.

2.4 Context

Optionally, observed contextual variables such as age, sex, BMI, smoking status, height, weight, systolic blood pressure, cholesterol, glucose, and hours of sleep can be included. In the data-generating process, each of these variables is assigned a coefficient, where γ and ζ represent the coefficients for the contextual variables C in the exposure and outcome models, respectively, capturing the influence of context on each stage.

When `use_context=True`, the exposure and outcome update as so:

$$X = G\pi + C\gamma + U + \varepsilon_X,$$
$$Y = \beta X + +C\zeta + U + \varepsilon_Y,$$

2.5 Parameters and defaults for Data

Unless changed:

- $n = 5000$ (number of individuals)
- $L = 15$ (number of instruments)
- $\beta = 0.30$ (correlation between X and Y)
- `pi_scale`$= 0.10$ (instrument strength)
- $\sigma_X = \sigma_Y = 1.0$ (noise in exposure and outcome)
- seed = 7
- γ: effects of context on exposure X, given by

  $$\gamma_c = \{0.03,\ 0.05,\ 0.14,\ 0.16,\ 0.00,\ 0.00,\ 0.08,\ 0.02,\ 0.12,\ -0.06\}$$

  corresponding to {age, sex, BMI, smoker, height, weight, systolic blood pressure, cholesterol, glucose, sleep}
- ζ: effects of context on outcome Y, given by

  $$\zeta_c = \{0.02,\ 0.03,\ 0.10,\ 0.05,\ 0.00,\ 0.00,\ 0.12,\ 0.15,\ 0.18,\ 0.04\}$$

  corresponding to the same contextual variables
- `use_context = False`

2.6 Interpretation of β

In simulated data, β is fixed by design ($\beta = 0.30$ by default), which represents the true causal effect of the exposure on the outcome. The purpose of fitting models to this simulated dataset is to evaluate whether the models can recover the known ground-truth effect.

In contrast, when analyzing real data, β is unknown and is precisely what the MR models aim to estimate. Here, β represents the causal effect of the exposure on the outcome, under the instrumental variable assumptions. The magnitude of β describes the strength of the causal relationship, while the sign indicates its direction (positive: the exposure increases the outcome; negative: the exposure decreases the outcome). Unlike in simulation, where correctness is judged by closeness to the true β, in real analyses the validity of $\hat{\beta}$ depends on the plausibility of the IV assumptions and sensitivity analyses.

2.7 Delivered dataset

The simulator returns a DataFrame with n rows (individuals) and the following columns:

- `X`: simulated exposure,
- `Y`: simulated outcome,
- `age`, `sex`, `BMI`, `smoking sta- tus`, `height`,`weight`, `systolic blood pressure`, `cholesterol`, `glucose`, and `hours of sleep`: observed context,
- `G0`,...,`G{L-1}`: SNP (0/1/2).

Additional data is also returned as a metadata object with the ground truth (β) π, U, MAFs, and L and n.

2.8 Data Simulation Procedure For Experiments

To generate the datasets used in my experiments, I systematically varied three key parameters: the number of individuals (N), the number of instruments (L), and whether contextual variables are included. All other parameters were fixed with their default values.

1. Number of individuals (N): I simulated datasets with different sample sizes to test scalability and robustness. The list of sample sizes was: $N \in \{1000, 3000, 5000, 10000, 20000\}$.
2. Number of instruments (L): To examine how the number of available instruments influenced model performance, I varied L across: $L \in \{5, 10, 15, 30, 50\}$.
3. Contextual information: Each dataset was generated both with and without contextual variables included in the predictors X and outcomes Y, corresponding to: Context $\in \{\text{True}, \text{False}\}$.

By taking the Cartesian product of these settings, I got a complete grid of datasets spanning all combinations of N, L, and context inclusion, which were then used to train and evaluate the models.

3 Baseline Models

Let $Y \in R^n$ be the outcome, $X \in R^n$ the exposure, $G \in R^{n \times L}$ the SNP instruments. An intercept is included in all regressions.

3.1 Ordinary Least Squares (OLS)

I regress Y directly on X:

$$Y = \alpha + \beta_{\text{OLS}} X + \varepsilon$$

estimating $\hat{\beta}_{\text{OLS}}$ by least squares. This serves as a benchmark when U induces confounding.

3.2 Two-Stage OLS

3.2.1 Stage 1 (first OLS):

Regress X on instruments:

$$\hat{X} = \alpha_1 + G\hat{\pi} + v$$

3.2.2 Stage 2 (second OLS):

Regress the outcome on the fitted exposure $\hat{X}$:

$$Y = \alpha_2 + \beta\hat{X} + u$$

3.3 Two-Stage OLS using LASSO

To mitigate many weak instruments, I first select a subset of SNPs via LASSO in Stage 1, then perform the same two-stage OLS using only the selected SNPs.

3.3.1 Stage 1 (LASSO)

Select relevant SNP instruments using LASSO:

$$\hat{\pi}^{\text{LASSO}} = \arg\min_{\pi,\,\alpha} \frac{1}{2n} \left\| X - \alpha\mathbf{1} - G\pi \right\|_2^2 + \lambda\|\pi\|_1,$$

with λ chosen by cross-validation. The fitted exposure is

$$\hat{X} = \alpha_1 + G\hat{\pi}^{\text{LASSO}} + v$$

3.3.2 Stage 2 (OLS)

Regress the outcome on the fitted exposure $\hat{X}$:

$$Y = \alpha_2 + \beta\,\hat{X} + u$$

4 Contextualized Models

Let $C \in R^{n \times p}$ denote the context matrix with p contextual covariates (e.g., age, sex, BMI). I extend the baseline models by allowing the effect of X on Y to vary with context C, using a Contextualized Regressor.

4.1 Contextualized Linear Regression

For each second stage of the models, I apply a contextualized linear regression model.

For each sample i, the contextualized linear regression model is

$$Y_i = f_\mu(C_i) + f_\beta(C_i)\, X_i + \varepsilon_i,$$

where,

- C_i are the contextual variables for sample i,
- $f_\mu(C_i)$ is the context encoder for the offset (context-specific intercept),
- $f_\beta(C_i)$ is the context encoder for the linear coefficient (context-specific slope),
- ε_i is the residual error.

I implement this contextulized model using the `ContextualizedRegressor` from the `contextualized` package [3].

This context encoder a multilayer perceptron (MLP), which maps the context variables C_i into sample-specific parameters. Thus, the regression coefficients and intercept are learned as functions of the context rather than as global constants. I experimented with two different encoders. The first uses the default parameters set by the `ContextualizedRegressor` for an MLP. The second set an MLP encoder specified by width=8 and depth=1, corresponding to a small MLP.

In this work, contextualization is applied only to the second stage. The first stage involves regressing the exposure on a potentially large number of genetic instruments, which introduces many more parameters and makes contextualization computationally demanding and potentially unstable. By contrast, the second stage has a simpler structure with fewer parameters to estimate, making it a more practical starting point for evaluating the benefits of contextulized MR.

4.2 One-Stage Contextualized Regression

I regress Y directly on X with context C, modeling the exposure effect $\beta(C)$ as a context-dependent parameter:

$$Y = f_\mu(C) + f_\beta(C)X + \varepsilon,$$

where $\beta(C)$ is learned as a function of the contextual variables. The fitted parameters are the contextualized exposure effects $f_\beta(C)$ and intercepts $f_\mu(C)$ for each sample. This serves as a contextualized benchmark when U induces confounding.

4.3 Two-Stage Contextualized Regression

4.3.1 Stage 1 (First-Stage OLS):

Regress exposure on instruments to obtain a fitted exposure:

$$\hat{X} = \alpha_1 + G\hat{\pi} + v.$$

4.3.2 Stage 2 (Contextualized Regression):

Regress outcome on the fitted exposure $\hat{X}$ with context C:

$$Y = f_\mu(C) + f_\beta(C)\hat{X} + \varepsilon,$$

where $f_\beta(C)$ and $f_\mu(C)$ varies with C and is estimated via the contextualized linear regressor.

4.4 Two-Stage Contextualized Regression with LASSO Instruments

4.4.1 Stage 1 (LASSO)

Select relevant SNP instruments by LASSO:

$$\hat{\pi}^{\text{LASSO}} = \arg\min_{\pi,\,\alpha} \frac{1}{2n} \|X - \alpha\mathbf{1} - G\pi\|_2^2 + \lambda\|\pi\|_1,$$

with λ chosen by cross-validation. The fitted exposure is

$$\hat{X} = \hat{\alpha} + G\hat{\pi}^{\text{LASSO}} + v.$$

4.4.2 Stage 2 (Contextualized Regression)

Regress outcome on the fitted exposure $\hat{X}$ with context C:

$$Y = f_\mu(C) + f_\beta(C)\hat{X} + \varepsilon,$$

where $f_\beta(C)$ and $f_\mu(C)$ varies with C and is estimated via the contextualized linear regressor.

5 Results

5.1 MSE

Add why I am using the MSE

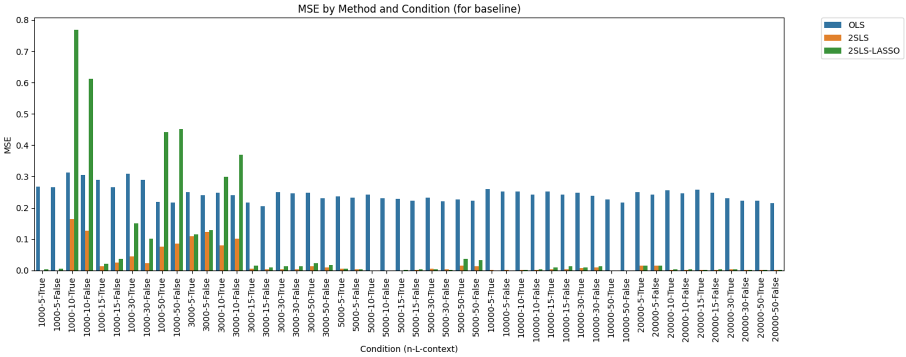
5.2 Baseline performance

Figure 5.2 illustrates the performance of the baseline models (OLS, 2SLS, and 2SLS-LASSO) across different sample sizes (n), numbers of instruments (L), and context settings. The OLS model consistently produces a high mean squared error (MSE) because it directly regresses the exposure X on the outcome Y. As a result, unmeasured confounders that influence both X and Y introduce bias; the unmeasured confounders added to the simulated data creates a spurious correlation between the two variables. This bias remains even as the sample size or number of instruments increases. In the context of Mendelian Randomization, this underscores the limitation of a naïve regression from exposure (X) to outcome (Y); without incorporating genetic instruments, the estimated causal effect remains biased by confounders.

2SLS addresses confounding through a two-stage procedure; it first regresses the exposure X on the genetic instruments (G) to obtain the predicted exposure $\hat{X}$, and then regresses the outcome Y on $\hat{X}$. By isolating the variation in X explained by G, this breaks the link between unmeasured confounders and the exposure, yielding an less biased estimate of the causal effect (β). As a result, 2SLS achieves much lower MSE overall, especially as sample size grows, although it exhibits higher MSEs in small samples sizes (1000 - 3000 n).

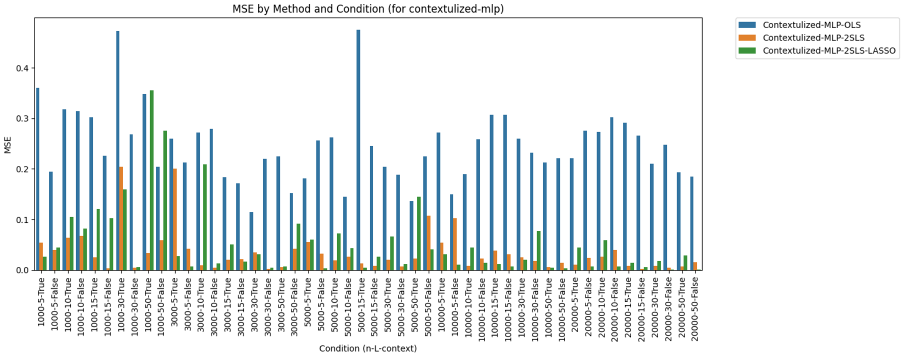
In contrast, 2SLS-LASSO modifies the first stage by applying LASSO regression to select a subset of relevant SNP instruments before constructing $\hat{X}$. This helps mitigate the problem of many weak instruments, but at small n the regularization could exclude important instruments and produce large spikes in MSE. With larger samples, however, the method stabilizes.

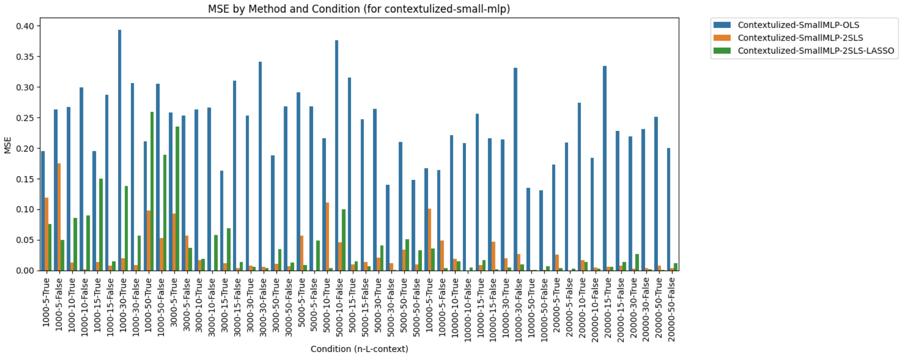
Taken together, OLS is systematically biased by confounding, 2SLS is unbiased but noisy in small samples, and 2SLS-LASSO is fragile without sufficient data.



5.3 Contextualized Models

Figures 5.3 and 5.3 compare the contextualized models to their baseline counterparts on the same datasets. Contextualized OLS shows little to no improvement and, in some cases, performs slightly worse. This indicates that adding contextualization cannot correct the bias introduced by confounding and may add extra variance. For 2SLS, the contextualized models behave almost identically to the baseline, with only minor differences across sample sizes. The clearest improvement comes from 2SLS-LASSO. At small n, contextualization reduces the large error spikes observed in the baseline, though as n grows the performance of contextualized and baseline 2SLS-LASSO becomes similar.





5.4 Comparing Between Model Types

Figure 5.4 shows the 2 stage OLS models (where the first stage uses OLS and the second stage is either a contextualized linear regressor or OLS (baseline)). At small sample sizes, all models show some instability, with higher variance in MSE, but the differences are relatively minor. As the sample size increases, there is little improvement in overall MSE; the contextualized models closely follow the performance of standard 2SLS, with minor reductions in error at larger n. This suggests that contextualization in the second stage provides only limited additional benefit.

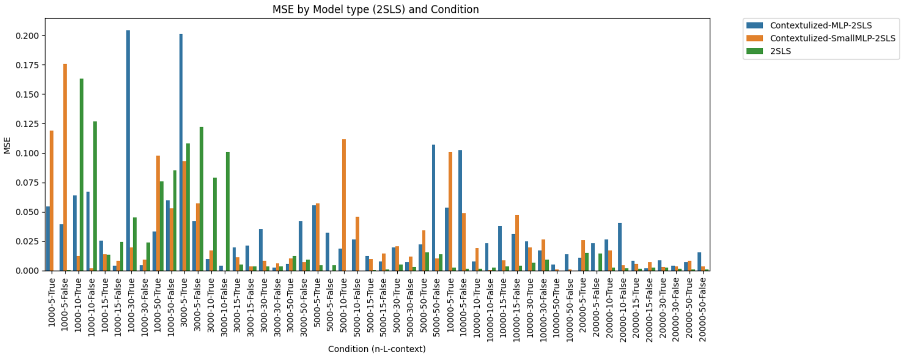


Figure 5.4 highlights the 2-stage LASSO models (where the first stage uses LASSO for instrument selection and the second stage is either a contextualized linear regressor or OLS (baseline)). In the baseline, 2SLS-LASSO is highly unstable at small n, with large spikes in error due to over-regularization. Adding contextualization in the second stage substantially stabilizes the estimator, producing consistently lower MSE across conditions. As sample size increases, the contextualized models follow the performance of 2SLS-LASSO, with only small reductions in error. Overall, contextualization makes this otherwise fragile approach far more reliable, particularly in small-sample settings.

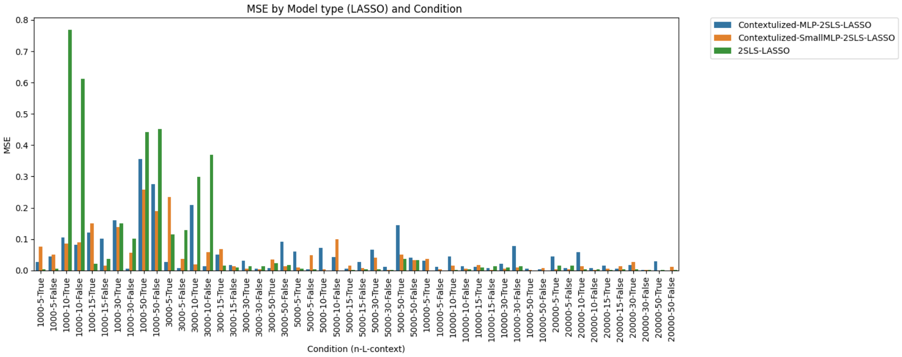
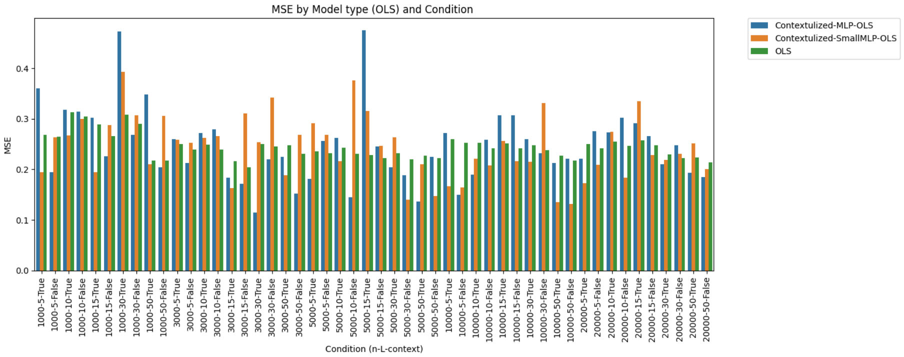


Figure 5.4 illustrates the OLS models (where the first stage is OLS or a contextualized linear regressor). Here, contextualization provides little benefit; MSE remains high across all sample sizes, closely resembling the baseline OLS performance. This reinforces that contextualization cannot overcome the fundamental bias induced by unmeasured confounding.



5.5 Comparison across models and datasets

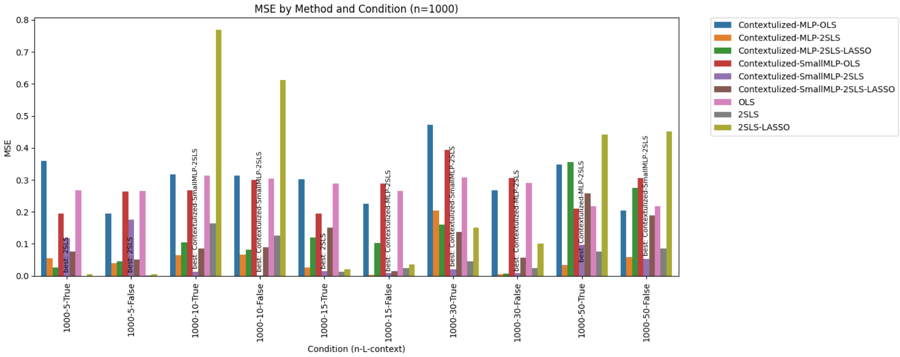
Figures 5.5–5.5 compare all models (baseline and contextualized) on the same datasets, highlighting which approach performs best under each condition. For every dataset, the lowest MSE model is marked, allowing us to see not just overall trends but also which estimator is "best" in specific scenarios.

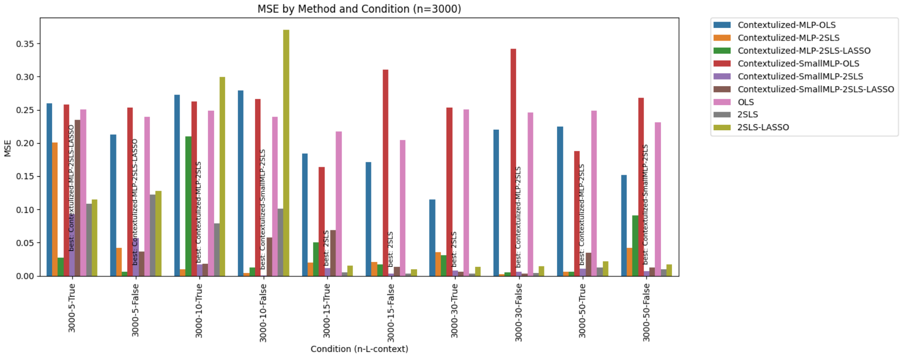
A consistent result across all sample sizes is that the single stage models are never the best performing model. Its structural bias from confounding means that, even with contextualization, it cannot match the performance of methods that leverage instruments.

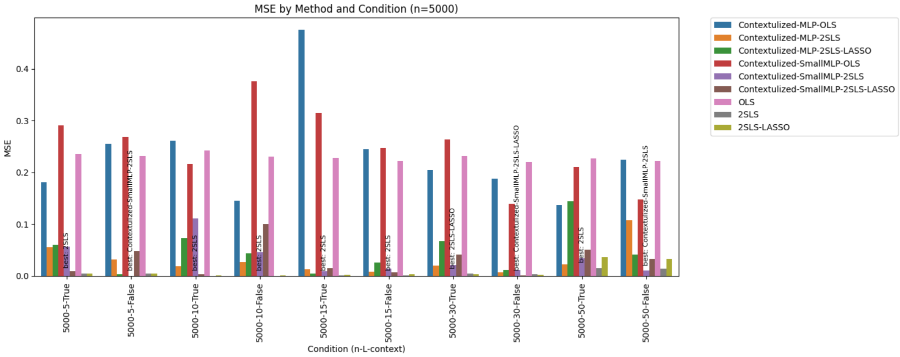
Sample-size dependent patterns

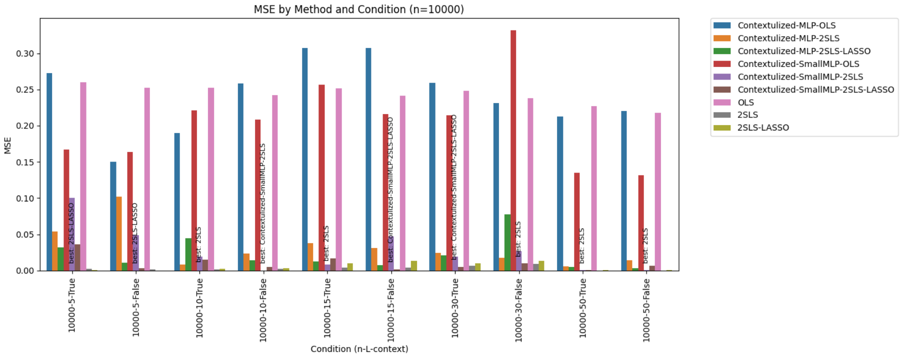
- At small n (e.g., $n = 1000$), contextualized 2SLS or 2SLS often emerges as the best choice.
- At moderate n (3000–5000), performance is mixed. Sometimes the contextualized 2-stage models perform best, but in other cases the 2-stage baseline models match or outperform them. They all shared similar performance. The benefit of contextualization in the second stage depends on the dataset.
- At large n (10000–20000), the differences narrow even further. Baseline 2SLS or 2SLS-LASSO can be just as competitive as their contextualized counterparts, and there is no consistent winner. Again, whether contextualization in the second stage provides an advantage ultimately depends on the dataset.

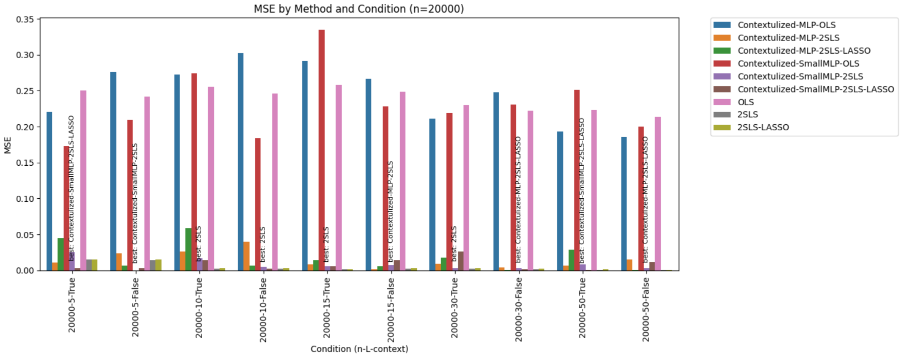
These results highlight that the "best" model depends on both the dataset and the sample size. Contextualization is most helpful when data are limited or when the estimator is fragile. However, because in the setup only the second stage is contextualized, these models are not always the best choice; in many cases, the baseline models with their global interpretation just as well, or nearly as well, as their contextualized counterparts. This means that while contextualization offers flexibility and context-specific parameter estimates in the second stage, the added complexity is not always rewarded with better estimation.











6 Discussion

This work provides an exploration into contextualized Mendelian randomization (MR) modeling. While the results show that contextualization in the second stage can capture context specific heterogeneity, the improvements are not always consistent across datasets. This suggests that contextualization only at the second stage for MR may be insufficient. A natural extension is to also apply contextualization at the first stage, where the exposure is regressed on the instruments. This would allow the model to account for context dependent instrument exposure relationships rather than assuming a single global structure for the first stage.

Another important direction concerns the choice of context encoders. In this study, only a limited amount of encoders were considered. Testing different context encoders systematically may reveal whether specific architectures are particularly suited for MR tasks.

A limitation of contextualized MR is the limited availability of real individual-level data that also include contextual variables. While individual-level data are more informative for modeling, they are often difficult to obtain due to privacy restrictions and cost barriers. In contrast, summary-level data are more widely available and freely shared, making them attractive for large-scale applications. Developing methods that extend contextualized MR to work directly with summary statistics would therefore increase its applicability. However freely available summary data generally lack contextual variables, which limits their ability to support contextualized modeling. An alternative path is to identify free or institutionally available individual-level datasets that include contextual information, which would provide valuable opportunities to test contextualized models at scale.

Overall, these directions highlight that contextualized MR has promise, but realizing its potential requires both methodological extensions and practical adaptations.

7 Code Availability

All code used for the simulations and analyses in this study is publicly available at: `https://github.com/ntalluri/contextualized-mr`

References

[1] B. Lengerich, C. N. Ellington, A. Rubbi, M. Kellis, and E. P. Xing, “Contextualized machine learning,” 2023.

[2] E. Sanderson, M. M. Glymour, M. V. Holmes, H. Kang, J. Morrison, M. R. Munaf, T. Palmer, C. M. Schooling, C. Wallace, Q. Zhao, and G. Davey Smith, “Mendelian randomization,” *Nature Reviews Methods Primers*, vol. 2, p. 6, Feb 2022. Published online 10 February 2022.

[3] C. N. Ellington, B. J. Lengerich, W. Lo, A. Alvarez, A. Rubbi, M. Kellis, and E. P. Xing, “Contextualized: Heterogeneous modeling toolbox,” *Journal of Open Source Software*, vol. 9, no. 97, p. 6469, 2024.