Estimation of Effects of Endogenous Time-Varying Covariates: A Comparison Of Multilevel Linear Modeling and Generalized Estimating Equations

Research Report

Ward B. Eiling (9294163)

Supervisors: Ellen Hamaker and Jeroen Mulder

Master's degree in Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

December 17, 2024

Word count: 3894/2500

FETC-approved: 24-2003

Candidate journal: Psychological Methods

1 Introduction

Across a wide range of disciplines, researchers analyze clustered longitudinal, observational data to investigate prospective causal relationships between variables. When analyzing such data, psychological researchers most commonly use the multilevel linear model[1] (MLM, Bauer & Sterba, 2011), which—in the context of longitudinal data analysis—partitions observed variance into stable between-person differences and within-person fluctuations (Hamaker & Muthén, 2020). In the application of the MLM, time invariant and time-varying covariates, the latter measured repeatedly over time, are often available. The inclusion of covariates is a common strategy to improve parameter precision (Boruvka et al., 2018) and address bias introduced by (time-varying) confounders (Daniel et al., 2013; Wodtke, 2020). Nevertheless, it is well-known that this approach is not universally beneficial, as conditioning on variables like colliders within the causal pathway can distort treatment effect estimates (Elwert & Winship, 2014).

Dating back to the work of Pepe & Anderson (1994), it has also been well-established that including endogenous time-varying covariates—those directly or indirectly influenced by prior exposure/treatment or outcome—in longitudinal studies without treatment can result in biased estimates of treatment effects. Despite its significance, this issue has received little attention in psychological research. Building on this foundation, a recent paper by Qian et al. (2020) examined the suitability of MLM for estimating the causal effect of a time-varying exposure or treatment. Specifically, they focused on settings where the exposure is randomly assigned at each occasion within individuals. Such randomized exposures may include, for example, prompts delivered through push notifications to remind participants of cognitive or mindfulness-based strategies (Nahum-Shani et al., 2021; Walton et al., 2018). While random assignment with a constant probability might seem sufficient to identify (the presence and absence of) causal effects, Qian et al. (2020) showed that model fitting issues and parameter bias can arise when

[1]The MLM is known by various names in different substantive fields, including: linear mixed model, hierarchical linear model, random-effect model and mixed-effects model.

a *time-varying endogenous covariate* is present.

However, due to a divide between the disciplines that employ the MLM, such critiques appear to have largely failed to reach the applied researcher in psychology. One specific reason might be that the technical jargon in other disciplines makes it difficult for researchers to recognize when and how these issues emerge. Therefore, this report aims to understand why Qian et al. (2020) found biased estimates of the treatment effect for some generative models containing endogenous covariates and not for others; and to explain this issue to an audience of psychologists. To achieve this aim, the study will first use graphical diagrams to evaluate relevant criteria, ensuring accessibility and rigor through a reliance on graphical rules rather than algebra. Next, data simulations, based on the original scenarios of Qian et al. (2020) as well as additional scenarios, will be conducted to pinpoint the issue and to assess whether these criteria effectively identify bias. Accordingly, the following research question will be addressed: *When does the inclusion of endogenous variables in multilevel linear models result in biased estimates of the treatment effect?*

2 Methods

To obtain a better understanding of the issue exposed by Qian et al. (2020), two methods were employed. First, graphical methods were used provide insight into the presence and extent of bias with potential violation of criteria: (a) path diagrams were used to evaluate the conditional independence assumption (Qian et al., 2020) and (b) directed acyclic graphs (DAGs) were used to evaluate the backdoor criterion (Pearl, 1988, 2009). Second, a simulation study was performed to reproduce the results for the generative models (GMs) from Qian et al. (2020) and to further isolate the issue using additional GMs. In this simulation, bias in the treatment effect was quantified using analytical multilevel models identical to the generative models. In this section, the GMs will be formulated and specifications will be defined for the simulation study.

2.1 Data Generation

We consider 2 GMs from Qian et al. (2020), one (GM A) being a special case of the general model (GM G) where bias was detected. To further isolate the source of bias, we introduce two additional special cases, labeled GM B and C. Table 1 summarizes the differences between the generative models. Compared to the general model G, GM A is not directly determined by the random intercept b_{i0}; GM B is does not have a random slope b_{i2} for treatment; and GM C does not have a fixed interaction effect β_1 between covariate and treatment.

Table 1: Generative Models: Summary of Differences

Generative Model	Name in Qian et al. (2020)	dependency b_{i0} and X_{it}	random slope treatment b_{i2}	interaction β_1
G(eneral)	3	✓	✓	✓
A	1	×	✓	✓
B	NA	✓	×	✓
C	NA	✓	✓	×

The details of the generative models are described below. We follow the symbol notation of Qian et al. (2020) to allow for direct comparison, but rewrite the equations into within- and between-person models (see Raudenbush & Bryk, 2002; Schoot, 2017).

2.1.1 Generative Model G

Following the original notation of Qian et al. (2020), the outcome of GM G was generated according to the following model:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + b_{i0} + A_{it}(\beta_0 + \beta_1 X_{it} + b_{i2}) + \epsilon_{it+1}$$

where Y_{it+1} is the outcome at time $t+1$, X_{it} is the covariate at time t, A_{it} is the treatment

at time t, b_{i0} is the random intercept, b_{i2} is the random slope for the treatment, and ϵ_{it+1} is the error term. We may rewrite this model into the repeated-observations or within-person model in the following steps:

$$\begin{aligned}
Y_{it+1} &= \alpha_0 + \alpha_1 X_{it} + b_{i0} + A_{it}(\beta_0 + \beta_1 X_{it} + b_{i2}) + \epsilon_{it+1} \\
&= \alpha_0 + \alpha_1 X_{it} + b_{i0} + \beta_0 A_{it} + \beta_1 A_{it} X_{it} + A_{it} b_{i2} + \epsilon_{it+1} \\
&= \alpha_0 + b_{i0} + \alpha_1 X_{it} + \beta_0 A_{it} + A_{it} b_{i2} + \beta_1 A_{it} X_{it} + \epsilon_{it+1} \\
&= (\alpha_0 + b_{i0}) + \alpha_1 X_{it} + (\beta_0 + b_{i2}) A_{it} + \beta_1 A_{it} X_{it} + \epsilon_{it+1} \\
&= \pi_{0i} + \pi_{1i} X_{it} + \pi_{2i} A_{it} + \pi_{3i} A_{it} X_{it} + \epsilon_{it+1}.
\end{aligned}$$

with the person-level or between-person model (level 2):

$$\begin{aligned}
\pi_{0i} &= \alpha_0 + b_{i0}, \quad \text{where} \quad b_{i0} \sim \mathcal{N}(0, \sigma^2_{b0}), \\
\pi_{1i} &= \alpha_1, \\
\pi_{2i} &= \beta_0 + b_{i2}, \quad \text{where} \quad b_{i2} \sim \mathcal{N}(0, \sigma^2_{b2}), \\
\pi_{3i} &= \beta_1.
\end{aligned}$$

We model fixed effects α_0, α_1, β_0, and β_1 as constants across individuals, while random effects b_{i0} and b_{i2} capture individual-specific deviations. The presence of the interaction term β_1 implies treatment heterogeneity: the effect of the treatment A_{it} on the outcome depends on the value of the covariate X_{it}. b_{i0} represents deviations from the population intercept α_0, and b_{i2} represents deviations from the population slope β_0. A higher b_{i0} indicates a higher initial outcome, while a higher b_{i2} indicates a stronger treatment effect. Following Qian et al. (2020), the random effects b_{i0} and b_{i2} are modeled independent of each other.

The covariate is generated as:

$$X_{it} = \begin{cases} b_{i0} + \epsilon_{X_{it}}, & \text{if } t = 1, \\ b_{i0} + Y_{it} + \epsilon_{X_{it}}, & \text{if } t \geq 2, \end{cases} \quad \text{where} \quad \epsilon_{X_{it}} \sim \mathcal{N}(0, 1)$$

The randomization probability of treatment $p_t = P(A_{it} = 1 \mid H_{it})$ is constant at 1/2. Thus, $A_{it} \sim$ Bernoulli(0.5) for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. In other words, for every given person i and every timepoint t, the probability that treatment is assigned is equivalent to a fair coinflip. The exogenous noise is $\epsilon_{it+1} \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

Figure 1a shows the path diagram for the first couple observations of GM G.

2.1.2 Generative Model A

GM A is a special case of GM G, where the effect of the random intercept b_{i0} on the covariate X_{it} is set to zero. This results in a model where the covariate X_{it} is not directly determined by the random intercept b_{i0} (see Figure 1b). Instead, the endogenous covariate X_{it} equals the previous outcome Y_{it} plus some random noise:

$$X_{it} = \begin{cases} \epsilon_{X_{it}}, & \text{if } t = 1, \\ Y_{it} + \epsilon_{X_{it}}, & \text{if } t \geq 2, \end{cases} \quad \text{where} \quad \epsilon_{X_{it}} \sim \mathcal{N}(0, 1)$$

2.1.3 Generative Model B

GM B is a special case of GM G, where the variance for the random slope b_{i2} is set to zero: $\sigma_{b2}^2 = 0$. Consequently, the random slope b_{i2} for the treatment A_{it} is removed (see Figure 1c). While the within-person model is the same as GM G, there is a slight alteration in the between-person model:

$$\pi_{2i} = \beta_0.$$

The single equation model then becomes:

$$Y_{it+1} = (\alpha_0 + b_{i0}) + \alpha_1 X_{it} + \beta_0 A_{it} + \beta_1 A_{it} X_{it} + \epsilon_{it+1}$$

2.1.4 Generative Model C

GM C is a special case of GM G, where the fixed interaction parameter $\beta_1 = 0$, which implies the removal of the interaction term $\beta_1 A_{it} X_{it}$ (see Figure 1d). This, in turn, removed π_{3i}, thereby creating a discrepancy in within-person model of GM C and GM G:

$$Y_{it+1} = \pi_{0i} + \pi_{1i} X_{it} + \pi_{2i} A_{it} + \epsilon_{it+1}.$$

Nevertheless, the between-person model of π_{0i}, π_{1i} and π_{2i} remains the same as GM G. The single equation model then becomes:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + b_{i0} + A_{it}(\beta_0 + b_{i2}) + \epsilon_{it+1}.$$

2.1.5 Parameter Values

The following parameter values were adapted from Qian et al. (2020):

$$\alpha_0 = -2, \quad \alpha_1 = -0.3, \quad \beta_0 = 1, \quad \beta_1 = 0.3,$$

$$\sigma^2_{b0} = 4, \quad \sigma^2_{b2} = 1, \quad \sigma^2_{\epsilon} = 1.$$

2.2 Data Analysis

In the simulation study, we evaluated the performance of the models across a total of 24 different settings, each replicated 1,000 times, by systematically varying the following factors:

- **Generative Models (GM):** G, A, B, C

- **Number of timepoints (T):** 10, 30
- **Sample size (N):** 30, 100, 200

All data generation and estimation was performed in `R`, version 4.4.2 (Team, 2024). After the generation of data generation for any given setting, analytical multilevel linear models were fit that are equivalent to each of the respective data-generating models. To fit the standard MLM, the `lmer` function from the R-package `lme4` (Bates et al., 2015) was employed with restricted maximum likelihood estimation.

3 Results

3.1 Conditional Independence and Path Diagrams

The first criterion for evaluating the presence of bias in treatment effect estimates is the *conditional independence assumption*, introduced by Qian et al. (2020) and based on the work of Sitlani et al. (2012). According to Qian et al. (2020), this assumption should identify whether estimators of the treatment effect are consistent and unbiased under randomized treatment assignment. The conditional independence assumption states that the covariate at time t (X_{it}) should be independent of the individual's random effects (b_{i0} and b_{i1}) once we account for their history of covariates (H_{it-1}), previous treatments (A_{it-1}), and prior outcomes (Y_{it}). This is implied from the following notation:

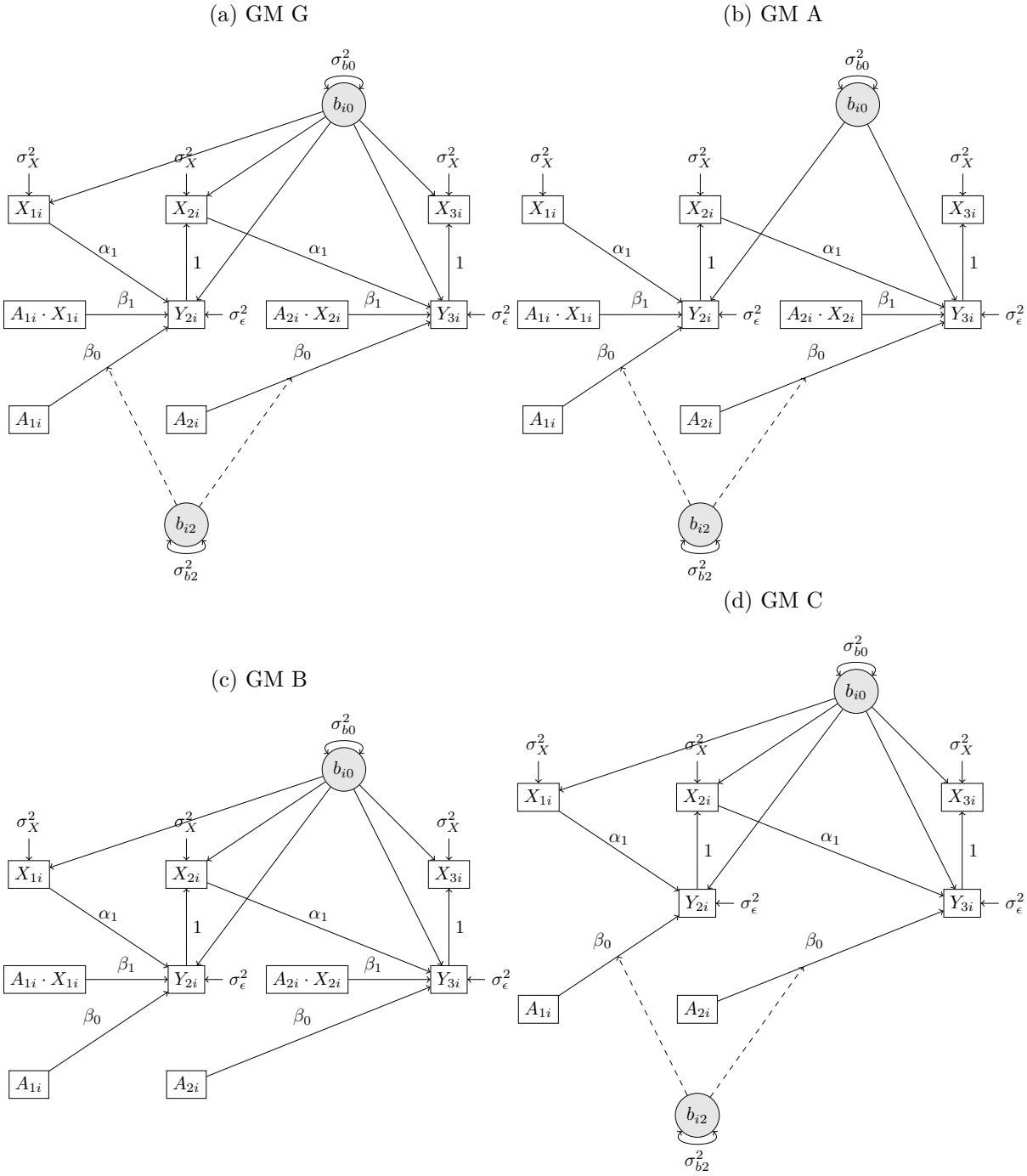
$$X_{it} \perp (b_{i0}, b_{i1}) \mid H_{it-1}, A_{it-1}, Y_{it}.$$

where b_{i0} and b_{i1} represent the random intercept and random slope(s), respectively, and H_{it-1} comprises all observations of that covariate before timepoint t. This assumption allows for X_{it} to be influenced by earlier variables (e.g., outcomes or treatments) but not directly by unobserved individual characteristics (i.e., random effects). If the included endogenous covariates are only

affected by prior outcomes and treatments, the assumption is automatically satisfied. However, as Qian et al. (2020) highlights, ensuring this assumption holds requires careful consideration of theory and domain knowledge.

To clarify the application of the conditional independence assumption, we pair the equations of the generative models (GMs) with path diagrams (Duncan, 1966; Wright, 1934) illustrating the first three timepoints (t) for each model (see Figure 1).

Figure 1: Path Diagrams for Generative Models G, A, B and C (t = 1, 2, 3)



Note. Random effects are represented by grey circles, observed variables by squares and relationships across variables by arrows, where dashed lines are reserved for random slopes.

Let's begin with the general model, GM G (Figure 1a), which Qian et al. (2020) identified as prone to bias. In GM G, the covariate X_{it} is directly influenced by unobserved individual

factors (represented by the random effects, b_{i0}). Consequently, conditioning on prior variables, such as the outcome at the previous timepoint Y_{it}, does not fully block or eliminate the influence of these unobserved factors. As a result, X_{it} remains dependent on the random effects, violating the assumption that X_{it} should be independent of these unobserved factors once we account for prior variables. This violation of the conditional independence assumption explains the biased estimates of the treatment effect observed in GM G, as identified by Qian et al. (2020).

In contrast, GM A, a special case of GM G where no bias was found by Qian et al. (2020), removes the direct link between X_{it} and the random effects b_{i0}. In this case, X_{it} is simply the previous outcome Y_{it} plus some random noise. While there remains an indirect connection between X_{it} and b_{i0} through Y_{it}, conditioning on Y_{it} effectively "breaks the link" between X_{it} and the random effects, satisfying the conditional independence assumption. This explains the unbiased treatment effect estimates found by Qian et al. (2020).

For the additional special cases, GM B and GM C, the direct link between the random effects and X_{it} remains, as in GM G. As a result, these models also violate the conditional independence assumption, leading to biased estimates of the treatment effect.

In summary, GM G, B, and C violate the conditional independence assumption, which suggests that we would expect biased treatment effect estimates for these models. In contrast, GM A satisfies the assumption, supporting unbiased estimates of the treatment effect.

3.2 Backdoor Criterion and DAGs

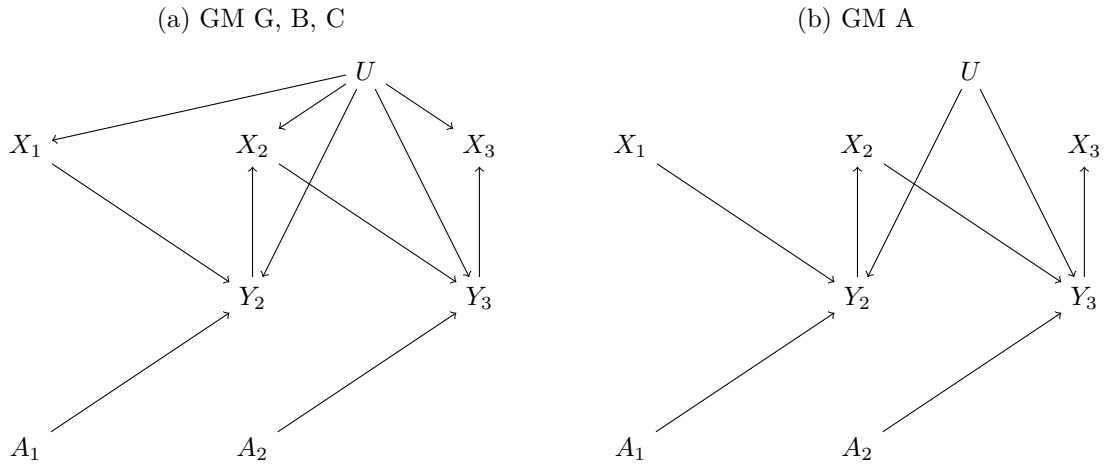
The second criterion for evaluating the presence of bias in treatment effect estimates is the *backdoor criterion* (Pearl, 1988, 2009). This criterion provides a partial solution to the classical problem in causal inference: *causal effects* cannot be directly observed and must instead be inferred from observed associations, which often represent a mixture of causal effects and various undesirable non-causal, or *spurious*, components (Holland, 1986). When it is possible, under ideal conditions (e.g., no measurement error, infinite sample size), to isolate the causal effect

from a combination of causal and spurious components, the causal effect is said to be identified. According to the backdoor criterion (Pearl, 1988, 2009), causal effects can be identified by blocking non-causal paths through conditioning on appropriate variables (e.g., controlling or matching). For instance, a causal effect of exposure on outcome can be identified by including in the analysis (i.e., controlling for) all relevant confounders—common causes that would otherwise induce spurious relationships. However, if spurious paths remain unblocked due to unmeasured variables or measurement error, the treatment and outcome remain linked via backdoor paths, leading to biased estimates of the treatment effect (Kim & Steiner, 2021).

To detect the presence of such backdoor paths, directed acyclic graphs (DAGs) (Pearl, 1995, 2009) are invaluable tools. DAGs generalize conventional linear path diagrams (Duncan, 1966; Wright, 1934) and operate in a fully nonparametric framework. Unlike traditional path diagrams, DAGs make no assumptions about distributional properties (e.g., multivariate normality) or functional forms (e.g., linearity). Instead, they encode qualitative causal assumptions about the data-generating process in the population. Arrows connecting nodes indicate direct causal effects, which may vary in magnitude across individuals (effect heterogeneity) or depend on the values of other variables (effect interaction or modification) (Elwert & Winship, 2014). Notably, random slopes from random-effects models and interaction effects are not explicitly represented in DAGs, which precludes their use for evaluating the conditional independence assumption.

Using the direct causal effects specified in each generative model (GM), we can formulate DAGs for the first three observations, representing the random disturbance b_{0i} as the node U (e.g., Kim & Steiner, 2021, see Figure 2). These diagrams confirm that random slopes and fixed interaction effects are absent. Indeed, this absence explains why the DAGs for GMs G, B, and C are equivalent.

Figure 2: DAGs for Generative Models G, A, B and C (t = 1, 2, 3)



Note. The red arrows show the biased backdoor path(s) in the treatment efffect (before controlling for X_{it}).

We now apply the backdoor criterion to these DAGs to assess potential bias in the treatment effect. For all GMs, there are no backdoor paths in the treatment effect $A_t \to Y_{t+1}$, as A_t lacks any parent nodes. Consequently, covariate X_t need not be controlled to obtain an unbiased total effect. Importantly, including X_{it} does not introduce identification issues, as it is neither a mediator (i.e., on the pathway from A_t to Y_{t+1}) nor a collider (i.e., a common effect of A_t and Y_{t+1}). Therefore, according to the backdoor criterion, controlling for X_{it} should not result in biased estimates of the treatment effect in any of the generative models.

3.3 Simulation Study

Table 2 and Figure 3 present the simulation results for each of the generative models. The β_0 bias in Table 2 refers to $\bar{\hat{\beta}}_0 - \beta_0$, representing the difference between the mean of the estimated parameter values $\bar{\hat{\beta}}_0$ and the prespecified treatment effect $\beta_0 = 1$. Thus, an absolute bias of 0.05 implies a 5% relative bias.

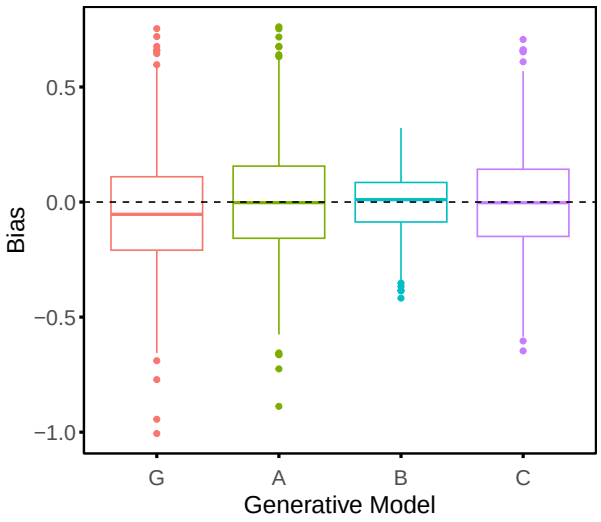
Table 2: Treatment effect bias for Generative Models G, A, B and C over 1000 replications

			β_0		
GM	T	N	Bias	SD	SR
G	10	30	-0.052	0.245	0.999
		100	-0.064	0.134	1.000
		200	-0.051	0.096	1.000
	30	30	-0.024	0.206	0.997
		100	-0.030	0.108	0.996
		200	-0.023	0.080	0.997
A	10	30	0.000	0.238	0.998
		100	-0.012	0.129	1.000
		200	0.003	0.093	0.999
	30	30	-0.001	0.203	0.998
		100	-0.007	0.107	0.996
		200	0.001	0.079	0.996
B	10	30	0.000	0.126	1.000
		100	0.004	0.073	1.000
		200	0.002	0.048	1.000
	30	30	-0.001	0.071	1.000
		100	0.000	0.040	1.000
		200	0.000	0.028	1.000
C	10	30	0.001	0.217	0.999
		100	-0.008	0.121	1.000
		200	0.005	0.087	1.000
	30	30	0.000	0.193	1.000
		100	-0.008	0.103	0.997
		200	0.001	0.075	0.999

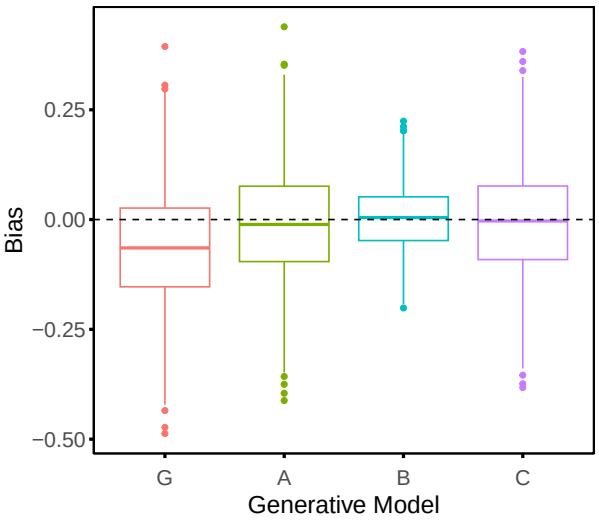
Note. GM: generative model. T: number of timepoints. N: sample size. SD: standard deviation of estimates across replications. SR: model fitting success rate. Bias: $\bar{\hat{\beta}}_0 - \beta_0$, which represents the difference between the mean of the estimated parameter values $\hat{\beta}_0$ and the prespecified treatment effect $\beta_0 = 1$

Figure 3: Estimation bias for the fixed treatment effect β_0 of each generative model for different combinations of sample size N and number of timepoints T over 1000 simulation replications

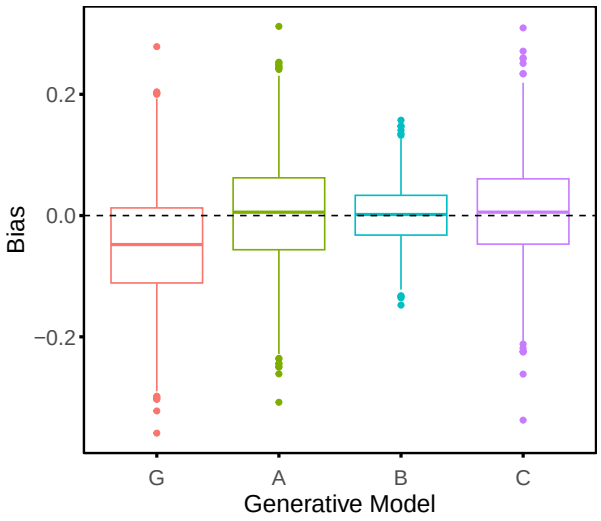
(a) T=10, N=30



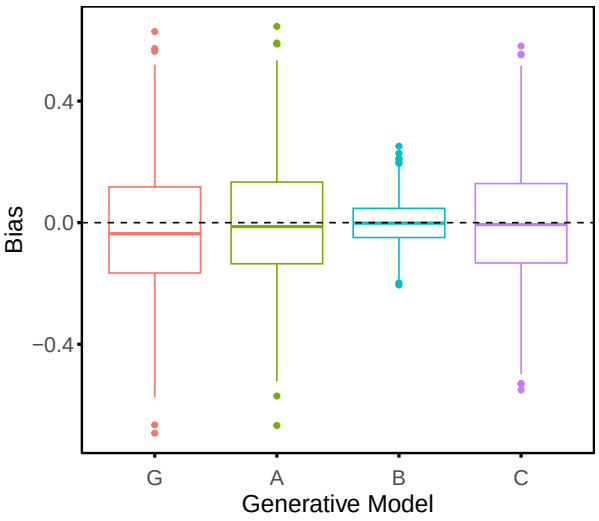
(b) T=10, N=100



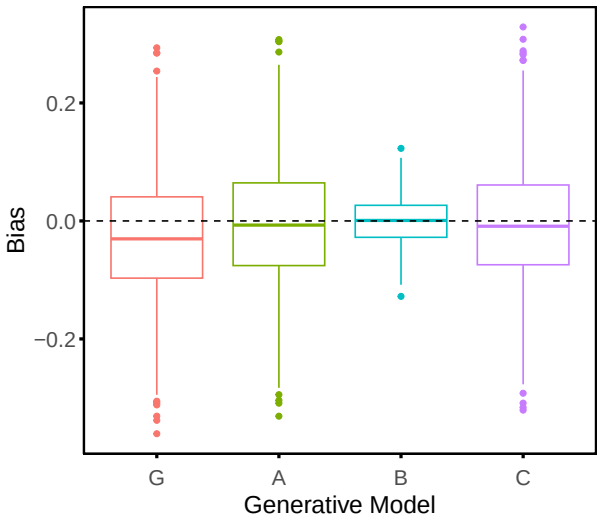
(c) T=10, N=200



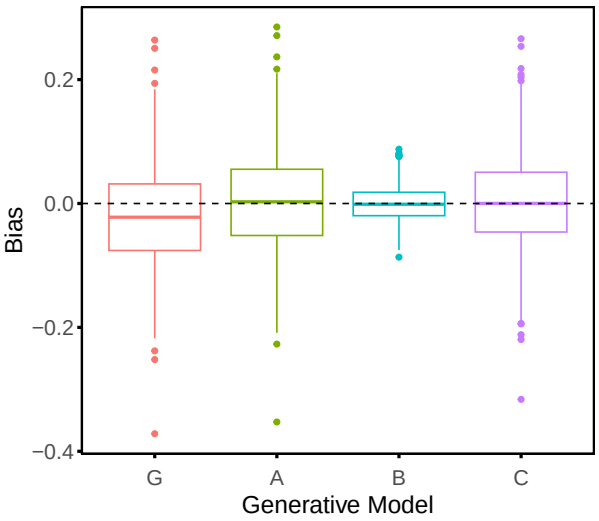
(d) T=30, N=30



(e) T=30, N=100



(f) T=30, N=200



In this reproduction of Qian et al. (2020), the overall pattern was consistent with the original study (see Figure 3): we observed substantial absolute bias ranging from 0.023 (2.3%) to 0.064 (6.4%) for the most general generative model (GM G), and much smaller bias of ≤ 0.015 (1.5%) for GM A. These results align with expectations based on the conditional independence assumption, which predicts that the treatment effect would be unbiased for GM A and biased for GM G. However, the findings contradict the backdoor criterion, which predicts no bias for any of the generative models (GMs). Notably, we found 4 times greater treatment effect bias for GM A in the scenario with $T = 10$ and $N = 100$ than reported by Qian et al. (2020), with a maximum bias of -0.012 (1.2%; (see Table 2), compared to -0.003 (0.3%) found by Qian et al. (2020). For GM G, the size of the bias decreased as the number of time points increased.

For the two additional special cases of GM G, namely GM B and GM C, we observed even smaller absolute bias than for GM A: ≤ 0.010 (1%) for GM C and ≤ 0.005 (0.5%) for GM B (see Table 2). These findings align with the backdoor criterion's prediction of no bias but contradict the expectations based on the conditional independence assumption, which suggests the presence of bias. Additionally, GM B exhibited the smallest absolute bias overall and showed much smaller variability across simulation replications compared to all other GMs. In contrast, the remaining models exhibited comparable levels of variability (see Figure 3 and Table 2).

In summary, these findings suggest that once we modify the most general model G by removing either the dependency of the random intercept on the covariate (GM A), the random slope b_{i2} (GM B), or the interaction term β_1 (GM C), the bias either disappears or becomes negligible. However, neither the backdoor criterion nor the conditional independence assumption provided consistent predictions of treatment effect bias across all models.

4 Discussion

4.1 Main Findings

In this research report, we evaluated several GMs to investigate when endogenous time-varying covariates bias treatment effect estimates in MLMs under randomized treatment. We also assessed the ability of the conditional independence assumption and the backdoor criterion to predict this bias. Consistent with Qian et al. (2020) and the conditional independence assumption, we observed substantially greater bias in the most general model (GM G) compared to a special case (GM A), where the direct effect of the random intercept on the covariate was removed. However, one unexpected result was the fourfold increase in maximum bias for GM A compared to Qian et al. (2020). This discrepancy may stem from differences in the simulation setup (e.g., random number generation or handling of warnings/errors). The two additional special cases of GM G, where the random slope for treatment (GM B) or the interaction term between treatment and covariate (GM C) was removed, exhibited even smaller bias in the treatment effect—despite violations of the conditional independence assumption. These findings tell us that—at least in this particular instance—the dependency between the covariate and treatment, the random slope for treatment, and the interaction effect are all essential for bias to exceed 5.

This naturally raises important questions: why was no discernible bias observed in GM B and GM C, as predicted by the conditional independence assumption, and can this pattern generalize beyond the current generative models? One possibility is that, while bias may exist as predicted, it was canceled out by parameter removal in these specific models. Alternatively, heterogeneity in the treatment effect—whether unexplained via the random slope or explained via the covariate—may both be necessary for bias to occur. This explanation aligns with the statement by Qian et al. (2020) that "applying linear mixed models is problematic because potential moderators of the treatment effect are frequently endogenous" (p. 375). If treatment

effect moderation by a covariate is indeed a prerequisite for bias, it could explain the absence of bias in GM C, where no interaction term was included. However, it remains unclear why Qian et al. (2020) does not explicitly mention this condition (e.g., when introducing the conditional independence assumption). Further research is needed to determine whether these findings generalize or are specific to the evaluated GMs, thereby informing practical recommendations for using MLMs with endogenous time-varying covariates.

Regarding the backdoor criterion and DAGs (Pearl, 1988, 2009), our results suggest that the classical non-parametric DAG may be insufficient to identify bias in GM G. While DAGs account for direct causal effects, they do not impose restrictions on random slopes or interaction effects[2], which are central to the conditional independence assumption. Similar concerns regarding the use of the DAG with Pearl's backdoor criterion in situations with interaction effects have been raised (Attia et al., 2022; Weinberg, 2007). Future research could explore to what extent proposed extensions of the DAG—that incorperate interaction effects—may allow the backdoor criterion to identify bias in the treatment effect estimates.

It should also be noted that the current investigation takes for granted that the marginal (population-averaged) interpretation of the treatment effect estimators of the MLM may not be valid due to the presence of endogenous time-varying covariates Pepe & Anderson (1994). In the context of clustered longitudinal data analysis in psychology, the conditional-on-the-random-effect interpretation of the parameters is often aligned with the scientific interest. Conversely, in fields like epidemiology, where the focus is on estimating the average causal effect of a treatment, the marginal interpretation of the parameters may be more relevant. Future research should integrate insights from the literature on marginal effects and causal inference to address this issue more comprehensively.

Another avenue for future investigation is the role of centering approaches (see Hamaker & Muthén (2020) for an overview). Namely, Antonakis et al. (2021) notes that the assumption

[2]Note that the term "effect modification", while often used interchangeably with "interaction", has a distinct definition in the counterfactual framework (VanderWeele, 2009).

of uncorrelatedness between the random effects and level 1 covariates can be relaxed by using Mundlak's contextual model[3] (Mundlak, 1978): adding the cluster means of each covariate as predictor of the random intercept. Explicitly modeling the source of endogeneity, as advocated by Bell & Jones (2015), may further clarify how it impacts treatment effect bias in GM G.

Finally, Qian et al. (2020) only considered independent random effects, a restrictive assumption that may be violated in practice. Exploring correlated random effects using structural equation modeling frameworks (Rovine & Molenaar, 2000) could provide further insight. Similarly, it would be valuable to examine the implications of endogenous covariates for other longitudinal data analysis methods, such as dynamic structural equation modeling (DSEM; a widely used approach in the social sciences based on the MLM).

4.2 Conclusion

This report is a first step towards understanding the implications of endogenous covariates in multilevel linear models. However, to recognize and understand completely when and why endogenous covariates may trouble an empirical investigation, further research is needed. Future studies should aim to generalize these findings, clarify the role of treatment effect moderation, and explore solutions such as DAG extensions, centering approaches, and alternative longitudinal modeling frameworks.

[3]This is referred to as the Correlated Random Effects (CRE) approach by Wooldridge (2002).

5 References

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods*, *24*(2), 443–483. https://doi.org/10.1177/1094428119877457

Attia, J., Holliday, E., & Oldmeadow, C. (2022). A proposal for capturing interaction and effect modification using DAGs. *International Journal of Epidemiology*, *51*(4), 1047–1053. https://doi.org/10.1093/ije/dyac126

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 148. https://doi.org/10.18637/jss.v067.i01

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*(4), 373–390. https://doi.org/10.1037/a0025813

Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, *3*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Boruvka, A., Almirall, D., Witkiewitz, K., & Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, *113*(523), 1112–1121. https://doi.org/10.1080/01621459.2017.1305274

Daniel, R. m., Cousens, S. n., De Stavola, B. l., Kenward, M. G., & Sterne, J. a. C. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, *32*(9), 1584–1618. https://doi.org/10.1002/sim.5686

Diggle, P. (2002). *Analysis of Longitudinal Data*. OUP Oxford.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, *72*(1), 1–16. https://doi.org/10.1086/224256

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*, 31–53. https://doi.org/10.1146/annurev-

soc-071913-043455

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.2307/2289064

Kim, Y., & Steiner, P. M. (2021). Causal graphical views of fixed effects and random effects models. *British Journal of Mathematical and Statistical Psychology*, *74*(2), 165–183. https://doi.org/10.1111/bmsp.12217

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69–85. https://doi.org/10.2307/1913646

Nahum-Shani, I., Potter, L. N., Lam, C. Y., Yap, J., Moreno, A., Stoffel, R., Wu, Z., Wan, N., Dempsey, W., Kumar, S., & al., et. (2021). The mobile assistance for regulating smoking (MARS) micro-randomized trial design protocol. *Contemporary Clinical Trials*, *110*, 106513. https://doi.org/10.1016/j.cct.2021.106513

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pepe, M. S., & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation*, *23*(4), 939–951. https://doi.org/10.1080/03610919408813210

Qian, T., Klasnja, P., & Murphy, S. A. (2020). Linear mixed models with endogenous covariates:

Modeling sequential treatment effects with application to a mobile health study. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *35*(3), 375–390. https://doi.org/10.1214/19-sts720

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). SAGE.

Rovine, M. J., & Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, *35*(1), 51–88. https://doi.org/10.1207/S15327906MBR3501_3

Schoot, J. H. M. M. R. van de. (2017). *Multilevel analysis: Techniques and applications, third edition* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Sitlani, C. M., Heagerty, P. J., Blood, E. A., & Tosteson, T. D. (2012). Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Statistics in Medicine*, *31*(16), 1738–1760. https://doi.org/10.1002/sim.4510

Team, R. C. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

VanderWeele, T. J. (2009). On the distinction between interaction and effect modification. *Epidemiology*, *20*(6), 863. https://doi.org/10.1097/EDE.0b013e3181ba333c

Walton, A., Nahum-Shani, I., Crosby, L., Klasnja, P., & Murphy, S. (2018). Optimizing digital integrated care via micro-randomized trials. *Clinical Pharmacology & Therapeutics*, *104*(1), 53–58. https://doi.org/10.1002/cpt.1079

Weinberg, C. R. (2007). Commentary: Can DAGs clarify effect modification? *Epidemiology*, *18*(5), 569–572. https://www.jstor.org/stable/20486428

Wodtke, G. T. (2020). Regression-based adjustment for time-varying confounders. *Sociological Methods & Research*, *49*(4), 906–946. https://doi.org/10.1177/0049124118769087

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3),

161–215. https://doi.org/10.1214/aoms/1177732676