Running head: CROSS-LAGGED PANEL MODEL IN MEDICAL RESEARCH

Cross-Lagged Panel Model in Medical Research: A Cautionary Note

Abstract

Longitudinal designs provide a strong inferential basis for uncovering reciprocal effects or causality between variables. For this analytic purpose, a cross-lagged panel model (CLPM) has been widely used in medical research, but the use of the CLPM has recently been criticized in methodological literature because parameter estimates in the CLPM conflate between-person and within-person processes. The aim of this study is to present some alternative models of the CLPM that can be used to examine reciprocal effects, and to illustrate potential consequences of ignoring the issue. A literature search, case studies, and simulation studies are used for this. We examined more than 300 medical papers published since 2009 that applied cross-lagged longitudinal models, finding that in all studies only a single model (typically, the CLPM) was performed and potential alternative models were not considered to test reciprocal effects. In 49% of the studies, only two time points were used, which makes it impossible to test such alternative models. Case studies and simulation studies showed that the CLPM often has worse model fit and markedly different estimates of cross-lagged parameters than alternative models, suggesting that research that relies on the CLPM only may draw erroneous conclusions regarding the presence, predominance, and sign of reciprocal effects as well as about causality.

Cross-Lagged Panel Model in Medical Research: A Cautionary Note

Collecting longitudinal data has become widely popular in medical research and other disciplines due to its statistical advantages over cross-sectional data. One of the biggest advantages of using a longitudinal design is that it can provide richer information for statistical inference aimed at uncovering reciprocal effects or causality between variables to answer questions such as how change (or growth, development) in one variable affects that of the other. More than 30 years ago, Nesselroade and Baltes[1] reviewed the benefits and drawbacks of using longitudinal data in psychology, noting that revealing causes (determinants) of intra-individual change is one of the major strengths of longitudinal data. Likewise, in the econometrics literature, Hsiao[2] argued that panel (i.e., longitudinal) data is effective for inferring dynamic relations between variables.

One of the most common methods for addressing reciprocal effects in medical research is use of a cross-lagged panel model (CLPM; Duncan[3]; also known as a dynamic panel model, autoregressive cross-lagged model, cross-lagged path model, or cross-lagged regression model), especially after the CLPM was integrated into the framework of structural equation modeling (e.g., Finkel[4], Marsh and Yeung[5]) . In these models, reciprocal effects are examined by testing the cross-lagged relations, which are the effect of variable X on variable Y after controlling for the previous effects of X.

The CLPM is a simple and powerful model to test reciprocal effects, and thus it has been widely used. However, the application of the CLPM has also recently been criticized. Notably, Hamaker, Kruiper, and Grasman[6] criticized the use of the CLPM because the cross-lagged estimates in the CLPM conflate between-person and within-person processes, and so the results do not represent the actual within-person relations over time. Between-person relations are the covariation of two variables in terms of individual differences (e.g., individuals with higher X tend to have higher Y relative to individuals with lower X), whereas within-person relation are the covariation within one person of

two variables across time points or situations. Obviously, these two types of relations are conceptually different. As such, the fact that estimates from traditional CLPM conflate between-person and within-person relations means that the cross-lagged estimates from the CLPM are conceptually difficult to interpret. Indeed, the importance of disaggregation to examine within-person processes has been widely acknowledged in the methodological literature (Curran & Bauer[7]; Hamaker[8]; Hoffman & Stawski[9]). Relying on the CLPM may draw erroneous conclusions regarding the presence, predominance, and sign of reciprocal effects as well as about causality.

To address this inherent problem with the CLPM, Hamaker et al[6] proposed a random-intercepts CLPM (RI-CLPM) as a possible analytic option. As discussed later, in the RI-CLPM, individual differences are effectively controlled by the inclusion of a latent variable that represents a time-invariant (but person-variant) trait-like factor; this allows testing the reciprocal effects within individuals. If this model is extended to include measurement errors, the model is equivalent to a so-called (bivariate) stable trait autoregressive trait and state (STARTS) model (Kenny & Zautra[10,11]). Usami, Murayama, and Hamaker[12] discussed the mathematical and conceptual relations between various cross-lagged models, including these models.

These recent studies are insightful and informative, providing applied medical researchers a basis for thinking about how to test reciprocal relations by longitudinal data. However, the arguments are limited mostly to mathematical and conceptual relations. As a result, we still know little about whether, when, and how the choice of different cross-lagged longitudinal models has substantive consequences for parameter estimates of reciprocal effects in practice, leading researchers to draw different conclusions from the same data in medical sciences. The aim of the current manuscript is to show the practical implications and importance of considering these alternative models when investigating reciprocal effects. This is approached through a literature search, case studies, and

statistical simulations. In the literature search, we first investigate the current common practice of longitudinal research in the medical literature, showing that medical researchers rely heavily and almost exclusively on the traditional CLPM when testing reciprocal relations or causality, and do not consider potential alternative models. Then, with case studies and statistical simulations, we illustrate the potential danger of this common practice, showing it can result in mistaken conclusions about reciprocal effects. In the end, we also provide some practical guidelines, hoping to help applied medical researchers who work on longitudinal data in the future.

Cross-Lagged Longitudinal Models

In this paper, we focus on three cross-lagged longitudinal models: the (traditional) CLPM, the RI-CLPM, and the STARTS model. Below, following Usami et al, [12] we describe these models by emphasizing the commonalities and differences among these cross-lagged models. Throughout the paper, we assume that researchers are interested in the reciprocal relation between two variables X and Y, although it is easy to expand the models in a way that include more than two variables (e.g., when examining mediating effects of variables is a main focus of the research).

CLPM

Let x_{it} and y_{it} be the measurements at time point t $(1 \dots t \dots T)$ for individual i $(1 \dots i \dots N)$. In the CLPM, x_{it} and y_{it} are first modeled as

$$\begin{aligned} x_{it} &= \mu_{xt} + x^*_{it}, \\ y_{it} &= \mu_{yt} + y^*_{it}. \end{aligned} \tag{1}$$

Here μ_{xt} and μ_{yt} are the temporal group means at time point t; x^*_{it} and y^*_{it} are temporal deviation terms from the temporal group means for individual i. With these equations, the trajectories of the temporal group mean are implicitly removed from the raw data. By

definition, the deviations have a mean of zero. Then, x_{it} and y_{it} for $t \geq 2$ are modeled as

$$\begin{aligned} x^*_{it} &= \beta_{xt} x^*_{i(t-1)} + \gamma_{xt} y^*_{i(t-1)} + d_{xit}, \\ y^*_{it} &= \beta_{yt} y^*_{i(t-1)} + \gamma_{yt} x^*_{i(t-1)} + d_{yit}, \end{aligned} \tag{2}$$

where β_{xt} and β_{yt} are autoregressive parameters and γ_{xt} and γ_{yt} are cross-lagged regression parameters at time point t. For these parameters, time-invariance can also be assumed (by using β_x and β_y, and γ_x and γ_y) if the cross-lagged relationships are assumed to be stable over time. Note that with $t = 1$, the initial observations x_{i1} and y_{i1} are modeled as exogenous variables.

From the view of Granger causality (Granger[13]), estimates of cross-lagged regression parameters (the longitudinal relationship between Y_{t-1} and X_t after controlling for the baseline X_{t-1}) are key for inferring reciprocal relations between the variables. The residuals d_{xit} and d_{yit} are usually assumed to be normally distributed and correlated:

$$\begin{pmatrix} d_{xit} \\ d_{yit} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega^2_{xt} & \\ \omega_{xyt} & \omega^2_{yt} \end{pmatrix} \right). \tag{3}$$

Here, ω^2_{xt} and ω^2_{yt} are time-variant residual variances and ω_{xyt} is a time-variant residual covariance. As with previous parameters, time-invariant residual variances and covariances can also be assumed (by using ω^2_x, ω^2_y, and ω_{xy}). A path diagram of the CLPM is provided in Figure 1a.

RI-CLPM

In the RI-CLPM (Hamaker et al[6]), x_{it} and y_{it} are modeled as

$$\begin{aligned} x_{it} &= \mu_{xt} + I_{xi} + x^*_{it} \\ y_{it} &= \mu_{yt} + I_{yi} + y^*_{it}, . \end{aligned} \tag{4}$$

Again, μ_{xt} and μ_{yt} are the temporal group means. Critically, the model also includes I_{xi} and I_{yi}, which are the defining characteristic of the RI-CLPM. These are (time-invariant)

trait factors that represent individual's trait-like deviations from temporal group means. Trait factors I_{xi} and I_{yi} have means of 0 and variance–covariance matrix $\boldsymbol{V}$. By accounting for trait factor scores, for each individual, x^*_{it} and y^*_{it} represent temporal deviations from the means of that individual because they are subtracted from the expected scores of individual i (i.e., $\mu_{xt} + I_{xi}$ and $\mu_{yt} + I_{yi}$). Accordingly, in the RI-CLPM, the time series x^*_{it} and y^*_{it} can be considered as within-person fluctuation. Due to this statistical property in temporal deviations, at $t = 1$ the initial deviation terms (x^*_{i1} and y^*_{i1}) are assumed to be uncorrelated with the trait factors. Using these within-person deviation terms, in the RI-CLPM the cross-lagged relations are modeled as in the Equation 2 for $t \geq 2$. A path diagram of the RI-CLPM is provided in Figure 1b.

Because the RI-CLPM accounts for trait factors and then separates stable between-person differences (i.e., trait factors) from within-person fluctuations over time, cross-lagged relations in the RI-CLPM can be considered as the one pertaining to a process that takes place at the within-person level. Therefore, in the RI-CLPM, γ_x and γ_y can be interpreted as the quantity that express the extent to which the two variables influence each other *within* individuals. Because longitudinal data typically include both quantitative information of within-person changes and its individual differences, the CLPM, which does not account for trait factors (i.e., individual differences), fails to disaggregate these two components. As such, the CLPM provides inaccurate estimates for within-person reciprocal effects.

Note that if substituting the cross-lagged relations of Equation 2 into 4, the trait factors, which are separated from independent variables ($x^*_{i(t-1)}$ and $y^*_{i(t-1)}$), can obviously be interpreted as random intercepts in the model. The model is named after this statistical fact. Obviously, the CLPM is a special case of the RI-CLPM, found by letting $I_{xi} = 0$ and $I_{yi} = 0$. The RI-CLPM requires two or more variables to have been measured at three or more time points, while the CLPM requires only two time points.

STARTS model

By extending the RI-CLPM to include measurement error, we obtain the STARTS model (Kenny & Zautra[10,11]). In the (bivariate) STARTS model, y_{it} and x_{it} are decomposed into latent true scores f_{xit} and f_{yit} and measurement errors ϵ_{xit} and ϵ_{yit}. That is,

$$\begin{aligned} x_{it} &= f_{xit} + \epsilon_{xit} \\ y_{it} &= f_{yit} + \epsilon_{yit}. \end{aligned} \tag{5}$$

These measurement errors are usually assumed to be normally distributed and possibly correlated, that is,

$$\begin{pmatrix} \epsilon_{xit} \\ \epsilon_{yit} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{xt}^2 & \\ \psi_{xyt} & \psi_{yt}^2 \end{pmatrix} \right). \tag{6}$$

Here, ψ_{xt}^2 and ψ_{yt}^2 are measurement error variances, and ψ_{xyt} is an error covariance. If needed, time-invariant measurement error (co)variances can be assumed. As in the RI-CLPM, f_{xit} and f_{yit} are modeled as

$$\begin{aligned} f_{xit} &= \mu_{xt} + T_{xi} + f_{xit}^* \\ f_{yit} &= \mu_{yt} + T_{yi} + f_{yit}^*. \end{aligned} \tag{7}$$

Here, f_{xit}^* and f_{yit}^* are the terms expressing temporal deviation from the expected scores of individual i, with accounting for measurement error.

Substituting the equation 7 into the equation 5 provides the specification of the STARTS model:

$$\begin{aligned} x_{it} &= \mu_{xt} + T_{xi} + f_{xit}^* + \epsilon_{xit} \\ x_{it} &= \mu_{yt} + T_{yi} + f_{yit}^* + \epsilon_{yit}. \end{aligned} \tag{8}$$

As in Eq. 2, temporal deviation terms are modeled as

$$\begin{aligned} f_{xit}^* &= \beta_{xt} f_{xi(t-1)} + \gamma_{xt} f_{yi(t-1)} + d_{xit} \\ f_{yit}^* &= \beta_{yt} f_{yi(t-1)} + \gamma_{yt} f_{xi(t-1)} + d_{yit}. \end{aligned} \tag{9}$$

A path diagram of the STARTS model is provided in Figure 1c. Obviously, in the STARTS model, cross-lagged relations are posited between latent true scores, rather than between observed scores, distinguishing it from the RI-CLPM and the CLPM. However, the STARTS model and the RI-CLPM share a common critical feature—the inclusion of trait factors. As such, like the RI-CLPM, cross-lagged parameters (γ_{xt} and γ_{yt}) in the STARTS model reflect within-person reciprocal effects. The STARTS model requires two or more variables to have been measured at four or more time points. This means that we can compare RI-CLPM and the STARTS to determine which of these models fits better to the data so long as more than three waves are available.

When observations may be influenced by measurement errors occurring for procedural reasons, accounting for measurement errors is desirable. However, the specification of measurement error when there is only one indicator variable (such as in the STARTS model) sometimes involves costs in terms of parameter estimation. Indeed, research has reported that the STARTS model often encounters estimation problems such as improper solutions and non-convergence. Conceptually, one primary reason is the fact that unlike trait factor variances (v^2) and residual variances (ω_t^2), the contribution from measurement error variances (ψ_t^2) is temporal: in the model-implied variance-covariance matrix, ψ_t^2 appears at time point t only. Because of this, unstable estimates of some parameters (particularly autoregressive parameters) caused by some aspects of the research design (e.g., small sample size) can easily inflate the variances of the deviation terms (x_t^*, y_t^*), increasing the risk of obtaining negative estimates of ψ_t^2.

Therefore, previous studies have also proposed models that incorporate multiple indicators (rather than a single indicator) to represent latent variables (see Cole et al[14]; Luhmann, Schimmack, & Eid[15]). In addition, research has also suggested the utility of a Bayesian approach to avoid unstable parameter estimation (Lüdtke, Robitzsch, & Wagner[16]).

Review of the literature

Method

To investigate recent trends in the use of cross-lagged longitudinal models in medical research, we conducted a literature search through the UTokyo REsource Explorer (TREE; http://tokyo.summon.serialssolutions.com/) web search engine in June of 2017. TREE aggregates information from many major databases (e.g., Web of Science, PubMed, PsycINFO, Engineering Village, ERIC, JSTOR) and electronic journals under contract with The University of Tokyo. TREE summarizes this collection of information in a single search window, allowing us to perform more comprehensive and efficient literature search than by using the individual databases separately. We first used the English keywords "cross lagged model" and "cross lagged relation", searching English papers published since 2009 in medical journals. In addition, we limited our search to only peer-reviewed papers. Therefore, news items, book reviews, and doctoral dissertations were not considered.

We found 324 medical papers by this method. Of these, we excluded 53 papers that did not apply any cross-lagged longitudinal models to actual data, leaving us with 271 papers. Most of the excluded papers were review papers, statistical simulations, or methodological and statistical discussion. See Table 1 for the complete list of retained papers.

Result

Among 271 papers, 106 (= 40%) papers collected longitudinal data at two time points. 89 (= 33%) papers collected data with three waves, 36 (= 13%) with four waves, 16 (= 6%) papers with five waves, and 24 (= 9%) at more than five time points. The = 40% proportion for two time points is close to the one = 45% reported by Hamaker et al[6] in the field of psychology. With regard to the statistical analysis they performed, 257 papers (= 95%) used the CLPM to analyze longitudinal data, and one paper used a model

similar to the RI-CLPM (see Telley et al, 2015 in Table 1; this model does not assume autoregressive parameters). Other papers applied different models, such as an autoregressive latent trajectory model (Poirier et al, 2016), a latent change score model (LCS; Baydar and Akcinar, 2018; Natsukai et al, 2013; Occhipinti et al, 2015; Usami et al, 2015), a model similar to the latent curve model with structured residuals (Baams et al, 2015; Mustillo et al, 2012; Williams et al, 2011), or a fixed-effects regression model (Baesemer et al, 2016; a model similar to the LCS). For the mathematical and conceptual relations between these models, see Usami et al.[12] Five papers used a multilevel-model framework (Arnett et al, 2016; Cooley et al, 2018; Daniel et al, 2018; Fuller-Tyszkiewicz et al, 2015; Kashdan et al, 2014) to account for individual differences in parameters of the cross-lagged model (see General Discussion on this point). Note that no research applied the STARTS model, and few studies compared analysis results from different cross-lagged models (one exception is a methodological paper of Usami et al, 2015, which compared analysis results from the LCS model and the CLPM).

These results indicate the heavy reliance on the traditional CLPM in the literature. It is also important to note that alternative cross-lagged longitudinal models (e.g., the RI-CLPM and the STARTS model) require at least three time points (with a stability assumption; the STARTS model requires at least four time points with an instability assumption) to fit the model (for the ALT model, we need four time points with a stability assumption). The fact that about 40% of the papers collected data at only two time points suggests that almost half of applied medical research implicitly precludes the option of using these alternative models.

Case studies

Method

To compare analysis results based on different cross-lagged longitudinal models, we focused on the 165 papers that collected longitudinal data with more than two time points. Among them, we randomly selected 50 papers and using the contact information provided in each of the paper we contacted the corresponding authors of the papers via email to request they share the dataset to help our research. In this contact, we emphasized that (1) our primary research purpose is simply to compare analysis results from different cross-lagged models, not to criticize their findings, (2) we would not provide any estimation results from the original paper or relevant information in the datasets to prevent identification of the source of the paper, (3) we would not share the dataset with any other researchers, and that (4) we did not need information about variables that are not relevant to cross-lagged analysis (e.g., personal information of participants).

To increase response rates from authors, we contacted the authors after one month if we had not received a reply from the first contact. As a result, we received a total of 21 responses from the authors (response rate: 42%), and among them, five authors (from five different papers) granted us access to their datasets. We were unable to obtain permissions from the authors of the other 16 papers, mainly because sharing with us might have violated the data sharing policy of their sources. Among the five datasets, two datasets were publicly available online without special permission from the authors, two datasets were provided directly by the authors, and one dataset was provided after a review of the data use agreement that we submitted. Note that one of the datasets provides us with the access only to the sample means and sample (co)variances information (rather than the raw data), which allowed us to estimate the parameters but not to fully account for missing data.

Among five datasets, two datasets have three time points and the others have more than three time points ($M_{\text{time-points}} = 6.0$). The average sample size of these datasets is large ($N = 2,741$). In this paper, we do not give the exact number of participants and time points for each study to prevent the identification of the studies. While all five studies applied CLPM, some of them specified the model in slightly different ways. Specifically, two studies assumed second-order autoregressive and cross-lagged parameters as well as first-order parameters. Another study assumed a mediator between two variables. In addition, one study assumed time-invariant parameters (i.e., stability), while the other four studies did not.

To ensure the comparability of the results between datasets, in the current analysis, we assume time-invariant parameters for autoregressive and cross-lagged coefficients (β and γ) and residual and error (co)variances (ω^2 and ψ^2). In addition, neither second-order parameters nor external variables (e.g., mediators) were included in any of the analyses. This setup also means that the results reported in the current paper are all different from those reported in the original papers. Note that one study collected multi-group data and applied the CLPM using multi-group analysis. For this dataset, we assumed group-invariant parameters for autoregressive and cross-lagged coefficients as well as residual and error (co)variances (i.e., measurement invariance between groups) while setting no constraints on the difference of temporal means between groups.

All analyses were conducted using Mplus version 7.4 (Muthen & Muthen[17]). However, we found improper solutions and non-convergence in four of the five datasets when using maximum likelihood (ML) estimation to fit the RI-CLPM or the STARTS model. In such cases, we instead used Bayes estimation, based on a Markov chain Monte Carlo method under the assumption of non-informative priors. With Bayes estimation, we obtained parameter estimates successfully without any convergence problems. For more detailed discussion about ML and Bayes estimation in terms of estimation problems in

applying the STARTS model, see Lüdtke, Robitzsch, & Wagner[16].

Result

Table 2 provides (unstandardized) autoregressive/cross-lagged parameter estimates and standard errors for the CLPM, the RI-CLPM, and the STARTS model. Except for the cross-lagged parameter estimates in Research 2, all autoregressive/cross-lagged parameter estimates with the CLPM were statistically significant with two-sided $\alpha = .05$. This can be partly attributed to the large sample sizes in these datasets, which increased the statistical power.

Although the RI-CLPM and the STARTS model also showed significant estimates in most cases, $\hat{\gamma}_x$ is not statistically significant in Research 4, while it is significant with the CLPM. Another different result is that the sign of $\hat{\gamma}_x$ in the STARTS model was different from that with the CLPM in Research 3.

We also found notable differences in the magnitudes of parameter estimates among cross-lagged models. The RI-CLPM provided smaller autoregressive parameter estimates ($\hat{\beta}$) than the CLPM did (approximately 0.49 times the size), while the STARTS model provided larger estimates on average (approximately 1.45 times the size).

The relation between parameter estimates from different cross-lagged longitudinal models must depend in complicated ways on the magnitude of the parameter values and on research design factors (e.g., N and T), and we need to be careful when generalizing the findings. But, one potential explanation for the increased autoregressive parameters in the STARTS model is the dissociation of measurement errors in the model because the autoregressive parameters are the major source of correlations (i.e., the variance–covariance matrix) between time points. For the RI-CLPM, in contrast, the decreased autoregressive parameter estimates may be a consequence of trait factors, which would explain a large portion of the correlations between time points.

The differences in estimates of autoregressive parameters between the RI-CLPM and the STARTS model also lead to differences between their cross-lagged parameter estimates and those found by the CLPM. In this case study, the RI-CLPM and the STARTS model showed smaller cross-lagged estimates (in absolute value, 0.66 and 0.62 times the size, respectively) from those with the CLPM. Although we need to be careful about the generalizability of findings, it is well-known that the magnitude of within-cluster (in this case, within-person) relations (i.e., cross-lagged parameters in the RI-CLPM and the STARTS model) is smaller than those of between-cluster (in this case, between-person) relations, when the between-cluster difference is larger than the within-cluster difference. The decreased cross-lagged effects could be explained by this so-called ecological fallacy (Robinson[18]).

With regard to standard errors, interestingly, the standard errors of $\hat{\gamma}$ in the RI-CLPM and the STARTS model are, on average, 1.6 and 2.7 times, respectively, the size of those with the CLPM. These results indicate that the inclusion of parameters that are specific to these models (i.e., trait factor (co)variances in the RI-CLPM and those and error (co)variances in the STARTS model) leads to an increase in standard errors. In combination with the observed upward or downward changes in autoregressive and cross-lagged parameter estimates, these results indicate that the RI-CLPM and the STARTS model will produce substantially different results on statistical tests than the CLPM will.

It is also important to note that, among the five datasets, the CLPM was chosen as the best model in terms of model fit only once, when the Bayesian Information Criterion was used in Research 2. This result indicates that many previous studies that applied only the CLPM may have drawn erroneous conclusions about the magnitude and presence of reciprocal effects.

The results described here indicate the importance of comparing alternative models

when testing for reciprocal effects, and the potential (in most cases, unintended) consequences of not considering multiple models. However, one might be concerned about the generalizability of the results due to the small number of studies (i.e., five) presented here. Another important issue is the improper solutions observed in two of the five datasets when applying the STARTS model. To address these issues more extensively, we conducted two statistical simulation studies, one focusing on the frequency of improper solutions and the other focusing on parameter estimates. Although the previous case studies indicated that these models could produce overtly different parameter estimates, to the best of our knowledge, no previous research has performed statistical simulation that directly compared the parameter estimates (and associated standard errors) produced by different cross-lagged longitudinal models we discussed here (i.e., the CLPM, the RI-CLPM, and the STARTS model). In addition, although some past studies have examined the frequency of improper solutions, focusing especially on the STARTS model (e.g., Cole et al[14]; Lüdtke et al[16]), no studies have systematically investigated the differences of longitudinal models used and examined the potential impact of model misspecification. Our statistical simulation also aims to extend the previous studies by addressing these points.

Simulation Study

Frequency of improper solutions

To systematically investigate the rate of improper solutions under various conditions, we performed Monte Carlo simulations, where both data generation model and analysis models were selected from the three models we have discussed, resulting in 9 ($= 3 \times 3$) combinations of data generation and analysis models. This way, we can examine the potential influence of model misspecification (as well as the correct model specification) on improper solutions. For simplicity of the simulations, the stability of

parameters was assumed.

For data generation, we systematically changed the number of total participants ($N = 200$, 600, $1,000$), the number of time points ($T = 4$, 6, 8), and the size of autoregressive parameters ($\beta = \beta_x = \beta_y = 0.5$, 0.7, 0.9). In this simulation, cross-lagged parameters γ were all fixed to 0.2. For the STARTS model, measurement error variances were set to ($\psi^2 = \psi_x^2 = \psi_y^2 = 0.2$, 0.5, 0.8). For the other models, ψ^2 is always set to zero. Variances of the temporal deviation terms at the first time point ($Var(x_{i1}^*)$ and $Var(y_{i1}^*)$), which are equivalent to those of observations in case of the CLPM, were fixed to $1 - \psi^2$. The size of β reflects the determination coefficients in cross-lagged regressions. For models with trait factors (i.e., the RI-CLPM and the STARTS model), we posited normal distribution for the trait factors and their variances were set to the half size of those of temporal deviation terms at the first time point (i.e., to $Var(x_{i1}^*)/2$ and $Var(y_{i1}^*)/2$).

Without loss of generality, the temporal group means were set to $\mu_{xt} = \mu_{yt} = t - 1$ for each time point. Correlation of the trait factors was set to 0.2. Correlation of temporal deviation terms at the first time point was set to 0.2, and in the STARTS model (time-invariant) correlations between measurement errors were set to 0.2. Finally, residual variances were fixed to $\omega^2 = \omega_x^2 = \omega_y^2 = 0.2$, and correlation of residuals between variables was fixed to 0.2 for each time point.

We generated simulated data (200 trials for each combination) by crossing these factors, resulting in 81 ($= 3(N) \times 3(T) \times 3(\beta) \times 3(\psi^2)$) combinations of factors for each pair of data generation model and data analysis model. Each simulated dataset was analyzed by the three types of analysis models, and we counted the number of improper solutions, which was defined as (1) out-of-range parameter estimates (e.g., negative variances parameters) or (2) a singular approximate Hessian matrix after termination of iteration. The whole simulation procedure, including data generation and analysis, was conducted in R (R Core Team[19]) using the lavaan (Rosseel[20]) package with the ML

estimation method. Simulation code is available in the Online Supporting Materials.

Table 3 presents the marginal proportions of improper solutions observed with each data analysis model under each level of the factors we manipulated. When the CLPM is used for analysis, it did not show improper solutions under any conditions. When CLPM is used for data generation, Table 3 shows that RI-CLPM and the STARTS model showed very large proportions of improper solutions (in the range of 40%–100%). Notably, in cases of the STARTS model, which posited measurement error (co)variances and residuals, 90% of the results exhibited improper solutions. Interestingly, the manipulated factors, such as the number of total participants (N) and number of time points (T) did not influence the results much. These results indicate that the impact of model misspecification dominates the risk of improper solutions, with the factors being manipulated playing a much smaller role. The same pattern was observed with different data generation models. Model misspecification was the biggest cause of improper solutions, and the STARTS model especially produced a higher number of improper solutions.

One particularly important observation is that improper solutions were still observed in the STARTS model even when the model was correctly specified. Indeed, the proportion of improper solutions was unacceptably high, at more than 70%. Note that, even compared with previous investigations (Cole et al[14]; Lüdtke et al[16]), our simulations showed larger number of improper solutions. This might be attributed to differences in the stability of measurements between the current simulations and the simulations in the previous studies. Instead of controlling the residual variances, the variances of all variables were set to 1 in the simulations of both Cole et al[14] and Lüdtke et al, [16], while we did not do this in the current investigation. In most of the current simulation conditions, the variances of variables are implicitly assumed to increase over time, as is often the case with longitudinal data in developmental/clinical research. Thus, the relative impacts of trait factor variances, (time-invariant) measurement error variances, and residual variances on

observations become smaller at later time points, increasing the risk of out-of-range estimates in these variance estimates. Another important difference is that such previous investigations have considered univariate (rather than bivariate) version of the STARTS model. The bivariate version of the STARTS model, which we simulated in the current study, might have a bigger risk of improper solutions caused by a singular Hessian matrix.

For correctly specified models, the RI-CLPM performed better, especially when sample size and the number of time points were larger. However, the proportion of improper solutions was still not negligible (at $10 - -15\%$). Therefore, although the RI-CLPM and the STARTS model can be considered as alternatives to the CLPM when investigating within-person reciprocal relations, these models might be susceptible to improper solutions, especially in the presence of model misspecification.

Statistical properties of estimates

To investigate the statistical properties of cross-lagged parameter estimates in each cross-lagged longitudinal model, we performed another Monte Carlo simulation. As in the previous simulation, the data generation model and analysis model were selected from the three types of models. For data generation, we systematically changed the number of total participants ($N = 200$, 600, $1,000$), the number of time points ($T = 4$, 6, 8), and the size of autoregressive parameters ($\beta = \beta_x = \beta_y = 0.5$, 0.7) and cross-lagged parameters ($\gamma = \gamma_x = \gamma_y = 0.0$, 0.1, 0.2). Other parameters were the same as in the previous simulation.

We generated simulated data (100 trials for each combination) by crossing these factors, resulting in 162 ($= 3(N) \times 3(T) \times 2(\beta) \times 3(\gamma) \times 3(\psi^2)$) combinations of factors for each pair of data generation model and data analysis model. Each simulated dataset was analyzed by the three types of analysis models. In this simulation, when improper solutions (e.g., out-of-range parameter estimates or a singular approximate Hessian

matrix) were observed, the results were discarded and the simulations were repeated until the total number of successful trials was 100 for each condition. The whole simulation procedure, including data generation and analysis, was conducted in R (R Core Team[19]) using the lavaan (Rosseel[20]) package with the ML estimation method. Simulation code is available in the Online Supporting Materials.

From the results of the previous simulation, we expected a large proportion of improper solutions when applying the RI-CLPM and the STARTS model (especially when the analysis model was misspecified), which would indicate that the parameter estimates in these models might be substantially biased by discarding results with improper solutions. Therefore, we limited our attention here mainly to the differences in the standard errors of the cross-lagged parameters estimates between models. This is because the standard errors should be less influenced by the occurrence of improper solutions, given that improper solutions are mainly caused by the magnitude of point estimates (e.g., out-of-range parameter estimates or a singular approximate Hessian matrix) rather than the magnitudes of associated standard errors.

Table 4 presents the marginal means of estimated standard errors for different data generation models and analysis models under the different conditions of N and T. Note that we aggregated the other factors (β, γ, and ψ^2) because we did not observe any notable influences of these factors on the estimates of standard errors. From Table 4, as we have observed from the five case studies, standard errors in the RI-CLPM and the STARTS model tend to be larger than those in the CLPM in most cases. Specifically, the standard errors were $1.1 - -2.2$ times the size of the CLPM in the RI-CLPM and $0.8 - -4.2$ times the size in the STARTS model. In applying the CLPM and the RI-CLPM, the standard errors decrease as T increases for correctly specified analysis models.

However, this was not the case when applying the STARTS model. Although it

shows similar magnitudes of standard errors under different conditions of T when the analysis model was correctly specified, we saw the opposite pattern when the analysis model was incorrectly specified: the standard errors became *larger* as T became larger. When the true model is either the RI-CLPM or the STARTS model, standard errors with the CLPM tend to be smaller than those with other models, indicating that (incorrectly) applying the CLPM without comparing alternative models entails a greater risk of committing a type-1 error when statistically testing for reciprocal relations.

Table 5 shows the marginal means of the proportions of models reaching inconsistent conclusions about the statistical significance of cross-lagged estimates when the true value of γ is (a) zero or (b) nonzero. For example, in a pair of the CLPM and the RI-CLPM, the models suggested different conclusions in two ways, with the CLPM showing significant results but the RI-CLPM not, and vice versa. In each condition, the upper row counts the sum of these two proportions, while the lower row counts only the first case, where the CLPM shows significant results but the RI-CLPM does not.

From Table 5(a), it is very obvious that different models tend to show inconsistent results (in terms of statistical significance) for cross-lagged estimates. Notably, when they show different results, in most cases only the simpler model (the CLPM being compared with the RI-CLPM and the STARTS model; the RI-CLPM being compared with the the STARTS model) showed a significant result. Note that the influences of T and N vary depending on the data generation models and analysis models. However, from Table 5(b), when γ=0 models tend to converge to agreement more frequently, although increasing T and N increased the risk of different statistical conclusions between models.

Although we have to take care about possible biased results here as a consequence of discarding the results when improper solutions were produced when applying the RI-CLPM and the STARTS model, this simulation clearly demonstrates that statistical tests of cross-lagged effects can often show substantially inconsistent results, regardless of

the number of participants or time points, especially when cross-lagged relations are actually present. One primary source of this should be the inflated standard errors of cross-lagged parameter estimates, as observed earlier.

General Discussion

In this manuscript, we discussed limitations of the commonly-used CLPM (specifically, the conflation of between-person and within-person effects) and the importance of considering alternatives such as the RI-CLPM and the STARTS model when investigating reciprocal effects within individuals. Through a literature search, case studies, and statistical simulations, we showed the current predominance of the CLPM for testing cross-lagged effects in the medical literature and demonstrated the risk of drawing inconsistent conclusions depending on the model tested. In addition, we showed the potential risk of improper solutions when applying alternative models (the STARTS model, in particular) with the ML method, especially when the model is misspecified.

One important observation was that many researchers implicitly precluded the option of using RI-CLPM or the STARTS model by collecting data from only two time points. Given the substantially different results obtained from different models, we recommend that applied researchers collect longitudinal data at more than two time points whenever possible. If we were to assume the instability of parameters across time points, more than three time points are required to compare model fits between RI-CLPM and the STARTS model. If collecting data from a larger number of time points, then performing model selection based on model fit indices is an important step in minimizing the risk of drawing erroneous conclusions about reciprocal effects. Parameter estimation may be a serious obstacle, though, especially when applying the STARTS model. Although improving research design (e.g., by choosing an appropriate sample size) is important, choosing a different estimation strategy, such as Bayesian estimation (Lüdtke,

Robitzsch, & Wagner[16]), and choosing a better specified analysis model via model selection seems to be more useful. Future research should more intensively investigate the utility of Bayesian estimation in applying various cross-lagged models.

Some limitations should be noted. First, the RI-CLPM and the STARTS model assume that autoregressive and cross-lagged parameters are fixed across participants, but we could incorporate random slopes for these effects. This would allow investigating the possible individual differences in within-person reciprocal effects. Such a model can be easily implemented with a multilevel modeling framework (e.g., Bringmann et al[21]; Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker[22]). We suspect that such new models may be more susceptible to improper solutions given the increased number of parameters and complicated covariance structure. Future investigations should provide clearer insights into how researchers can choose the appropriate analysis model in practice. A second point relates to the extension of the current discussion to other statistical models. For example, medical researchers are often interested in testing mediation effects to understand the mechanism by which one variable influences another (e.g., Richiardi, Bellocco, & Zugna[23]; Ten Have & Joffe[24]; VanderWeele[25]), and they are often assessed in a longitudinal design (e.g., Huang & Yuan[26]; Preacher[27]). The issue of the current paper applies especially to longitudinal mediation models that include cross-lagged relationships (e.g., a dynamic autoregressive mediation model; Maxwell, Cole & Mitchell[28]). If researchers fail to account for stable individual differences, then the estimated mediation effects conflate between-person and within-person processes. The current discussion is useful for considering possible alternatives when evaluating longitudinal mediation effects, and investigating the statistical properties of estimates and the frequency of estimation problems should be intriguing topics for future research. Finally, although the current study focused only on the medical literature, future study should examine common practices for testing reciprocal effects in other fields. This would give us more empirical

insights into the similarities and differences in these cross-lagged models.

Data Availability

All analysis and simulation codes used during this study are included in this published article (and its Supplementary Information files). The some numerical datasets analysed during the current study are not publicly available due to the data sharing policy of their sources, these datasets are however available from the authors upon reasonable request and with permission of each third party.

Ethical Approval and Informed Consent

We have not carried out any experiments during this study.

Additional Information (competing interests)

The authors declare no competing interests.

References

1. Nesselroade JR, Baltes PB. *Longitudinal research in the study of behavior and development*. New York: Academic Press; 1979.

2. Hsiao C. *Analysis of Panel Data*. London: Cambridge University Press; 2014.

3. Duncan OD. Some linear models for two-wave, two-variable panel analysis. *Psychol. Bull* 1969;72:177-182.

4. Finkel SE. *Causal Analysis with Panel Data*. Thousand Oaks, CA: Sage; 1995.

5. Marsh HW, Yeung AS. Causal effects of academic self-concept on academic achievement - structural equation models of longitudinal data. *J. Educ. Psychol.* 1997;89:41-54. doi:10.1037/0022-0663.89.1.41

6. Hamaker EL, Kuiper RM, Grasman RPPP. A critique of the cross-lagged panel model. *Psychol. Methods* 2015;20:102-116. http://dx.doi.org/10.1037/a0038889

7. Curran PJ, Bauer DJ. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu. Rev. Psychol.* 2011;62:583-619.

8. Hamaker EL. Why researchers should think "within-person": A paradigmatic rationale. in *Handbook of research methods for studying daily life* (ed. Matthias, M and Tamlin, C.) Guilford Press; 2012;43-61.

9. Hoffman L, Stawski RS. Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Res. Hum. Dev.* 2009;6: 97-120.

10. Kenny DA, Zautra A. The trait-state-error model for multiwave data. *J. Consult. Clin. Psychol.* 1995;63:52-59.

11. Kenny DA. Zautra A. Trait-state models for longitudinal data in *New methods for the analysis of change* (ed. Collins, L. and Sayer, A.) Washington, DC: American Psychological Association; 2001:243-263.

12. Usami S, Murayama K, Hamaker EL. A unified framework of cross-Lagged longitudinal models. Paper presented at the 82nd Annual Meeting of Psychometric Society 2017.

13. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 1969;37:424-438.

14. Cole DA, Martin NC, Steiger JH. Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychol. Methods* 2005;10:3-20.

15. Luhmann M, Schimmack U, Eid M. Stability and variability in the relationship between subjective well-being and income. *J. Res. Pers.* 2011;45:186-197.

16. Lüdtke O, Robitzsch A, Wagner J. More stable estimation of the STARTS model : A Bayesian approach using Markov Chain Monte Carlo techniques. *Psychol. Methods* 2018;23:570-593.

17. Muthén LK, Muthén BO. *Mplus User's Guide*. Los Angeles, CA: Muthén Muthén: 1998-2017.

18. Robinson WS. Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.* 1950;15:351-357.

19. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2016. https://www.R-project.org/.

20. Rosseel Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* 2012;48:1-36. http://www.jstatsoft.org/v48/i02/

21. Bringmann LF, *et al.* A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*. 2013;8:e60188. doi:10.1371/journal.pone.0060188

22. Schuurman NK, Ferrer, E, de Boer-Sonnenschein, M. & Hamaker, E.L. How to compare cross-lagged associations in a multilevel autoregressive model. *Psychol. Methods* 2016;21:206-221.

23. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int. J. Epidemiol.* 2013;42:1511-1519.

24. Ten Have TR, Joffe MM. A review of causal estimation of effects in mediation analyses. *Stat. Methods Med. Res.* 2012;21:77-107.

25. VanderWeele TJ. Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* 2016;37:17-32.

26. Huang J, Yuan Y. Bayesian dynamic mediation analysis. *Psychol. Methods* 2017;22:667-686.

27. Preacher K. Advances in mediation analysis: a survey and synthesis of new developments. *Annu. Rev. Psychol.* 2015;66:825-852.

28. Maxwell SE, Cole DA. & Mitchell, M.A. Bias in crosssectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behav. Res.* 2011;46:816-841.

Table 1. The list of 271 papers that applied cross-lagged models

ID	Authors	Year	Journal	the number of time point (T)
1	Adachi & Willoughby	2016	Child development	4
2	Andrade	2014	Journal of Adolescence	2
3	Arnett et al.	2012	Journal of Abnormal Child Psychology	4
4	Arnett et al.	2016	Journal of Child Psychology and Psychiatry	10
5	Avalon et al.	2016	Psychology and Aging	3
6	Baams et al.	2015	Archives of Sexual Behavior	3
7	Baesemer et al.	2016	Journal of Abnormal Child Psychology	8
8	Baneriee et al.	2011	Child Development	3
9	Bavdar & Akcinar	2018	Journal of Abnormal Child Psychology	5
10	Beauiean et al.	2013	Social Psychiatry and Psychiatric Epidemiology	2
11	Bekkhus et al.	2011	Journal of Abnormal Child Psychology	4
12	Bennett et al.	2015	Journal of Child Psychology and Psychiatry	3
13	Bentley et al.	2013	Quality of Life Research	4
14	Best et al.	2016	Journal of the American Geriatrics Society	2
15	Birkeland et al.	2016	International Archives of Occupational and Environmental Health	2
16	Bohlmann et al.	2015	Child Development	3
17	Bolhuis et al.	2014	Psychological Medicine	2
18	Bolhuis et al.	2017	Journal of the American Academy of Child and Adolescent	2
19	Bondü et al.	2016	Journal of Adolescence	2
20	Bonvanie et al.	2016	Pain	2
21	Bourque et al.	2016	Journal of the American Academy of Child & Adolescent	4
22	Boves et al.	2014	Journal of Abnormal Child Psychology	2
23	Bovlan et al.	2010	Journal of the American Academy of Child & Adolescent	3
24	Breeman et al.	2015	Journal of Abnormal Child Psychology	3
25	Brière et al.	2014	Comprehensive Psychiatry	3
26	Brinke et al.	2017	Journal of Abnormal Child Psychology	4
27	Brown et al.	2011	Journal of Aging and Health	2
28	Burns et al.	2016	Annals of Behavioral Medicine	5
29	Calvete et al 1	2015	Journal of Child and Family Studies	2
30	Calvete et al 2	2015	Journal of Adolescence	3
31	Chang & Shaw	2016	Child Psychiatry and Human Development	2
32	Chen et al.	2012	Journal of Child Psychology and Psychiatry	4
33	Chen et al.	2015	PLoS ONE	2
34	Cheng et al.	2016	Child: Care. health and development	2
35	Chi et al.	2014	AIDS and Behavior	3
36	Choi et al.	2012	Tobacco Control	10
37	Christensen & Knardahl	2012	Pain	2
38	Conway et al.	2017	Child Psychiatry and Human Development	2
39	Cooley et al.	2018	Journal of Abnormal Child Psychology	3
40	Cowlishaw et al.	2013	Aging & Society	2
41	Crocetti et al.	2016	PLoS ONE	6
42	Crocetti et al.	2017	Child Development	5
43	Crosnoe et al.	2012	Journal of Health and Social Behavior	2
44	Dakanalis et al.	2015	European Child & Adolescent Psychiatry	2
45	Dakanalis et al.	2016	Journal of Clinical Psychology	2
46	Daniel et al.	2014	Journal of Adolescence	3
47	Daniel et al.	2018	Child Development	3
48	Danzo et al.	2017	Journal of Adolescence	4
49	Das & Sawin	2016	Archives of Sexual Behavior	2
50	De Laet et al.	2014	Child Development	3

Tables and Figures