PLOS ONE

RESEARCH ARTICLE

Modeling reciprocal effects in medical research: Critical discussion on the current practices and potential alternative models

Satoshi Usami[1]*, **Naoya Todo**[2], **Kou Murayama**[3,4]

1 Department of Education, University of Tokyo, Bunkyo-ku, Tokyo, Japan, **2** Department of Psychology, University of Tsukuba, Tsukuba, Ibaraki, Japan, **3** Department of Psychology, University of Reading, Reading Berkshire, United Kingdom, **4** Research Institute, Kochi University of Technology, Kami, Kochi, Japan

* usami@ct.u-tokyo.ac.jp

Citation: Usami S, Todo N, Murayama K (2019) Modeling reciprocal effects in medical research: Critical discussion on the current practices and potential alternative models. PLoS ONE 14(9): e0209133. https://doi.org/10.1371/journal.pone.0209133

Editor: Timo Gnambs, Leibniz Institute for Educational Trajectories, GERMANY

Received: November 27, 2018

Accepted: June 4, 2019

Published: September 27, 2019

Data Availability Statement: The simulation codes underlying the results presented in Tables 3, 4, 5, 6, 7, 8 and 9 are provided in S1 File. The data underlying Table 2 were gathered from five authors of different studies. The authors did not have special access privileges that others would not have. By referring to the contact information provided in each of the papers listed in Table 1 and following the steps shown in Fig 2, interested researchers can access datasets in the same manner as the authors.

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 purpose. 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 alternative models. Case studies and simulation studies showed that the CLPM and alternative models often produce different (or even inconsistent) parameter estimates for reciprocal effects, suggesting that research that relies only on the CLPM may draw erroneous conclusions about the presence, predominance, and sign of reciprocal effects. Simulation studies also showed that alternative models are sometimes susceptible to improper solutions, even when reseachers do not misspecify the model.

Introduction

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

Funding: This research was supported in part by JSPS Kakenhi, Grant Number 26885007 and 16K17305 (Satoshi Usami), and Leverhulme Trust Research Leadership Award, Award Number RL-2016-030, and JSPS Kakenhi, Grant Number 16H06406 (Kou Murayama). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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, Kuiper, 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. Therefore, the CLPM can be a possible option when within-person variances are negligible, or when researchers are not interested in uncovering within-person relations because predicting/forecasting outcomes is the main analytic purpose.

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 within-person reciprocal effects 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 (within-person) 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 importance of considering these alternative models and the potential problems in current practices to infer 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, and do not consider potential alternative models. Such reliance on the traditional CLPM makes it difficult to infer within-person reciprocal effects. Then, with case studies and statistical simulations, we illustrate the potential danger of this common practice (i.e., applying only the CLPM), 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 effect 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 ($1 \ldots t \ldots T$) for individual i ($1 \ldots i \ldots N$). 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 relations 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 (i.e., their variances and covariance are assumed).

From the view of Granger causality (Granger [13]), estimates of cross-lagged regression parameters (the longitudinal relation between Y_{t-1} and X_t after controlling for the baseline X_{t-1}) are key for inferring reciprocal effects 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 Fig 1a.

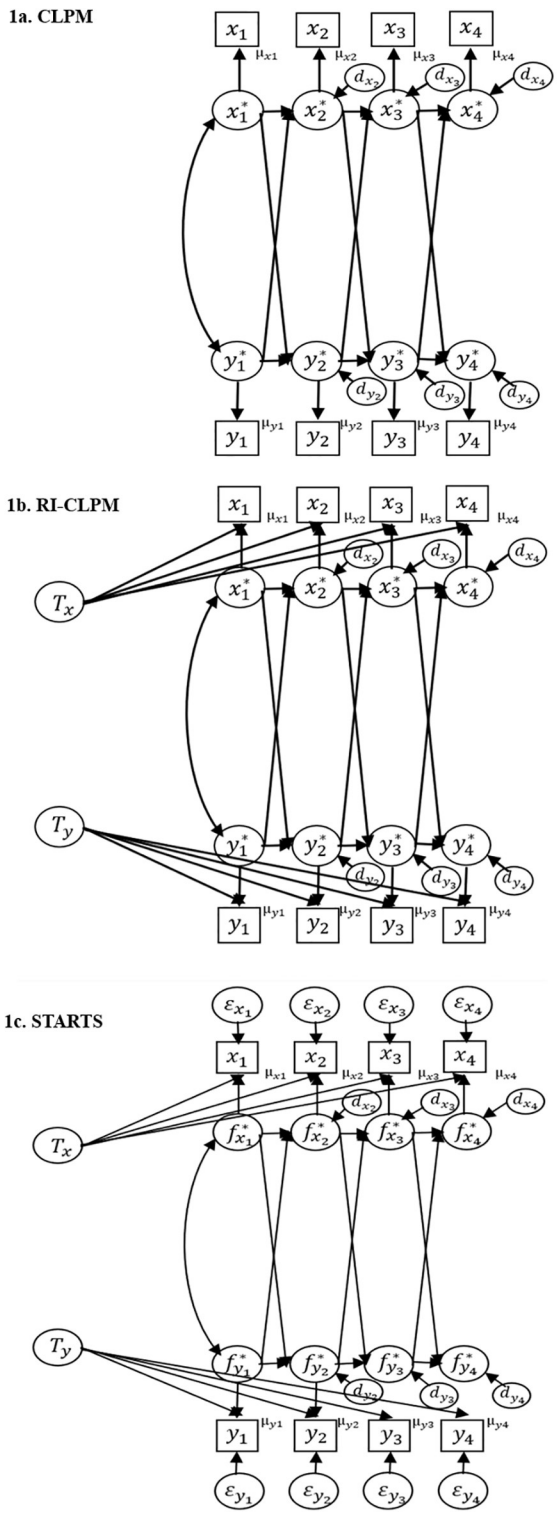

Fig 1. Path diagrams of cross-lagged models. (a) the CLPM. (b) the RI-CLPM. (c) the STARTS model.

https://doi.org/10.1371/journal.pone.0209133.g001

Residuals and error covariances and variance and covariances between trait factors are all omitted for clarity of presentation. Variances and covariance of latent true scores at $t = 1$ (i.e., exogeneous variables) are also omitted for the same purpose. Means of trait factors are set to zero in the RI-CLPM and the STARTS model. In all three cross-lagged models, means are modeled through temporal group means (μ_{xt} and μ_{yt}).

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. Notably, 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 $\mathbf{V}$. 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 Eq 2 for $t \geq 2$. A path diagram of the RI-CLPM is provided in Fig 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 when substituting the cross-lagged relations of Eq 2 into Eq 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 $I_{yi} = 0$ (i.e., trait factors variances are zero). 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,

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

Here, ψ^2_{xt} and ψ^2_{yt} 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 Eq 7 into the Eq 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 Fig 1c. Although measurement errors are not assumed in the CLPM or RI-CLPM, 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 (ν^2) and residual variances (ω^2_t), the contribution from measurement error variances (ψ^2_t) is temporal: in the model-implied variance-covariance matrix, ψ^2_t 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^2_t.

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 323 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 270 papers. Most of the excluded papers were review papers, statistical simulations, or methodological and statistical discussion. Table 1 lists the papers retained for the investigation (authors, publication year, journal, the number of time points). Full references are available in Table A in S1 File.

Result

Among 270 papers, 106 (= 39%) 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 proportion for two time points (= 39%) is close to the one reported by Hamaker et al [6] (= 45%) 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). Unfortunately, almost 40% of the papers collected data with only two time points, indicating many applied medical research implicitly precludes the option of using these alternative models.

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	Ayalon 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	Banerjee et al.	2011	Child Development	3
9	Baydar & Akcinar	2018	Journal of Abnormal Child Psychology	5
10	Beaujean 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.	2015	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 Psychiatry	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 Psychiatry	4
22	Boyes et al.	2014	Journal of Abnormal Child Psychology	2
23	Boylan et al.	2010	Journal of the American Academy of Child & Adolescent Psychiatry	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

(*Continued*)

Table 1. (Continued)

ID	Authors	Year	Journal	the number of time point (T)
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
51	de Leeuw et al.	2011	Pediatrics	3
52	de Wilde et al.	2016	Journal of Abnormal Child Psychology	3
53	Dempsy et al.	2016	Journal of Clinical Psychology in Medical Settings	2
54	Deschenes et al.	2016	Journal of Diabetes	4
55	Diamantopoulou et al.	2011	Europian Child & Adolscent Psychiatry	5
56	Ding et al_1	2014	BMC Neuroscience	2
57	Ding et al_2	2014	Behavioral and Brain Functions	2
58	Doane et al.	2016	Journal of Religion and Health	3
59	Eggers et al.	2017	AIDS and Behavior	2
60	Fabbri et al.	2015	Journals of Gerontology: Biological Sciences	3
61	Faller et al.	2017	Psycho-oncology	2
62	Fanti & Munoz Centifanti	2014	Child Psychiatry & Human Development	2
63	Fátima et al.	2014	Journal of Abnormal Child Psychology	4
64	Feldt et al.	2016	Scandinavian Journal of Work, Environment & Health	5
65	Fielder et al.	2014	Journal of Sex Research	4
66	Fletcher & Johnson	2016	Journal of Child and Family Studies	2
67	Flouri et al.	2015	Child: Care, health and development	3
68	Flouri et al.	2016	Journal of Abnormal Child Psychology	3
69	Flournoy et al.	2016	Child Development	2
70	Foti et al.	2010	American Journal of Psychiatry	5
71	Freedman et al.	2015	European Journal of Psychotraumatology	2
72	French et al.	2014	Child Development	3
73	Frijins et al.	2010	Journal of Adolescence	4
74	Fuller-Tyszkiewicz et al.	2015	Journal of Adolescence	34
75	Garbarski	2014	Journal of Health and Social Behavior	9
76	Garon-Carrier et al.	2016	Child Development	3
77	Gershoff et al.	2012	Child Development	2
78	Giard et al.	2016	Journal of Abnormal Child Psychology	2
79	Girard et al.	2017	European Child and Adolescent Psychiatry	4
80	Girard et al.	2014	PLoS ONE	5
81	Good et al.	2017	Psychiatry Research	2
82	Goodman et al.	2014	Infant Mental Health Journal	3
83	Greven et al.	2012	Journal of Child Psychology and Psychiatry	2
84	Greven et al.	2011	Journal of Abnormal Child Psychology	2
85	Gudmundsson et al.	2015	Acta Psychiatrica Scandinavica	3
86	Gutenbrunner et al.	2018	Journal of Abnormal Child Psychology	3
87	Hale et al.	2011	Journal of Child Psychology and Psychiatry	3
88	Hale III et al.	2016	European child and adolescent psychiatry	6
89	Hall et al.	2015	PLoS ONE	3
90	Hallett et al.	2010	American Journal of Psychiatry	2
91	Hamama-Raz et al.	2015	European Journal of Cancer Care	2
92	Hannigan et al.	2017	Journal of Child Psychology and Psychiatry	2
93	Hanson et al.	2016	PLoS ONE	4
94	Hanson et al.	2017	Scandinavian Journal of Work, Environment & Health	4

(*Continued*)

Table 1. (Continued)

ID	Authors	Year	Journal	the number of time point (T)
95	Harlaar et al.	2011	Child Development	2
96	Harris et al.	2015	Child Development	5
97	Harvey et al.	2016	Journal of Abnormal Psychology	4
98	Henchoz et al.	2014	Quality of Life Research	2
99	Hiemstra et al.	2013	PLoS ONE	5
100	Hietanen et al.	2016	Ageing & Society	4
101	Hill et al.	2013	Journal of Adolescence	2
102	Hinnant et al.	2013	Child Development	3
103	Hipwell et al.	2011	Journal of Child Psychology and Psychiatry	9
104	Holmes et al.	2016	Journal of Abnormal Child Psychology	4
105	Hopwood et al.	2010	Psychological Medicine	6
106	Houkes et al.	2011	BMC Public Health	3
107	Howarth et al.	2016	Child development	4
108	Huizink et al.	2014	Journal of Psychosomatic Obstetrics & Gynecology	3
109	Husby & Wichstrom	2017	Journal of Abnormal Child Psychology	4
110	Huyghebaert et al.	2016	International Journal of Stress Management	2
111	Ibrahim et al.	2009	Social Science & Medicine	3
112	In-Albon et al.	2017	Child Psychiatry and Human Development	3
113	Jackson & Cunningham	2017	Preventive Medicine	5
114	Jäggi et al.	2016	Journal of Adolescence	4
115	Jansen et al.	2013	Pediatrics	4
116	Kashdan et al.	2014	Archives of Sexual Behavior	21
117	Keijsers et al.	2012	Child Development	3
118	Keles et al.	2017	Journal of Abnormal Child Psychology	3
119	Kilian et al.	2012	Social Psychiatry and Psychiatric Epidemiology	3
120	Kim et al.	2018	Child Development	3
121	Kimonis et al.	2015	Journal of Abnormal Child Psychology	3
122	Kiviruusu et al.	2016	PLoS ONE	4
123	Klass et al.	2017	Psychological Medicine	8
124	Klimstra et al.	2014	Social Psychiatry and Psychiatric Epidemiology	4
125	Kochel et al.	2012	Child Development	3
126	Koen et al.	2012	Journal of Adolescent Health	2
127	Koleck et al.	2017	Quality of Life Research	2
128	Konttinen et al.	2014	International Journal of Obesity	3
129	Kuijpers et al.	2015	Journal of Child and Family Studies	2
130	Kuja-Halkola et al.	2015	Journal of Child Psychology and Psychiatry	4
131	Labhart et al.	2017	Behavioral Medicine	2
132	Lange et al.	2017	Child & Adolescent Mental Health	5
133	Lanz & Tagliabue	2014	Journal of Adolescence	2
134	Lavigne et al.	2015	Journal of Abnormal Child Psychology	3
135	Leadbeater & Jacqueline	2015	Journal of Abnormal Child Psychology	7
136	Leadbeater et al.	2009	Child Development	4
137	Lewis et al.	2014	European Child & Adolescent Psychiatry	2
138	Li & Zhang	2015	Social Science & Medicine	3
139	Liat et al.	2009	Child Development	2
140	Lifshitz-Vahav et al.	2017	Aging & Mental Health	2
141	Lindwall et al.	2011	Health Psychology	2

(*Continued*)

Table 1. (Continued)

ID	Authors	Year	Journal	the number of time point (T)
142	Liu et al.	2016	Journal of Health and Social Behavior	2
143	Loukas	2009	Journal of Abnormal Child Psychology	2
144	Lowe et al.	2014	Journal of Abnormal Psychology	3
145	Lucy et al.	2013	Journal of Adolescence	2
146	Luengo Kanacri et al.	2017	Child Development	2
147	Luo et al.	2012	Social Science & Medicine	3
148	Luyckx et al.	2012	Journal of Adolescent Health	2
149	Luyckx et al.	2010	Diabetes Care	4
150	Magee et al.	2014	Acta Pædiatrica	3
151	Mannering et al.	2011	Child Development	2
152	Marschall-Lévesque et al.	2017	Journal of Adolescent Health	3
153	Marshall et al.	2014	Child Development	2
154	Marsiglio et al.	2014	Journal of Child & Adolescent Trauma	2
155	Martinent & Nicolas	2017	International Journal of Stress Management	2
156	Martz et al.	2016	JAMA Psychiatry	3
157	Masquillier et al.	2015	AIDS and Behavior	2
158	Mauno et al.	2011	International Archives of Occupational and Environmental Health	3
159	McAdams et al.	2014	Journal of Adolescence	3
160	McAdams et al.	2015	Psychological Medicine	3
161	Meier et al.	2015	Family Practice	2
162	Micalizzi et al.	2016	Journal of Abnormal Child Psychology	2
163	Miller et al_1	2017	Journal of Abnormal Child Psychology	3
164	Miller et al_2	2017	Psychoneuroendocrinology	3
165	Mitchison et al.	2015	PLoS ONE	5
166	Moberg et al.	2011	Behavior Genetics	2
167	Mrug et al.	2009	Journal of Abnormal Child Psychology	2
168	Muratori et al.	2016	Comprehensive Psychiatry	3
169	Murphy et al.	2017	Journal of Clinical Psychology	3
170	Mustillo et al.	2012	Journal of Health and Social Behavior	9
171	Natsukai et al.	2013	Child Development	2
172	Neece et al.	2012	American Journal on Intellectual and Developmental Disabilities	7
173	Negriff et al.	2015	Journal of Adolescent Health	3
174	Newland et al.	2015	Journal of Child and Family Studies	3
175	Nielsen et al.	2017	International Archives of Occupational and Environmental Health	2
176	Nishiguchi et al.	2016	Psychiatry Research	2
177	Occhipinti et al.	2015	PLoS ONE	6
178	Olesen et al.	2013	BMC Psychiatry	9
179	Paek et al,	2016	Annals of Behavioral Medicine	3
180	Palosaari et al.	2013	Journal of Abnormal Psychology	3
181	Palosaari et al.	2016	Journal of Abnormal Child Psychology	3
182	Pastorelli et al.	2016	Journal of Child Psychology and Psychiatry	2
183	Patalay et al.	2015	Journal of Child Psychology and Psychiatry	3
184	Pearl et al.	2014	Journal of Child and Family Studies	4
185	Peter et al.	2016	Social Science & Medicine	2
186	Pettersson et al.	2011	BMC Public Health	2
187	Peyre et al.	2016	BMC Psychiatry	2
188	Pickard et al.	2017	Journal of the American Academy of Child and Adolescent Psychiatry	3

(*Continued*)

Table 1. (Continued)

ID	Authors	Year	Journal	the number of time point (T)
189	Poirier et al.	2016	European Child & Adolescent Psychiatry	5
190	Pössel & Black	2014	Journal of Clinical Psychology	3
191	Preckel et al.	2013	Journal of Adolescence	3
192	Priest et al.	2017	BMC Psychiatry	2
193	Rappe	2009	Journal of Abnormal Child Psychology	2
194	Rawal et al.	2014	Journal of Child Psychology and Psychiatry	2
195	Rhodes et al.	2015	Annals of Behavioral Medicine	3
196	Ribeiro et al.	2011	BMC Pediatrics	2
197	Richardson et al.	2011	Social Psychiatry and Psychiatric Epidemiology	2
198	Richie et al.	2015	Child Development	5
199	Richter et al.	2015	International Archives of Occupational and Environmental Health	2
200	Rivas-Drake et al.	2017	Child Development	3
201	Rommel et al.	2015	PLoS ONE	3
202	Ruttle et al.	2015	Psychoneuroendocrinology	3
203	Salihovic et al.	2012	Journal of Abnormal Child Psychology	4
204	Savage et al.	2015	Journal of the American Academy of Child & Adolescent Psychiatry	4
205	Senste et al.	2017	Journal of Abnormal Child Psychology	3
206	Seymour et al.	2014	Journal of Abnormal Child Psychology	3
207	Shaffer et al.	2013	Journal of Abnormal Child Psychology	6
208	Shields & Beaver	2011	Journal of Adolescent Health	2
209	Shimazu et al.	2009	Social Science & Medicine	3
210	Skalická et al.	2015	Child Development	2
211	Solberg et al.	2016	Psychological Medicine	3
212	Song et al.	2012	Helthcare Informativs Research	3
213	Spanos et al.	2010	Journal of Abnormal Psychology	3
214	Spilt et al.	2014	Child Development	4
215	Stavrakakis et al.	2012	Journal of Adolescent Health	3
216	Stinglhamber et al.	2015	PLoS ONE	2
217	Stratton et al.	2014	The Journal of Pain	3
218	Sturaro et al.	2011	Child Development	4
219	Sutin & Zonderman	2012	Psychological Medicine	2
220	Szabo et al.	2014	Journal of Crohn's and Colitis	4
221	Tabri et al.	2015	Psychological Medicine	104
222	Tang et al.	2009	Ageing International	2
223	Taylor et al.	2013	Psychological Medicine	2
224	Taylor et al.	2014	Journal of Autism and Developmental Disorders	2
225	Telley et al.	2015	Journal of Health and Social Behavior	3
226	Teppers et al.	2014	Journal of Adolescence	2
227	Tiet et al.	2010	Journal of Child and Family Studies	2
228	Tiggelman et al.	2015	Quality of Life Research	3
229	Timmermans et al.	2010	Psychological Medicine	7
230	Ting-Lan & Bellmore	2012	Journal of Abnormal Child Psychology	3
231	Trucco et al.	2014	Journal of Child Psychology and Psychiatry	2
232	Tsai et al.	2017	Journal of Abnormal Child Psychology	2
233	Tseng et al.	2015	Journal of Abnormal Child Psychology	3
234	Tucker et al_1	2013	Journal of Adolescent Health	2
235	Tucker et al_2	2013	Journal of Adolescent Health	4

(*Continued*)

Table 1. (Continued)

ID	Authors	Year	Journal	the number of time point (T)
236	Usami et al.	2015	Multivariate Behavioral Research	6
237	Van Dorn et al.	2017	Psychological Medicine	11
238	van Dulmen et al.	2012	Journal of Adolescent Health	3
239	van Zalk & Tillfors	2017	Child and Adolescent Psychiatry and Mental Health	3
240	Vanhalst et al.	2013	Journal of Abnormal Child Psychology	5
241	Vaz et al.	2014	PLoS ONE	2
242	Vella et al.	2017	Medicine and Science in Sports and Exercise	2
243	Vitezova et al.	2015	Maturitas	2
244	von Salisch et al.	2017	Journal of Abnormal Child Psychology	2
245	von Stumm & Deary	2013	Psychology and Aging	2
246	Voss et al.	2016	European Journal of Ageing	2
247	Waller et al.	2015	Journal of Abnormal Child Psychology	2
248	Wang & Fredricks	2014	Child Development	3
249	Wang & Kenny_1	2014	Journal of Abnormal Child Psychology	3
250	Wang & Kenny_2	2014	Child Development	2
251	Wang et al.	2012	Child: Care, health and development	2
252	Webb et al.	2016	Journal of Adolescent Health	5
253	Weinstein et al.	2017	PeerJ	2
254	Welp et al.	2016	Critical Care	3
255	Whelan et al.	2015	Journal of Child Psychology and Psychiatry	2
256	Wichstrøm et al.	2016	Journal of Adolescence	4
257	Wickrama et al.	2010	Journal of Aging and Health	3
258	Williams et al.	2011	Child Development	7
259	Wolf et al.	2016	Psychological Medicine	2
260	Wolff	2011	Dyslexia	3
261	Wols et al.	2015	Journal of Adolescence	2
262	Wood et al.	2012	Child Development	7
263	Wouters et al.	2016	AIDS and Behavior	2
264	Yan & Dix	2014	Journal of Child Psychology and Psychiatry	4
265	Yu et al.	2015	Social Science & Medicine	9
266	Zahl et al.	2017	Pediatrics	3
267	Zavos et al.	2012	Behavior Genetics	2
268	Zhou et al.	2014	PLoS ONE	3
269	Zhou et al.	2015	Psychiatry Research	3
270	Zhu et al.	2017	Journal of Abnormal Child Psychology	3
271	van den Eijnden et al.	2010	Journal of Abnormal Child Psychology	2

https://doi.org/10.1371/journal.pone.0209133.t001

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 these papers, 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. To summarize the procedure for case studies as well as literature review so far, a flow diagram is provided in Fig 2. 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

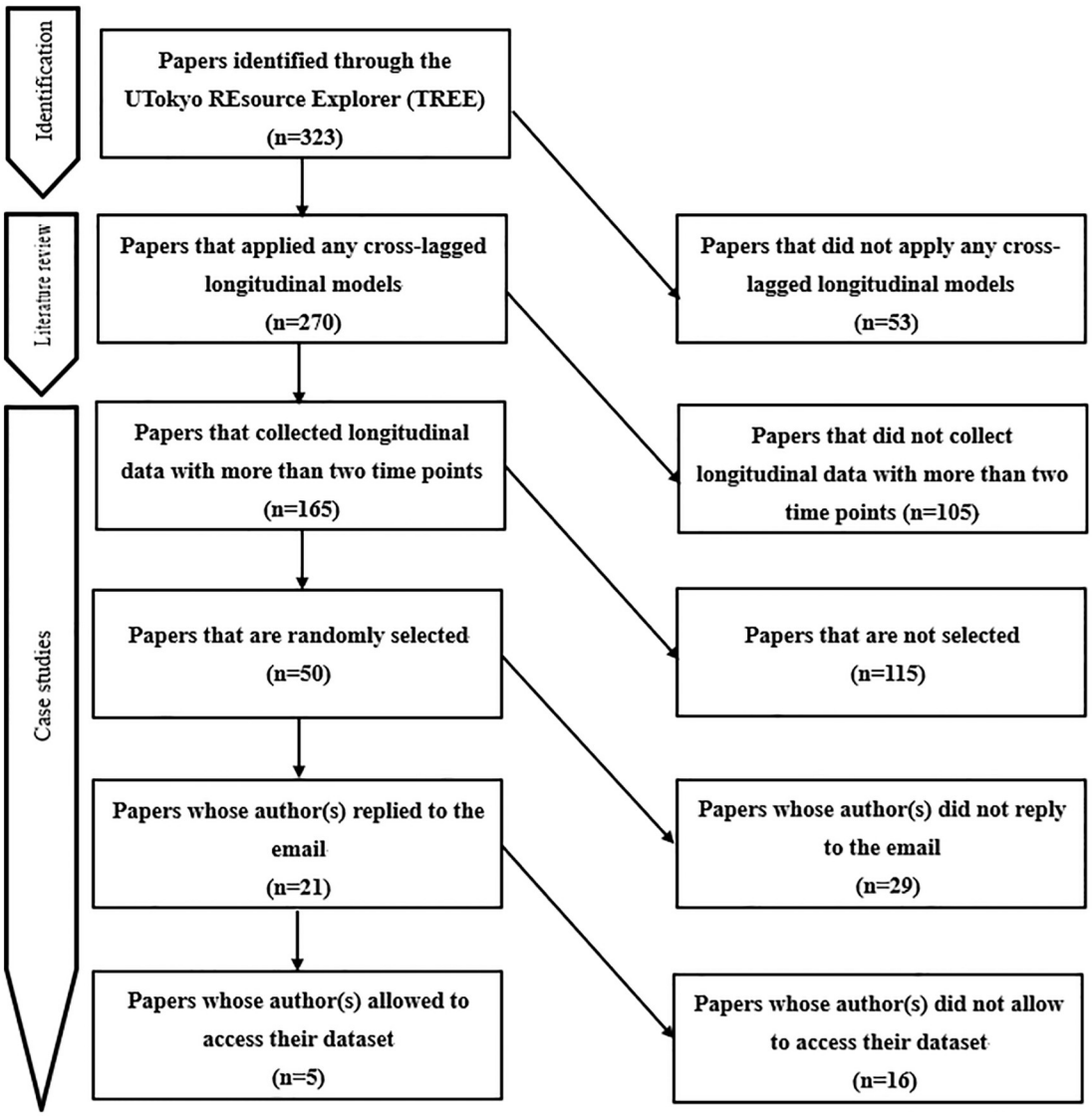

Fig 2. Flow diagram for literature review and case studies.

https://doi.org/10.1371/journal.pone.0209133.g002

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 (mean of the number of time points is 6.0). The average sample size of these datasets is large (= 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 (i.e., negative variance for trait factor or singular Hessian matrix was produced) and non-convergence in four of the five datasets when using maximum likelihood (ML) estimation to fit the RI-CLPM or the STARTS model. One potential reason is that large auto-regressive parameters might have adversely affected the risk of obtaining negative estimates of trait factor variances (as well as other variances) of these models. 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).

Table 2. Autoregressive/Cross-lagged parameter estimates and standard errors from the RI-CLPM or the STARTS model to those from the CLPM.

		Research 1 $N>1000$ $T=3$ maximum likelihood (ML)						Research 2 $N<1000$ $T=3$ Bayes						Research 3 $N>1000$ $T>3$ Bayes						Research 4 $N>1000$ $T>3$ Bayes						Research 5 $N>1000$ $T>3$ Bayes					
		CLPM		RI-CLPM		STARTS		CLPM		RI-CLPM		STARTS		CLPM		RI-CLPM		STARTS		CLPM		RI-CLPM		STARTS		CLPM		RI-CLPM		STARTS	
		Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
β_x		1.05	0.01	0.74	0.04	-	-	0.67	0.05	0.32	0.09	-	-	0.70	0.01	0.40	0.02	1.69	0.04	0.66	0.01	0.20	0.01	0.89	0.03	0.60	0.01	0.24	0.01	0.76	0.02
γ_x		0.38	0.13	0.53	0.22	-	-	-0.68	0.48	-0.90	0.86	-	-	0.03	0.01	0.02	0.01	-0.08	0.02	-0.06	0.00	0.00	0.01	-0.01	0.02	0.02	0.00	0.01	0.00	0.01	0.00
β_y		0.46	0.02	0.14	0.03	-	-	0.52	0.03	0.48	0.05	-	-	0.55	0.01	0.29	0.02	0.86	0.03	0.66	0.01	0.15	0.01	0.91	0.03	0.99	0.00	0.93	0.01	1.02	0.00
γ_y		0.01	0.00	0.01	0.00	-	-	0.00	0.00	0.00	0.01	-	-	0.18	0.02	0.11	0.03	0.30	0.05	-0.22	0.01	-0.11	0.02	-0.10	0.04	0.08	0.01	0.08	0.01	0.09	0.02
ML	AIC	25774.04		25706.86*																											
	BIC	25892.65		25847.04*																											
	RMSEA	0.091		0.079*																											
	CFI	0.953		0.969*																											
	SRMR	0.094		0.092*																											
Bayes	DIC							17439.74		17434.83*				82280.84		81476.47		80266.74*		71558.99		70109.62*		70160.40		109428.88		107221.91		106354.77*	
	BIC							17498.64*		17508.56				82422.04		81632.90		80681.61*		71679.28		70249.66*		70323.92		109635.95		107448.26		106602.48*	

CLPM. . .cross-lagged panel model; RI-CLPM. . .random intercepts CLPM; STARTS. . .stable trait autoregressive trait and state; AIC. . .Akaike Information Criterion; BIC. . .Bayesian information criterion; RMSEA. . .root mean square error of approximation; CFI. . .comparative fit index; SRMR. . .standardized root mean square residual; DIC. . .deviance information criterion

Bold value indicates that associated parameters are statistically significant ($p<.05$); "-" indicates that estimates can not be obtained due to the limited number of time point in the dataset

* indicates that associated model was preferred based on the corresponding model fit index

https://doi.org/10.1371/journal.pone.0209133.t002

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 largely 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 × 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) × 3(T) × 3(β) × 3(ψ^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 S1 File.

Table 3 presents the marginal proportions (i.e. proportion after aggregating across all the other factors) of improper solutions observed with each data analysis model under each level of the factors we manipulated. We mainly inspected marginal proportions in order to have an overall grasp of the factors that relate to the frequency of improper solutions. 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

Table 3. Marginal proportions of improper solutions observed at each data analysis model under each level of the factors.

	Data generation model								
	CLPM			RI-CLPM			STARTS		
Analysis model	CLPM	RI-CLPM	STARTS	CLPM	RI-CLPM	STARTS	CLPM	STARTS	CLPM
$\beta = 0.5$	.00	.66	.95	.00	.01	.98	.00	.00	.71
$\beta = 0.7$	.00	.59	.99	.00	.08	1.00	.00	.00	.89
$\beta = 0.9$	.00	.67	1.00	.00	.05	1.00	.00	.20	.94
$\psi^2 = 0.2$	.00	.65	.98	.00	.05	.98	.00	.00	.84
$\psi^2 = 0.5$	.00	.64	.99	.00	.04	1.00	.00	.00	.82
$\psi^2 = 0.8$	.00	.63	.98	.00	.05	1.00	.00	.20	.88
$N = 200$	.00	.65	.98	.00	.11	1.00	.00	.07	.90
$N = 400$	.00	.64	.98	.00	.05	1.00	.00	.07	.87
$N = 800$	.00	.63	.98	.00	.02	.99	.00	.07	.83
$N = 1600$	.00	.64	.98	.00	.01	.99	.00	.07	.78
$T = 4$	.00	.64	.98	.00	.10	.98	.00	.01	.75
$T = 6$	.00	.63	.98	.00	.03	1.00	.00	.09	.81
$T = 8$	.00	.64	.99	.00	.01	1.00	.00	.11	.98

https://doi.org/10.1371/journal.pone.0209133.t003

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 of developmental changes and growths. Standard latent growth model (LGM) also implicitly has that assumption (e.g., in linear LGM, variance of true score increases over time). 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 showed smaller proportions of improper solutions than the STARTS model, 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 effects, 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 S1 File.

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. Standard errors might 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 Comparing the magnitudes of standard errors among models is useful because this might suggest the reason why inconsistent results are obtained among models in testing statistical significance of cross-lagged parameters, as we will discuss later.

Table 4 presents the marginal means (i.e. means aggregated across all of the other factors) of estimated standard errors for different data generation models and analysis models.

Table 4. Marginal means of standard errors estimated at each model.

	Data generation model								
	CLPM			RI-CLPM			STARTS		
Analysis model	CLPM	RI-CLPM	STARTS	CLPM	RI-CLPM	STARTS	CLPM	RI-CLPM	STARTS
$\gamma = 0$	0.02	0.03	0.08	0.01	0.03	0.12	0.02	0.03	0.41
$\gamma = 0.1$	0.02	0.03	0.09	0.01	0.03	0.13	0.02	0.03	0.40
$\gamma = 0.2$	0.02	0.03	0.10	0.01	0.03	0.15	0.02	0.03	0.42
$\beta = 0.5$	0.02	0.03	0.06	0.01	0.03	0.09	0.02	0.03	0.32
$\beta = 0.7$	0.02	0.03	0.12	0.01	0.03	0.18	0.02	0.03	0.50
$\psi^2 = 0.2$	0.01	0.03	0.06	0.01	0.03	0.08	0.02	0.03	0.12
$\psi^2 = 0.4$	0.02	0.03	0.08	0.01	0.03	0.13	0.02	0.03	0.30
$\psi^2 = 0.6$	0.02	0.03	0.13	0.01	0.03	0.18	0.02	0.03	0.82
$N = 200$	0.02	0.04	0.14	0.02	0.04	0.20	0.03	0.04	0.67
$N = 600$	0.01	0.02	0.07	0.01	0.02	0.11	0.02	0.02	0.35
$N = 1000$	0.01	0.02	0.06	0.01	0.02	0.09	0.01	0.02	0.22
$T = 4$	0.02	0.04	0.11	0.02	0.04	0.16	0.02	0.04	0.53
$T = 6$	0.02	0.02	0.08	0.01	0.02	0.12	0.02	0.03	0.39
$T = 8$	0.01	0.02	0.08	0.01	0.02	0.12	0.02	0.02	0.32

https://doi.org/10.1371/journal.pone.0209133.t004

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 1.3–2.6 times the size of the CLPM in the RI-CLPM and 3.3–38.7 times the size in the STARTS model. The standard errors decrease as T and N increase in cross-lagged models. In addition, β and ψ^2, which relate to the (relative) magnitudes of measurement error variances, explain the magnitudes of the estimated standard errors in the STARTS model. Although these estimated standard errors in the RI-CLPM and the STARTS model might be somewhat biased by discarding the results with improper solutions, Table 4 provides an important suggestion for practice: 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 runs a great risk of committing a type-1 error when statistically testing for reciprocal effects.

Tables 5, 6 and 7 shows the marginal means of the proportions of models reaching consistent/inconsistent conclusions about the statistical significance of cross-lagged parameters for different data generation models and analysis models.

From these tables, it is obvious that different models tend to show inconsistent results (in terms of statistical significance) for cross-lagged parameters when $\gamma \neq 0$. 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, and when $\gamma = 0$ models tend to converge to agreement more frequently. Note, however, that our simulations used relatively small values for

Table 5. Marginal means of proportions that models suggest consistent/inconsistent conclusions about reciprocal relations (CLPM vs RI-CLPM).

	Data generation model											
	CLPM				RI-CLPM				STARTS			
Analysis model	both non-sig	STARTS only	RI-CLPM only	both sig	both non-sig	STARTS only	RI-CLPM only	both sig	both non-sig	STARTS only	RI-CLPM only	both sig
$\gamma = 0$	0.94	0.01	0.04	0.01	0.92	0.01	0.07	0.00	0.92	0.03	0.04	0.01
$\gamma = 0.1$	0.01	0.00	0.46	0.53	0.07	0.02	0.38	0.53	0.31	0.02	0.51	0.16
$\gamma = 0.2$	0.00	0.00	0.20	0.80	0.00	0.00	0.18	0.82	0.06	0.00	0.53	0.40
$\beta = 0.5$	0.32	0.00	0.15	0.52	0.34	0.02	0.12	0.52	0.47	0.02	0.27	0.23
$\beta = 0.7$	0.31	0.00	0.31	0.37	0.32	0.00	0.30	0.38	0.38	0.01	0.46	0.15
$\psi^2 = 0.2$	0.31	0.00	0.25	0.43	0.32	0.02	0.21	0.45	0.38	0.02	0.33	0.27
$\psi^2 = 0.4$	0.32	0.00	0.25	0.44	0.33	0.01	0.21	0.45	0.43	0.01	0.38	0.18
$\psi^2 = 0.6$	0.32	0.00	0.21	0.47	0.33	0.01	0.20	0.45	0.48	0.02	0.38	0.12
$N = 200$	0.32	0.00	0.36	0.32	0.38	0.02	0.29	0.32	0.56	0.02	0.32	0.10
$N = 600$	0.32	0.00	0.19	0.49	0.31	0.01	0.20	0.49	0.39	0.01	0.39	0.21
$N = 1000$	0.31	0.00	0.15	0.53	0.31	0.00	0.14	0.55	0.34	0.01	0.38	0.27
$T = 4$	0.32	0.00	0.44	0.24	0.36	0.01	0.40	0.24	0.48	0.02	0.45	0.05
$T = 6$	0.31	0.00	0.18	0.50	0.32	0.01	0.16	0.51	0.43	0.01	0.38	0.18
$T = 8$	0.31	0.00	0.08	0.60	0.31	0.01	0.07	0.60	0.37	0.02	0.26	0.35

"both non-sig" indicates that both models showed non-significant estimates for cross-lagged relations.

"RI-CLPM only" indicates that only the RI-CLPM showed significant estimates for cross-lagged relations.

"CLPM only" indicates that only the CLPM showed significant estimates for cross-lagged relations.

"both sig" indicates that both models showed significant estimates for cross-lagged relations.

https://doi.org/10.1371/journal.pone.0209133.t005

Table 6. Marginal means of proportions that models suggest consistent/inconsistent conclusions about reciprocal relations (CLPM vs STARTS).

	Data generation model											
	CLPM				RI-CLPM				STARTS			
Analysis model	both non-sig	STARTS only	RI-CLPM only	both sig	both non-sig	STARTS only	RI-CLPM only	both sig	both non-sig	STARTS only	RI-CLPM only	both sig
$\gamma = 0$	0.82	0.13	0.04	0.01	0.86	0.07	0.07	0.01	0.94	0.01	0.05	0.00
$\gamma = 0.1$	0.01	0.00	0.77	0.22	0.08	0.01	0.81	0.09	0.32	0.00	0.67	0.01
$\gamma = 0.2$	0.00	0.00	0.79	0.21	0.00	0.00	0.91	0.09	0.06	0.00	0.93	0.01
$\beta = 0.5$	0.30	0.02	0.44	0.24	0.32	0.03	0.53	0.12	0.49	0.00	0.49	0.01
$\beta = 0.7$	0.25	0.07	0.62	0.06	0.31	0.02	0.66	0.01	0.39	0.00	0.61	0.00
$\psi^2 = 0.2$	0.23	0.08	0.42	0.26	0.28	0.06	0.52	0.14	0.39	0.01	0.59	0.02
$\psi^2 = 0.4$	0.28	0.03	0.55	0.13	0.32	0.02	0.62	0.04	0.44	0.00	0.56	0.00
$\psi^2 = 0.6$	0.31	0.01	0.62	0.06	0.34	0.00	0.64	0.02	0.50	0.00	0.50	0.00
$N = 200$	0.29	0.03	0.57	0.11	0.37	0.03	0.55	0.05	0.58	0.00	0.42	0.00
$N = 600$	0.27	0.05	0.52	0.16	0.29	0.03	0.62	0.07	0.40	0.00	0.59	0.01
$N = 1000$	0.26	0.06	0.51	0.17	0.28	0.03	0.62	0.08	0.35	0.00	0.64	0.01
$T = 4$	0.32	0.00	0.65	0.02	0.36	0.00	0.63	0.01	0.50	0.00	0.50	0.00
$T = 6$	0.26	0.06	0.50	0.18	0.30	0.04	0.58	0.08	0.44	0.00	0.55	0.00
$T = 8$	0.24	0.08	0.44	0.24	0.28	0.04	0.58	0.10	0.38	0.00	0.60	0.01

"both non-sig" indicates that both models showed non-significant estimates for cross-lagged relations.

"STARTS only" indicates that only the STARTS showed significant estimates for cross-lagged relations.

"CLPM only" indicates that only the CLPM showed significant estimates for cross-lagged relations.

"both sig" indicates that both models showed significant estimates for cross-lagged relations.

https://doi.org/10.1371/journal.pone.0209133.t006

Table 7. Marginal means of proportions that models suggest consistent/inconsistent conclusions about reciprocal relations (RI-CLPM vs STARTS).

	Data generation model											
	CLPM				RI-CLPM				STARTS			
Analysis model	both	STARTS only	RI-CLPM only	both sig	both	STARTS only	RI-CLPM only	both sig	both	STARTS only	RI-CLPM only	both sig
$\gamma = 0$	0.84	0.14	0.01	0.01	0.91	0.08	0.01	0.00	0.96	0.01	0.04	0.00
$\gamma = 0.1$	0.44	0.03	0.34	0.19	0.44	0.01	0.45	0.10	0.82	0.00	0.17	0.01
$\gamma = 0.2$	0.19	0.00	0.59	0.21	0.18	0.00	0.73	0.09	0.59	0.00	0.40	0.01
$\beta = 0.5$	0.44	0.03	0.30	0.23	0.42	0.03	0.43	0.12	0.74	0.00	0.24	0.01
$\beta = 0.7$	0.54	0.08	0.33	0.04	0.60	0.02	0.37	0.01	0.84	0.00	0.16	0.00
$\psi^2 = 0.2$	0.45	0.11	0.20	0.24	0.47	0.06	0.33	0.14	0.70	0.01	0.27	0.02
$\psi^2 = 0.4$	0.52	0.04	0.31	0.12	0.53	0.02	0.41	0.04	0.80	0.00	0.20	0.00
$\psi^2 = 0.6$	0.51	0.02	0.43	0.05	0.53	0.01	0.45	0.01	0.86	0.00	0.14	0.00
$N = 200$	0.63	0.05	0.23	0.09	0.64	0.03	0.29	0.05	0.88	0.00	0.12	0.00
$N = 600$	0.45	0.06	0.34	0.15	0.48	0.03	0.43	0.07	0.78	0.00	0.21	0.01
$N = 1000$	0.40	0.06	0.37	0.16	0.42	0.03	0.48	0.07	0.71	0.00	0.28	0.01
$T = 4$	0.75	0.00	0.22	0.02	0.75	0.00	0.23	0.01	0.94	0.00	0.06	0.00
$T = 6$	0.43	0.07	0.34	0.16	0.44	0.04	0.44	0.08	0.81	0.00	0.19	0.00
$T = 8$	0.30	0.10	0.38	0.22	0.34	0.05	0.52	0.10	0.62	0.01	0.36	0.01

"both non-sig" indicates that both models showed non-significant estimates for cross-lagged relations.

"STARTS only" indicates that only the STARTS showed significant estimates for cross-lagged relations.

"RI-CLPM only" indicates that only the RI-CLPM showed significant estimates for cross-lagged relations.

"both sig" indicates that both models showed significant estimates for cross-lagged relations.

https://doi.org/10.1371/journal.pone.0209133.t007

Table 8. Marginal means of the proportions of models preferred by Akaike Information Criterion.

	Data generation model								
	CLPM			RI-CLPM			STARTS		
Analysis model	CLPM	RI-CLPM	STARTS	CLPM	RI-CLPM	STARTS	CLPM	RI-CLPM	STARTS
$\gamma = 0$	0.91	0.08	0.00	0.46	0.53	0.00	0.00	0.98	0.02
$\gamma = 0.1$	0.93	0.07	0.00	0.42	0.58	0.00	0.00	0.98	0.02
$\gamma = 0.2$	0.95	0.04	0.00	0.41	0.59	0.00	0.00	0.99	0.01
$\beta = 0.5$	0.92	0.07	0.01	0.09	0.91	0.00	0.00	0.98	0.02
$\beta = 0.7$	0.94	0.06	0.00	0.78	0.22	0.00	0.00	0.99	0.01
$\psi^2 = 0.2$	0.93	0.06	0.01	0.35	0.65	0.00	0.00	0.99	0.00
$\psi^2 = 0.4$	0.93	0.06	0.00	0.44	0.56	0.00	0.00	0.99	0.01
$\psi^2 = 0.6$	0.92	0.08	0.00	0.50	0.50	0.00	0.00	0.97	0.03
$N = 200$	0.94	0.06	0.00	0.46	0.54	0.00	0.00	0.98	0.01
$N = 600$	0.93	0.07	0.00	0.42	0.58	0.00	0.00	0.98	0.02
$N = 1000$	0.93	0.07	0.00	0.41	0.59	0.00	0.00	0.98	0.02
$T = 4$	1.00	0.00	0.00	0.55	0.45	0.00	0.00	1.00	0.00
$T = 6$	0.94	0.06	0.00	0.40	0.59	0.00	0.00	0.99	0.01
$T = 8$	0.86	0.14	0.01	0.34	0.66	0.00	0.00	0.97	0.03

https://doi.org/10.1371/journal.pone.0209133.t008

trait factor variances and measurement error variances, and as in the previous simulation, this may have contributed to the results that simpler models were favored. For example, a larger size of β indicates relative smaller impact of trait factor variances over time, which might have made parameter estimates somewhat unstable.

Note that when the true model is either the RI-CLPM or the STARTS model, (mostly, negative) biased point estimates were observed even when models were correctly specified. Marginal means of (standardized) point estimates and corresponding biases for different data generation models and analysis models are provided in Table B in S1 File. Although we have to take care about possible biased results here as a consequence of discarding the results when improper solutions were produced, in 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.

Tables 8 and 9 shows the marginal means of the proportions of models preferred by information criteria (Akaike Information Criterion: AIC, and Bayesian Information Criterion: BIC) under different data generation models and analysis models.

With both AIC and BIC, when the true model was the STARTS model, the RI-CLPM was preferred in most of the cases. When the true model was the RI-CLPM, the CLPM was often preferred. It should be noted, however, again that there may be a bias in the results as we discarded the results with improper solutions and our simulations used relatively small values for trait factor variances and measurement error variances.

General discussion

In this manuscript, we discussed the importance of considering alternative models such as the RI-CLPM and the STARTS model to infer reciprocal effects, and presented potential problems of applying commonly-used CLPM (specifically, the conflation of between-person and within-person effects). 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

Table 9. Marginal means of the proportions of models preferred by Bayesian Information Criterion.

	Data generation model								
	CLPM			RI-CLPM			STARTS		
Analysis model	CLPM	RI-CLPM	STARTS	CLPM	RI-CLPM	STARTS	CLPM	RI-CLPM	STARTS
$\gamma = 0$	0.93	0.06	0.00	0.53	0.47	0.00	0.01	0.98	0.01
$\gamma = 0.1$	0.95	0.05	0.00	0.49	0.51	0.00	0.00	0.98	0.01
$\gamma = 0.2$	0.96	0.04	0.00	0.46	0.54	0.00	0.00	0.99	0.01
$\beta = 0.5$	0.94	0.05	0.00	0.15	0.85	0.00	0.00	0.98	0.02
$\beta = 0.7$	0.95	0.05	0.00	0.83	0.17	0.00	0.00	0.99	0.01
$\psi^2 = 0.2$	0.95	0.05	0.00	0.41	0.59	0.00	0.00	0.99	0.00
$\psi^2 = 0.4$	0.95	0.05	0.00	0.50	0.50	0.00	0.00	0.99	0.01
$\psi^2 = 0.6$	0.94	0.06	0.00	0.57	0.43	0.00	0.01	0.96	0.03
$N = 200$	0.96	0.04	0.00	0.57	0.43	0.00	0.01	0.98	0.01
$N = 600$	0.94	0.06	0.00	0.47	0.53	0.00	0.00	0.98	0.01
$N = 1000$	0.94	0.06	0.00	0.44	0.56	0.00	0.00	0.98	0.01
$T = 4$	1.00	0.00	0.00	0.66	0.34	0.00	0.01	0.99	0.00
$T = 6$	0.96	0.04	0.00	0.45	0.55	0.00	0.00	0.99	0.01
$T = 8$	0.88	0.11	0.00	0.37	0.63	0.00	0.00	0.97	0.03

https://doi.org/10.1371/journal.pone.0209133.t009

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, even if the time lag between occasions is set to be optimal to effectively capture the theoretical process (see Dormann & Griffin [21] on this point). 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.

One potential limitation of the alternative models is the large number of improper solutions observed in our study. Although we acknowledge that large number of improper solutions might be caused by the specific true parameter values used in our simulations, our results indicate that, when researcher encounters improper solutions in applying the RI-CLPM and the STARTS model, this might suggest the possibility of model misspecification. This is especially the case in applying the RI-CLPM, because in this model, a dominant factor that caused improper solutions was model misspecification (Table 3). However, we also observed that alternative models produce improper solutions even when researchers do not misspecify the true model. Future studies should examine how this is caused and effective ways to address the problem.

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 [22]; Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker [23]; they both used the framework of multilevel vector autoregression model). 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 [24]; Ten Have & Joffe [25]; VanderWeele [26]), and they are often assessed in a longitudinal design (e.g., Huang & Yuan [27]; Preacher [28]). The issue of the current paper applies especially to longitudinal mediation models that include cross-lagged relations (e.g., a dynamic autoregressive mediation model; Maxwell, Cole & Mitchell [29]). 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.

Supporting information

S1 File.
(PDF)

Author Contributions

Conceptualization: Satoshi Usami.

Data curation: Satoshi Usami, Naoya Todo.

Formal analysis: Satoshi Usami, Naoya Todo.

Funding acquisition: Satoshi Usami.

Methodology: Satoshi Usami.

Writing – original draft: Satoshi Usami.

Writing – review & editing: Satoshi Usami, Kou Murayama.

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. https://doi.org/10.1037/h0027876
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. https://doi.org/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 PMID: 25822208
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. https://doi.org/10.1146/annurev.psych.093008.100356 PMID: 19575624
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. https://doi.org/10.1080/15427600902911189
10. Kenny DA, Zautra A. The trait-state-error model for multiwave data. *J. Consult. Clin. Psychol*. 1995; 63:52–59. https://doi.org/10.1037//0022-006x.63.1.52 PMID: 7896990
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 longitudinal models to examine reciprocal relations. *Psychol. Methods* in press. https://doi.org/10.1037/met0000210 PMID: 30998041
13. Granger CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 1969; 37:424–438. https://doi.org/10.2307/1912791
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. https://doi.org/10.1037/1082-989X.10.1.3 PMID: 15810866
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. https://doi.org/10.1016/j.jrp.2011.01.004
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. https://doi.org/10.1037/met0000155 PMID: 29172612
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. https://doi.org/10.2307/2087176
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. Dormann C. & Griffin M. Optimal time lags in panel studies. *Psychol. Methods* 2015; 20:489–505. https://doi.org/10.1037/met0000041 PMID: 26322999
22. Bringmann LF, et al. A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*. 2013; 8:e60188. https://doi.org/10.1371/journal.pone.0060188 PMID: 23593171
23. 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. https://doi.org/10.1037/met0000062 PMID: 27045851
24. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int. J. Epidemiol*. 2013; 42:1511–1519. https://doi.org/10.1093/ije/dyt127 PMID: 24019424
25. Ten Have TR, Joffe MM. A review of causal estimation of effects in mediation analyses. *Stat. Methods Med. Res*. 2012; 21:77–107. https://doi.org/10.1177/0962280210391076 PMID: 21163849
26. VanderWeele TJ. Mediation analysis: A practitioner's guide. *Annu. Rev. Public Health* 2016; 37:17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402 PMID: 26653405
27. Huang J, Yuan Y. Bayesian dynamic mediation analysis. *Psychol. Methods* 2017; 22:667–686. https://doi.org/10.1037/met0000073 PMID: 27123750
28. Preacher K. Advances in mediation analysis: a survey and synthesis of new developments. *Annu. Rev. Psychol*. 2015; 66:825–852. https://doi.org/10.1146/annurev-psych-010814-015258 PMID: 25148853
29. 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. https://doi.org/10.1080/00273171.2011.606716 PMID: 26736047