Analysing matched continuous longitudinal data: A review

Peer-reviewed author version

Delporte, Margaux; AERTS, Marc; VERBEKE, Geert & MOLENBERGHS, Geert (2024) Analysing matched continuous longitudinal data: A review. In: Statistical methods in medical research,.

DOI: 10.1177/09622802241300823
Handle: http://hdl.handle.net/1942/45043

Analysing matched continuous longitudinal data: A review

Abstract

Longitudinal data is frequently encountered in medical research, where participants are followed throughout the time. Additional structure and hence complexity occurs when there there is pairing in the participants (e.g., matched case-control studies) or within the participants (e.g., analysis of participants' both eyes). Various modelling approaches, identified through a systematic review, are discussed, including (un)paired t-tests, multivariate analysis of variance (MANOVA), difference scores, linear mixed models (LMM), and new statistical methods. Next, highlighting the importance of selecting appropriate models based on the data's characteristics, the methods are applied on both a real-life case study in ophthalmology and a simulated case-control study. Key findings include the superiority of the conditional linear mixed model and multilevel models in handling paired longitudinal data in terms of precision. Moreover, the article underscores the impact of accounting for intra-pair correlations and missing data mechanisms. Focus will be on discussing the advantages and disadvantages of the approaches, rather than on the mathematical or computational details.

Some Keywords: longitudinal data; paired data; random effects model

1 Introduction

Longitudinal studies are fundamental in medical research, providing valuable insights into the progression of diseases, treatment effectiveness, and patient outcomes over time. Longitudinal data, where measurements are collected on the same subjects at multiple time points, offer the opportunity to investigate within-subject changes while controlling for inter-subject variability. Clearly, it should be taken into account that the measurements from different subjects are independent, while observations within subjects are correlated. Due to this dependence, traditional techniques like classical (generalised) linear regression models are not suitable. Still, a large number of approaches to model longitudinal data have been developed and implemented in standard statistical software (Verbeke and Molenberghs, 2000).

In practice, the subjects are not always independent, as there can be a meaningful one-to-one relationship between them. This is evident in various scenarios, such as case-control studies, where each control is carefully matched to a case based on multiple attributes to minimise confounding. Twin studies, widely utilised in both biomedical and psychological research, serve as another example, aiming to account for genetic influences. Additionally, the pairing can also be present within individuals, such as, for example, hearing thresholds measured on both ears of a set of subjects. In, for example, ophthalmology research, a case can serve as its own control when treatment is administered to only one eye. Leveraging these intra-pair correlations within pairs can yield more precise estimates of the

effect under investigation. However, our literature study found that this is often not the case, and the pairing feature is ignored.

In longitudinal studies, missing data poses additional challenges to statistical analysis and inference. The nature of longitudinal data collection, spanning multiple time points, heightens the probability of missingness due to, amongst others, participant dropout. Addressing missing data is vital for maintaining the validity of study findings. Various techniques, such as multiple imputation methods, likelihood-based or Bayesian approaches, and modelling strategies accounting for missing data mechanisms, have been developed to handle this issue (Molenberghs and Kenward, 2007). However, selecting an appropriate method requires careful consideration of the underlying missing data mechanism. When data is Missing Completely at Random (MCAR), the missingness is independent of both observed and unobserved outcomes (Rubin, 1976). When missingness is linked to observed data, it falls under the category of Missing at Random (MAR). Conversely, if it is further associated with unobserved data, it is labelled as Missing Not at Random (MNAR). Ignoring missing data or applying inadequate handling techniques can lead to biased estimates.

In Section 2, a motivating real-life dataset, the Ophthalmology data, is introduced. Section 3 delves into the modelling techniques identified through a literature review. These approaches are then contrasted using the Ophthalmology data in Section 4 and a simulated case-control datasets in Section 5. Finally, Section 6 presents concluding thoughts.

2 Ophtomology data

The dataset at hand, provided by the DRCR Retina Network, was collected in the context of a clinical trial comparing the efficacy and safety of three treatments for central-involved Diabetic Macular Edema (DME).The study spanned 2 years with four-weekly follow-up visits. In the original study, 660 eyes were included, but our analysis was restricted to the 497 patients who had DME in one eye, but not in the other. Our research question involves comparing the evolution of the visual acuity in the eye exhibiting DME with the unaffected eye, taking into account both the correlation induced by the repeated measurements as well as the correlation due to the pairing of eyes within a subject. The mean visual acuity in our dataset over time is depicted in Figure 1, at this point ignoring both correlations.

3 Modelling approaches

To identify the methods that were used in the literature for the analysis of paired longitudinal data, a systematic review was conducted, which resulted in 56 articles that employed various methods. These methods are grouped in different categories, which are discussed in the sections below. A more detailed overview of the methodology and the results of the systematic review can be found in Appendix A.

In the remainder of the paper the notation will be as follows: the measurements of subject j of pair i at time k can be defined as y_{ijk} with $i = 1, \ldots, N/2$, $j = 1, 2$ and $k = 1, \ldots, n_{ij}$.

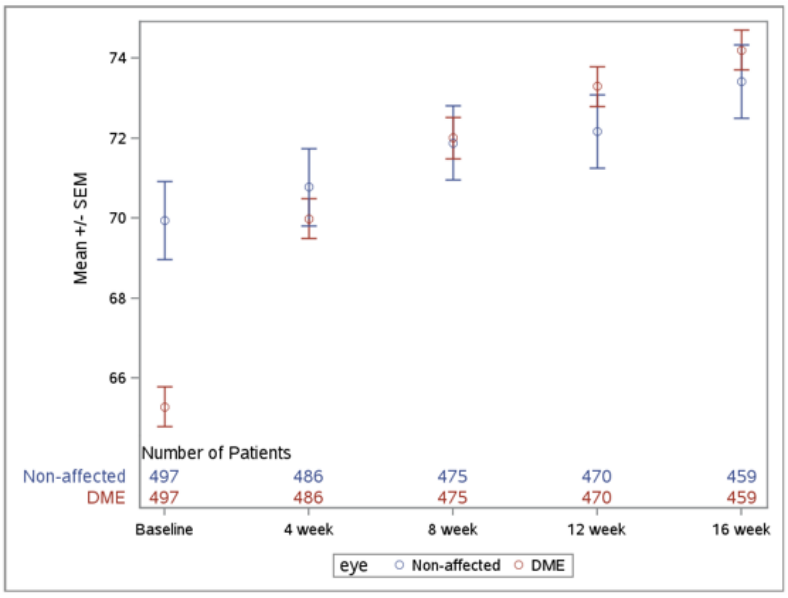

Figure 1: Mean visual acuity over time

3.1 Paired t-tests

A first approach is to employ paired t-tests or the non-parametric Wilcoxon rank sum tests to compare the pairs at specific time points. In the paired t-test, differences are used: $w_{ik} = y_{i1k} - y_{i2k}$, which result in a single longitudinal sequence. Next, the one-sample t-test is performed at a single time point i:

$$\frac{\bar{W}_k - \mu_{w_k}}{s_{w_k}}$$

follows a Student's t-distribution with $N/2 - 1$ degrees of freedom when the differences w_k can be considered to be normally distributed and:

$$\bar{W}_k = \frac{\sum_i W_{ik}}{N/2}, \qquad s^2_{w_k} = \frac{\frac{\sum_i (W_{ik} - \bar{W}_j)^2}{N/2-1}}{N/2}$$

Eight studies used this approach due to its benefits in terms of a straightforward interpretation, simplicity, and use of all available data. The method has, however, considerable disadvantages: it does not consider 'overall' differences across all time points, and more importantly, it does not allow to study differences in evolution. In addition, corrections for multiple testing have to be applied to keep the Type I error at bay.

3.2 Unpaired t-tests

The independent samples t-test, or unpaired t-test can detect whether two independent groups have a different mean at a specific time point k. Under the assumption of equality of variances and normality of Y_{i1k} and Y_{i2k}, the test statistics $\frac{\bar{Y}_{i1k} - \bar{Y}_{i2k}}{s_p \sqrt{\frac{2}{N}}}$ follows a Student's t-distribution with $2N - 2$ degrees of freedom, where $s_p = \sqrt{\frac{s^2_{y_{i1k}} - s^2_{y_{i2k}}}{2}}$.

Four studies (Shih et al., 2019; Mok et al., 2023; Dayal et al., 2017; Rebibo et al., 2013) performed an unpaired t-test or the non-parametric Mann-Whitney U-test to compare pairs at specific time points. Notably, in each study there was a 1:1 matching of a cases and controls, as opposed to pre-existing pairs such as siblings. However, when the pairs are positively correlated, the paired t-test has more power. Conditional on the null hypothesis being false, the size of this correlation is positively associated with the amount of statistical power. In addition, the unpaired t-test suffers from the same drawbacks as the paired t-test: it is not possible to test for 'overall' differences, nor differences in evolution. Similarly to the paired t-test,a correction for multiple testing should be administered.

3.3 Multivariate analysis of variance

Multivariate analysis of variance (MANOVA) is the multivariate extension of one-way analysis of variance (ANOVA). MANOVA is a statistical method that examines the effect of one or multiple factors on several dependent variables simultaneously. It allows for the assessment of null hypotheses regarding the effects of factor variables on the means of different groupings of dependent variables. In our scenario specifically, it can be used to test for pair (group) differences for the response at all time points simultaneously.

Beutel et al. (1996) used this methodology. While 27% of the dyads had missing data, they only included the complete cases in the MANOVA analysis. Since the resulting conclusions are only valid under the strict assumption of MCAR, extra steps such as multiple imputation should be implemented. A second disadvantage is that while this method can test for 'overall' differences, no conclusions can be made about evolution.

3.4 Difference scores

A fourth approach, adopted by three studies, is to calculate difference scores within the subjects between two specific time points. Let the two time points be a and b and the new difference scores $z_i = y_{i1a} - y_{i1b}$ and $v_i = y_{i2a} - y_{i2b}$. Ruhdorfer et al. (2015) administered a paired t-test on z_i and v_i to draw inference on differences between the pairs, while Goodman and Must (2011) used the non-parametric Wilcoxon test. This method allows studying differences in the evolution of the paired groups. Still, choices have to be made about which interval to use: Goodman and Must (2011) only considered the difference from baseline to the last timepoint, discarding data from intermediate time points. In contrast, Goodman and Must (2011) compared multiple pairs of difference scores between subsequent time points, necessitating correction for multiple testing.
The third study (LoCascio et al., 1998) calculated differences between baseline and each follow-up measurement for each patient, but later ignored both the longitudinal and paired nature of the data by applying ANCOVA on all difference scores simultaneously.

Closely related to the subject-specific difference scores is the calculation of a summary measure of the evolution for each subject and subsequently comparing these for the paired groups. (Schlee et al., 2021) first calculated slopes for each subject via linear regression and then calculated the Wilcoxon rank-sum test to test if the distributions of the slope estimates are equal in the study group and the matched control group. An advantage of this method is that each subject with more than one measurement can be included in the analysis, and it does not necessitate regular measurement intervals. In addition, no multiple testing issues arise. Still, this method treats the slopes as observed values and does not take the standard errors of the slopes into account. As a consequence, the

standard errors and the corresponding p-values of this method are incorrect.

3.5 Linear mixed models

Linear mixed models were introduced by Laird and Ware (1982) for the analysis of clustered continuous responses. Using the notation introduced at the onset of this section and ignoring the existence of pairs, let Y_{ijk} denote the kth measurement of subject i of pair j. The mixed model of a longitudinal sequence is specified as:

$$\begin{aligned} \boldsymbol{y_{ij}}|\boldsymbol{b}_{ij} &\sim N(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{b}_{ij}, \boldsymbol{\Sigma}_{ij}), \\ \boldsymbol{b}_{ij} &\sim N(\boldsymbol{0}, \boldsymbol{D}), \end{aligned} \tag{1}$$

where $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ denote, respectively, $(n_{ij} \times p)$ and $(n_{ij} \times q)$ dimensional matrices of known covariates. $\boldsymbol{\beta}$ is the p-dimensional vector containing fixed effects and $\boldsymbol{b}_{ij}$ denotes the q-dimensional vector of random effects. Finally, $\boldsymbol{\Sigma}_{ij}$ equals the $(n_{ij} \times n_{ij})$-dimensional residual covariance matrix and $\boldsymbol{D}$ denotes the variance-covariance matrix of the random effects. The marginal density of $\boldsymbol{Y_{ij}}$ equals

$$\begin{aligned} \boldsymbol{y_{ij}} &= \int f(\boldsymbol{y_{ij}}|\boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) d\boldsymbol{b}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*_{ij}, \\ \boldsymbol{\epsilon}^*_{ij} &\sim N(\boldsymbol{0}, \boldsymbol{V}_{ij}), \end{aligned} \tag{2}$$

where the covariance matrix $\boldsymbol{V}_{ij} = \boldsymbol{Z}_{ij}\boldsymbol{D}\boldsymbol{Z}'_{ij} + \boldsymbol{\Sigma}_{ij}$.

One key benefit of employing (generalized) linear mixed models lies in their fully parametric nature, enabling the use of both maximum likelihood and Bayesian estimation. This implies ignorability when data are incomplete, as outlined in Rubin (1976), under the assumption of Missingness at Random and mild regularity conditions. In essence, this means that the inferences drawn from a linear mixed model remain valid even in scenarios where missing data is dependent upon observed data, as long as the missingness is further to that independent of unobserved data. Still, as we will discuss in Section 3.5.3, ignorability does not hold when direct likelihood is not used, which is for example the case when the robust variance, or 'sandwich,' estimator is applied.

The literature review identified linear mixed models as the predominant method; the method was used in 37 of the 56 included studies. However, it should be noted that the random effects structure was unclear in Bouwmans et al. (2015). The different strategies found in the literature review are discussed below.

3.5.1 Random subject effect

24 studies used exclusively random effect on the level of the subject: $\boldsymbol{b}_{ij}$. Next, a fixed effect of group membership, and an interaction between group and time was used to assess the impact of being in the case or the paired control group.

However, two comments should be made. Firstly, all of these studies stated they 'matched' participants in the case and control group. However, in the literature study 'matching' was sometimes used to indicate that the distributions of age and gender were alike, instead of case-by-case matching of individual participants based on several attributes. Still, some of the studies in this category

described a 1:1 matching process (e.g. Iaffaldano et al. (2021)), and subsequently did not take into account the pairing. Secondly, six of these studies indicated that they used repeated measures ANOVA. While very similar to a linear mixed model, repeated measures ANOVA assumes a common set of time points or a time schedule and time is regarded as a factor with n levels, with subjects as subplots (Krueger and Tian, 2004). As a consequence, mixed models are superior to the repeated measures ANOVA in handling multiple missing data points.

The primary disadvantage is that the method does not take into account the correlation induced by the pairing and hence assumes that the pairs are independent. As a consequence, the standard errors of this method are incorrect. Still, the parameters estimates will be unbiased since the fixed effect estimates are independent of the chosen variance-covariance structure of the random effects (Lange and Ryan, 1989).

3.5.2 Random pair effect

In three studies, the model exclusively contained random effects for pairs $\boldsymbol{b}_i$, omitting a random effect for individual members within each pair. A study conducted by Ahmed et al. (2018) took the form of a case-control study, while in another study (Shek and Dou, 2020) an inherent link was present between members of a pair. In this particular study, a child repeatedly completed two identical questionnaires about maternal control on the one hand, and paternal control on the other. However, this approach implicitly assumes that all the measures of the dyad are independent, given the random effect of pair.

In contrast, a case-control study conducted by Border et al. (2020) employed a random pair effect while also incorporating an autoregressive residual correlation structure to relax the conditional independence assumption. Still, changing the residual structure will affect the parameter estimates of the fixed effects, as will be discussed in Section 5.

3.5.3 Robust variance estimation

Another possibility found in the literature (Sibbel et al., 2016) was to use a random effect on the subject level, as described in Section 3.5.1, and combine this with a robust variance estimator to take into account the pairing. An asymptotically consistent estimator, the so-called sandwich estimator, described in Huber (1967), White (1980), and Liang and Zeger (1986) is the following:

$$(\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-}\left(\sum_{i=1}^{N}\boldsymbol{X}_i'\widehat{\boldsymbol{V}_i}^{-1}\widehat{\boldsymbol{\epsilon}_i}\widehat{\boldsymbol{\epsilon}_i}'\widehat{\boldsymbol{V}_i}^{-1}\boldsymbol{X}_i\right)(\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-},$$

where $\widehat{\boldsymbol{\epsilon}_i} = \boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$. The estimator is consistent when the mean is correctly specified in the model (Verbeke and Molenberghs, 2000). Hence, when interest lies in the estimation of average longitudinal evolution and the dataset is sufficiently large, the sandwich estimator is a solution with minimal effort. But importantly, when missing data is present, very strict assumptions have to be made regarding the underlying process of missingness in order to obtain valid conclusions regarding the fixed effect based on the sandwich estimator. More specifically, it is assumed that the missing data is missing completely at random (MCAR), which means that the missingness is independent on observed as well as unobserved data. A second drawback is that efficiency is gained if an appropriate

covariance model can be specified (Diggle and Zeger, 1994).

3.5.4 Marginal linear mixed model

A versatile method is employing a marginal linear mixed model, as denoted in (2), where the random effects are integrated out of the density of the hierarchical model. Here, the population mean is modelled via the mean structure, and the dependence in the data is modelled via the marginal positive-definite matrix $\boldsymbol{V}_i$. Note that this model is more flexible, since it only imposes positive-definiteness on $\boldsymbol{V}_i$, in contrast to the hierarchical model that needs positive definiteness of both $\boldsymbol{\Sigma}_i$ and $\boldsymbol{D}$ (Verbeke and Molenberghs, 2000). This method was utilised by Benestad et al. (2022), who incorporated an unstructured covariance matrix to account for the pairing and repeated nature of his matched case-control study.

3.5.5 Nested random effects

Six studies used nested random effects or so-called three-level multilevel models to analyse their paired longitudinal data. These models are described by (Fitzmaurice et al., 2004, Ch. 22) as follows:

$$_{ijk} = X_{ijk}\beta + Z^{(3)}_{ijk}b^{(3)}_k + Z^{(2)}_{ijk}b^{(2)}_{jk} + \epsilon_{ijk},$$

where $Z^{(3)}_{ijk}$ and $Z^{(2)}_{ijk}$ are the design matrices for the random effects at the level of the pair and the subject, respectively. Here, the notation of the superscript signals the levels at which the random effects vary. The model allows that the random effects are correlated within a given level, but assumes that there are no correlations between levels. In addition, ϵ_{ijk}, the random component at the lowest level, is assumed to be independent within their level, with variance σ^2.

3.5.6 Conditional linear mixed model

A last possible option in the linear mixed model family, is the conditional linear mixed model. This model was employed by one study in our literature review: Gerber et al. (2016) investigated the effect of early antibiotic exposure on weight in twins during the first 8 years of life. The researchers specifically chose twins who were discordant in their early-life antibiotic exposure. Their model predicted the difference in growth trajectories in twins (weight of the exposed twin minus the weight of the unexposed twin) with a linear mixed model. As a consequence, the fixed slope of time represents the effect of antibiotic exposure on the growth evolution. It is worth noting that this method is equivalent to the conditional linear mixed model, discussed in (Verbeke and Molenberghs, 2000, Ch. 13).

Verbeke and Molenberghs (2000) indicate as a main advantage of the conditional linear mixed model that inferences about the longitudinal effects can be made without making assumptions about the cross-sectional components. Ignoring the pairing, the model is as follows:

$$\boldsymbol{Y_i} = \mathbf{1}_{n_i}b^*_i + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \epsilon_{(1)i}, \tag{3}$$

where b^*_i represents the cross-sectional components and is considered a nuisance. The matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ and the vectors $\boldsymbol{b}_i$ and $\boldsymbol{\beta}$ are submatrices of their original counterparts in (1), after the deletion of the cross-sectional effects. In a conditional linear model, the model fitting proceeds in two

steps. First, there is conditioning on sufficient statistics for the nuisance parameters b_i^*. Second, the remaining parameters in the conditional density of $\boldsymbol{Y_i}$ given these sufficient statistics are estimated via (restricted) maximum likelihood. The details can be found in Verbeke and Molenberghs (2000).

Importantly, the conditional model can be obtained via first taking the difference between the measurements within a pair $z_{ik} = y_{i1k} - y_{i2k}$ and then employing a standard linear mixed model as shown in (1). As a result, the intercept can be interpreted as the baseline difference between the groups, while the slope of time denotes the treatment effect on the evolution.

3.6 New methods

Two studies described new statistical methods to use for paired longitudinal data. The study by Wilson (1979) focuses on examining individualised growth trajectories in longitudinal twin data using repeated measures ANOVA. The total variance is partitioned into various substantial sources of variance, and the magnitude of their effects are calculated. The author subsequently formulates a hypothesis test concerning twin concordance, exploring whether twins within each pair exhibit greater similarity than they do with twins from different pairs. This directly leads to the calculation of intraclass correlations, representing the concordance within pairs in the form of correlation coefficients.

A second paper by Kim (2006) is in the context of longitudinal ophthalmology data, where the paired eyes are assigned to different treatments. He constructs methods to test the hypothesis that an interaction exists between the treatment (eye-specific factor) and race (person-specific factor). Two methods are described: a large sample-based non-parametric test statistic and a non-parametric bootstrap test analogy. He compares the results of his methods with generalized estimating equations (GEE) with different working correlation structures.

4 Analysis of the ophthalmology data

In the previous section, several approaches to model paired continuous longitudinal data have been presented. To demonstrate how to choose the right approach for a specific scenario, we revisit the ophthalmology data presented in Section 2. Here, our main emphasis is on selecting the appropriate modelling approach rather than delving into the results and insights gained from the statistical analysis. The main research question is to study the impact of the medication on the evolution of visual acuity in eyes with diabetic macular edema (DME$=0$) and without diabetic macular edema (DME$=1$).

The results of the different approaches can be found in Table 1. Scrutinising the results of analyses of the complete dataset, it is clear that the research question cannot be answered by the paired or the unpaired t-test. Based on these tests, we can only conclude that the acuity is different at baseline, but the differences are no longer significant at subsequent time points. As expected, the standard errors of the paired t-test are considerably smaller compared to those of the unpaired t-test. MANOVA confirms these results and shows that there is an overall difference in visual acuity in the first five time points. Note that the latter results are only valid under the assumption of MCAR.

The remaining five methods can answer the research question at hand. For instance, based on the paired t-test on the slopes of the subject-specific regression models, the conclusion can be drawn that the visual acuity has a better progression in the DME eyes. However, as the slopes are treated as

Table 1: Analysis of the ophthalmology data.

		Full data			Complete cases		
Method	Parameter	Estimate	SE	p-value	Estimate	SE	p-value
Paired t-test	Baseline	-4.646	0.956	<.0001	-4.618	0.999	<.0001
	4 weeks	-0.788	0.962	0.413	-0.873	0.991	0.379
	8 weeks	0.126	0.926	0.892	0.251	0.957	0.793
	12 weeks	1.128	0.911	0.216	1.230	0.941	0.192
	16 weeks	0.787	0.899	0.382	0.906	0.916	0.323
Unpaired t-test	Baseline	-4.646	1.098	<.0001	-4.618	1.161	<.0001
	4 weeks	-0.788	1.088	0.469	-0.873	1.121	0.436
	8 weeks	0.126	1.062	0.905	0.251	1.086	0.817
	12 weeks	1.128	1.050	0.283	1.230	1.084	0.257
	16 weeks	0.787	1.042	0.450	0.906	1.060	0.393
MANOVA	Wilks lambda	0.905		<.0001	0.905		<.0001
Comparison slopes		0.005	0.002	0.004	0.046	0.005	<.0001
LMM naive	DME	-0.342	0.284	0.227	-3.097	0.593	<.0001
	time*DME	0.005	0.001	<.0001	0.043	0.008	<.0001
LMM sandwich	DME	-0.342	0.858	0.690	-3.097	0.993	0.002
	time*DME	0.005	0.001	<.0001	0.043	0.006	<.0001
LMM nested	DME	-0.363	0.982	0.656	-3.255	0.939	0.001
	time*DME	0.004	0.001	<.0001	0.046	0.004	<.0001
Conditional LMM	DME	-0.366	0.813	0.653	-3.256	0.935	0.001
	time*DME	0.004	<0.001	<.0001	0.046	0.004	<.0001

'observed' and the standard errors are not taken into account, the p-value of this analysis is incorrect. This is also true for the standard error of the 'naive' linear mixed model, where the pairing was not taken into account, and only a random intercept of the subject was included. In other words, it is assumed that all measurements of the eyes within a participant are independent, conditional on the person-specific random effect. Still, the estimated fixed effects are very similar, and in each analysis it can be concluded that when treated, there is a beneficial evolution in visual acuity in the eyes with DME. When comparing the standard errors of the three approaches that correctly take into account the pairing, they are the lowest in the conditional linear mixed model.

Next, in two columns on the right of Table 1, the analysis is repeated on a dataset restricted to the first five measurements of cases who have no missing data on these time points. It is apparent that the results of the MANOVA analysis are exactly equal, as by default only complete cases are considered in the analysis. All other methods take into account incomplete profiles, but differ in their assumptions with regards to the missingness. Specifically, the (un)paired t-test, MANOVA, and sandwich LMM assume missingness to be completely at random (MCAR), whereas the nested and conditional LMM uphold validity under the more lenient Missing at Random (MAR) assumption. In contrast, the standard errors with regards to the slope comparisons and the naive LMM are invalid.

Table 2: Average parameter estimates, average standard errors and standard deviation of the estimates of the treatment effect at baseline and the treatment effect on the evolution of 100 simulated datasets.

Method	Parameter	Avg. Estimate	Avg. SE	SD estimates
LMM naive	treated	-0.5372	0.4637	0.2100
	time*treated	-0.2411	0.1028	0.1714
LMM sandwich	treated	-0.5372	0.2967	0.2067
	time*treated	-0.2411	0.1710	0.1711
LMM marginal	treated	-0.5377	0.2956	0.2050
	time*treated	-0.2414	0.1708	0.1698
LMM nested	treated	-0.5372	0.2243	0.2067
	time*treated	-0.2411	0.1714	0.1711
LMM combination	treated	-0.5303	0.3380	0.2271
	time*treated	-0.2423	0.1404	0.1761
Conditional LMM	treated	-0.5372	0.2243	0.2067
	time*treated	-0.2411	0.1711	0.1711

5 Simulation study

To compare the performance of various methods across a wide array of standardized datasets, simulated data from case-control studies are employed. Specifically, we simulated 100 datasets, each comprising five measurements from 200 pairs of subjects. While the treatment effect remains fixed across datasets, the variances of both the random effects and the residuals were varying. The resulting estimates and standard errors from the different categories of linear mixed models are averaged and presented in Table 2.

Comparing the parameter estimates, it is clear that some are slightly different from the others. This is the case for both the marginal model and the 'combination' method, where a random effect of the pair is combined with an autoregressive correlation structure in the residual variance-covariance matrix. Previous studies (Lange and Ryan, 1989) showed that in the absence of missing data, fixed effects do not depend on the chosen variance-covariance structure of the random effects, but the same does not hold true for the variance-covariance structure of the residuals. While the naive, sandwich, nested, and conditional models assume that the residuals are uncorrelated given the random effects, this is not the case for the marginal and 'combination' model. In the marginal model, the residual variance-covariance matrix is unstructured, and in the combination model, autoregressive correlation is assumed.

Next, the differences in average standard errors of the treatment effect on the evolution are negligible, while larger differences exist in the average standard errors of the baseline effects. The smallest standard errors are found in the conditional LMM and the nested random effects model. Comparing the averaged SE to the standard deviation of the estimates, it is apparent that these estimates are also more accurate compared to the other models. Notably, the standard errors of the naive model are proven to be incorrect since they do not take the intra-pair correlation into account.

6 Concluding Remarks

In this research, we focus on the analysis of paired longitudinal data, where the pairing is either within the participant, or between the participants. First, a systematic review has been conducted to identify the methods that are used in the literature. Next to showing the broad range of methods that are used for this kind of data, the systematic review demonstrated that most studies ignored the pairing, while it could be used in favour of getting more precise estimates.

We presented the various methods that emerged from the systematic review and discussed the possible research questions they can answer, along with their respective advantages and limitations. For instance, while MANOVA and (un)paired t-tests are suitable for comparing pairs at different time-points, they fall short in assessing differences in progression over time. The questions can be answered by linear mixed models, or alternatively, by deriving summary measures (like slopes) for comparison among groups. However, it is crucial to account for standard errors of the summary statistics in the analysis. Furthermore, in linear mixed models with only a random effect at the subject level, standard errors can be misleading since pairs are assumed to be independent. These nuances underscore the importance of selecting the appropriate modelling approach.

In addition, special attention has been given to missing data, which is frequently encountered in longitudinal studies. Some methods are only valid under the very restrictive assumption of Missing Completely at Random (MCAR), which assumes that the missingness does not depend under observed nor unobserved data. This is the case for MANOVA, as well as linear mixed models with the robust sandwich estimator. In contrast, in linear mixed models that do not employ the sandwich estimator, ignorability holds under the less restrictive Missing at Random assumption.

Following the methodological exploration, we applied these techniques to a real-life ophthalmology dataset, where both eyes of participants were examined concurrently. Interestingly, neither the (un)paired t-test nor MANOVA could effectively address the specific research question concerning the differences in evolution between eyes with and without diabetic macular edema. Moreover, we concluded that the standard errors in analyses comparing slopes via the paired t-test and the linear mixed model with only a random subject effect were flawed. Our analysis demonstrated that the most precise estimates are obtained via the conditional linear mixed model.

Next, we compared the linear-mixed model based techniques on 100 simulated datasets of a case-control study with identical treatment effects. Our analysis revealed slight disparities in parameter estimates attributable to variations in residual covariance structures. However, these differences proved inconsequential, as did variations in standard errors regarding the estimation of treatment effects on evolution. Notably, the standard errors of the treatment effect at baseline were the most accurately estimated when employing either the multilevel ('nested') linear mixed model or the conditional linear mixed model.

References

Aguilar-Mediavilla, E., Buil-Legaz, L., López-Penadés, R., Sanchez-Azanza, V., and Adrover-Roig, D. (2019). Academic outcomes in bilingual children with developmental language disorder: A longitudinal study. *Frontiers in Psychology*, 10.

Ahmed, B., King, W., Gourash, W., Belle, S., Hinerman, A., Pomp, A., Dakin, G., and Courcoulas, A. (2018). Long-term weight change and health outcomes for sleeve gastrectomy (sg) and matched

roux-en-y gastric bypass (rygb) participants in the longitudinal assessment of bariatric surgery (labs) study. *Surgery (United States)*, 164(4):774–783.

Alfieri, P., Scibelli, F., Montanaro, F., Digilio, M., Ravà, L., Valeri, G., and Vicari, S. (2022). Differences and similarities in adaptive functioning between children with autism spectrum disorder and williams–beuren syndrome: A longitudinal study. *Genes*, 13(7).

Ancoli-Israel, S., Liu, L., Rissling, M., Natarajan, L., Neikrug, A., Palmer, B., Mills, P., Parker, B., Sadler, G., and Maglione, J. (2014). Sleep, fatigue, depression, and circadian activity rhythms in women with breast cancer before and after treatment: A 1-year longitudinal study. *Supportive Care in Cancer*, 22(9):2535–2545.

Andreas, J., Fals-Stewart, W., and O'Farrell, T. (2006). Does individual treatment for alcoholic fathers benefit their children? a longitudinal assessment. *Journal of Consulting and Clinical Psychology*, 74(1):191–198.

Benestad, M., Drageset, J., Eide, G., Vollsæter, M., Halvorsen, T., and Vederhus, B. (2022). Development of health-related quality of life and subjective health complaints in adults born extremely preterm: a longitudinal cohort study. *Health and Quality of Life Outcomes*, 20(1).

Bergström, K., Klatte, M., Steinbrink, C., and Lachmann, T. (2016). First and second language acquisition in german children attending a kindergarten immersion program: A combined longitudinal and cross-sectional study. *Language Learning*, 66(2):386–418.

Beutel, M., Willner, H., Deckardt, R., Von Rad, M., and Weiner, H. (1996). Similarities and differences in couples' grief reactions following a miscarriage: Results from a longitudinal study. *Journal of Psychosomatic Research*, 40(3):245–253.

Border, W., Sachdeva, R., Stratton, K., Armenian, S., Bhat, A., Cox, D., Leger, K., Leisenring, W., Meacham, L., Sadak, K., Sivanandam, S., Nathan, P., and Chow, E. (2020). Longitudinal changes in echocardiographic parameters of cardiac function in pediatric cancer survivors. *JACC: CARDIOONCOLOGY*, 2(1):26–37.

Bouwmans, M., Bos, E., Booij, S., van Faassen, M., Oldehinkel, A., and de Jonge, P. (2015). Intra- and inter-individual variability of longitudinal daytime melatonin secretion patterns in depressed and non-depressed individuals. *CHRONOBIOLOGY INTERNATIONAL*, 32(3):441–446.

Büttner, F., Howell, D., Severini, G., Doherty, C., Blake, C., Ryan, J., and Delahunt, E. (2021). Using functional movement tests to investigate the presence of sensorimotor impairment in amateur athletes following sport-related concussion: A prospective, longitudinal study. *Physical Therapy in Sport*, 47:105–113.

Ceroni, D., Martin, X., Lamah, L., Delhumeau, C., Farpour-Lambert, N., De Coulon, G., and Ferrière, V. (2012). Recovery of physical activity levels in adolescents after lower limb fractures: A longitudinal, accelerometry-based activity monitor study. *BMC Musculoskeletal Disorders*, 13.

Dall, P., Ellis, S., Ellis, B., Grant, P., Colyer, A., Gee, N., Granat, M., and Mills, D. (2017). The influence of dog ownership on objective measures of free-living physical activity and sedentary behaviour in community-dwelling older adults: a longitudinal case-controlled study. *BMC PUBLIC HEALTH*, 17.

Dayal, D., Soni, V., Das, G., Bhunwal, S., Kaur, H., and Bhalla, A. (2017). Longitudinal observations on growth patterns of obese infants: Developing country perspectives. preliminary study. *Pediatria Polska*, 92(4):397–400.

De Paúl, J. and Domenech, L. (2000). Childhood history of abuse and child abuse potential in adolescent mothers: A longitudinal study. *Child Abuse and Neglect*, 24(5):701–713.

Diggle, P.J., L. K.-Y. and Zeger, S. (1994). *Analysis of Longitudinal Data.* Clarendon Press, Oxford.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied longitudinal analysis.* Wiley series in probability and statistics. Wiley, Hoboken.

Gardner, S. and Boellaard, R. (2007). Does youth relationship education continue to work after a high school class? a longitudinal study. *Family Relations*, 56(5):490–500.

Gerber, J., Bryan, M., Ross, R., Daymont, C., Parks, E., Localio, A., Grundmeier, R., Stallings, V., and Zaoutis, T. (2016). Antibiotic exposure during the first 6 months of life and weight gain during childhood. *JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION*, 315(12):1258–1265.

Goodman, E. and Must, A. (2011). Depressive symptoms in severely obese compared with normal weight adolescents: Results from a community-based longitudinal study. *JOURNAL OF ADOLESCENT HEALTH*, 49(1):64–69.

Gurucharri, C., Phelps, E., and Selman, R. (1984). Development of interpersonal understanding: A longitudinal and comparative study of normal and disturbed youths. *Journal of Consulting and Clinical Psychology*, 52(1):26–36.

Hancock, K., Craig, A., Dickson, H., Chang, E., and Martin, J. (1993). Anxiety and depression over the first year of spinal cord injury: A longitudinal study. *Paraplegia*, 31(6):349–357.

Hands, B. (2008). Changes in motor skill and fitness measures among children with high and low motor competence: A five-year longitudinal study. *Journal of Science and Medicine in Sport*, 11(2):155–162.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press*, 1:221–233.

Hull, D., Powell, M., Fagan, M., Hobbs, C., and Williams, L. (2020). Positive youth development: A longitudinal quasi-experiment in jamaica. *Journal of Applied Developmental Psychology*, 67.

Iaffaldano, P., Lucisano, G., Caputo, F., Paolicelli, D., Patti, F., Zaffaroni, M., Morra, V., Pozzilli, C., De Luca, G., Inglese, M., Salemi, G., Maniscalco, G., Cocco, E., Sola, P., Lus, G., Conte, A., Amato, M., Granella, F., Gasperini, C., Bellantonio, P., Totaro, R., Rovaris, M., Salvetti, M., Clerici, V., Bergamaschi, R., Maimone, D., Scarpini, E., Capobianco, M., Comi, G., Filippi, M., and Trojano, M. (2021). Long-term disability trajectories in relapsing multiple sclerosis patients treated with early intensive or escalation treatment strategies. *THERAPEUTIC ADVANCES IN NEUROLOGICAL DISORDERS*, 14.

Isberg, A., Ren, Y.-F., Henningsson, G., and Mcwilliam, J. (1993). Facial growth after pharyngeal flap surgery in cleft palate patients: A five-year longitudinal study. *Scandinavian Journal of Plastic and Reconstructive Surgery and Hand Surgery*, 27(2):119–126.

Keresztes, P., Merritt, S., Holm, K., Penckofer, S., and Patel, M. (2003). The coronary artery bypass experience: gender differences. *HEART LUNG*, 32(5):308–319.

Kim, J. (2006). Hypothesis testing problems in an unbalanced longitudinal ophthalmology study. *Communications in Statistics - Theory and Methods*, 35(3):461–476.

Kleinbub, J., Palmieri, A., Broggio, A., Pagnini, F., Benelli, E., Sambin, M., and Sorarù, G. (2015). Hypnosis-based psychodynamic treatment in als: A longitudinal study on patients and their caregivers. *Frontiers in Psychology*, 6:1–14.

Kretzschmar, M., Heilmeier, U., Yu, A., Joseph, G., Liu, F., Solka, M., McCulloch, C., Nevitt, M., and Link, T. (2016). Longitudinal analysis of cartilage t2 relaxation times and joint degeneration in african american and caucasian american women over an observation period of 6 years - data from the osteoarthritis initiative. *OSTEOARTHRITIS AND CARTILAGE*, 24(8):1384–1391.

Krueger, C. and Tian, L. (2004). A comparison of the general linear mixed model and repeated measures anova using a dataset with multiple missing data points. *Biological Research For Nursing*, 6(2):151–157. PMID: 15388912.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4).

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17(2):624–642.

Langer, S., Abrams, J., and Syrjala, K. (2003). Caregiver and patient marital satisfaction and affect following hematopoietic stem cell transplantation: A prospective, longitudinal investigation. *Psycho-Oncology*, 12(3):239–253.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liu, Y., Zhang, J., Chau, S., Yu, M., Chan, N., Chan, J., Li, S., Huang, B., Wang, J., Feng, H., Zhou, L., Mok, V., and Wing, Y. (2022). Evolution of prodromal rem sleep behavior disorder to neurodegeneration: A retrospective longitudinal case-control study. *Neurology*, 99(6):E627–E637.

LoCascio, V., Ballanti, P., Milani, S., Bertoldo, F., LoCascio, C., Zanolin, E., and Bonucci, E. (1998). A histomorphometric long-term longitudinal study of trabecular bone loss in glucocorticoid-treated patients: Prednisone versus deflazacort. *CALCIFIED TISSUE INTERNATIONAL*, 62(3):199–204.

Loher, S., Fatzer, S., and Roebers, C. (2014). Executive functions after pediatric mild traumatic brain injury: A prospective short-term longitudinal study. *Applied Neuropsychology: Child*, 3(2):103–114.

Luo, H., Qiu, L., Wu, Y., and Zhang, X. (2019). Growth in syphilis-exposed and -unexposed uninfected children from birth to 18 months of age in china: a longitudinal study. *Scientific Reports*, 9(1).

Magnusson, E. and Nauclér, K. (1990). Reading and spelling in language-disordered children linguistic and metalinguistic prerequisites: Report on a longitudinal study. *Clinical Linguistics and Phonetics*, 4(1):49–61.

Metallinou, D., Karampas, G., Lazarou, E., Iacovidou, N., Pervanidou, P., Lykeridou, K., Mastorakos, G., and Rizos, D. (2021). Serum activin a as brain injury biomarker in the first three days of life. a prospective case—control longitudinal study in human premature neonates. *Brain Sciences*, 11(9).

Mok, E., Kam, K., and Young, A. (2023). Corneal nerve changes in herpes zoster ophthalmicus: a prospective longitudinal in vivo confocal microscopy study. *Eye (Basingstoke)*, 37(14):3033–3040.

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies.* Wiley series in probability and statistics. Wiley, Hoboken.

Moyle, M., Weismer, S., Evans, J., and Lindstrom, M. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children. *Journal of Speech, Language, and Hearing Research*, 50(2):508–528.

Nickols-Richardson, S., O'Connor, P., Shapses, S., and Lewis, R. (1999). Longitudinal bone mineral density changes in female child artistic gymnasts. *Journal of Bone and Mineral Research*, 14(6):994–1002.

Oertel, F., Outteryck, O., Knier, B., Zimmermann, H., Borisow, N., Bellmann-Strobl, J., Blaschek, A., Jarius, S., Reindl, M., Ruprecht, K., Meinl, E., Hohlfeld, R., Paul, F., Brandt, A., Kümpfel, T., and Havla, J. (2019). Optical coherence tomography in myelin-oligodendrocyte-glycoprotein antibody-seropositive patients: a longitudinal study. *Journal of Neuroinflammation*, 16(1).

Oshima, Y., Sato, S., Chen-Yoshikawa, T., Yoshioka, Y., Shimamura, N., Hamada, R., Nankaku, M., Tamaki, A., Date, H., and Matsuda, S. (2020). Quantity and quality of antigravity muscles in patients undergoing living-donor lobar lung transplantation: 1-year longitudinal analysis using chest computed tomography images. *ERJ Open Research*, 6(2):1–11.

Peetsma, T., Vergeer, M., Roeleveld, J., and Karsten, S. (2001). Inclusion in education: comparing pupils' development in special and regular education. *EDUCATIONAL REVIEW*, 53(2):125–135.

Rebibo, L., Verhaeghe, P., Cosse, C., Dhahri, A., Maréchal, V., and Regimbeau, J.-M. (2013). Does longitudinal sleeve gastrectomy have a family "halo effect"? a case-matched study. *Surgical Endoscopy*, 27(5):1748–1753.

Roberts, H., Grant, M., Hubber, N., Super, P., Singhal, R., and Chapple, I. (2018). Impact of bariatric surgical intervention on peripheral blood neutrophil (pbn) function in obesity. *OBESITY SURGERY*, 28(6):1611–1621.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Ruhdorfer, A., Wirth, W., Dannhauer, T., and Eckstein, F. (2015). Longitudinal (4 year) change of thigh muscle and adipose tissue distribution in chronically painful vs painless knees - data from the osteoarthritis initiative. *Osteoarthritis and Cartilage*, 23(8):1348–1356.

Schlee, W., Simoes, J., and Pryss, R. (2021). Auricular acupressure combined with self-help intervention for treating chronic tinnitus: A longitudinal observational study. *Journal of Clinical Medicine*, 10(18).

Scholten-Peeters, G., Coppieters, M., Durge, T., and Castien, R. (2020). Fluctuations in local and widespread mechanical sensitivity throughout the migraine cycle: A prospective longitudinal study. *Journal of Headache and Pain*, 21(1).

Shek, D. and Dou, D. (2020). Perceived parenting and parent-child relational qualities in fathers and mothers: Longitudinal findings based on hong kong adolescents. *International Journal of Environmental Research and Public Health*, 17(11):1–20.

Shih, V., Banks, E., Bonine, N., Harrington, A., Stafkey-Mailey, D., Yue, B., Ye, J., Fuldeore, R., and Gillard, P. (2019). Healthcare resource utilization and costs among women diagnosed with uterine fibroids compared to women without uterine fibroids. *CURRENT MEDICAL RESEARCH AND OPINION*, 35(11):1925–1935.

Sibbel, S., Hunt, A., Laplante, S., Beck, W., Gellens, M., and Brunelli, S. (2016). Comparative effectiveness of dialyzers: A longitudinal, propensity score-matched study of incident hemodialysis patients. *ASAIO Journal*, 62(5):613–622.

Torgalsbøen, A.-K., Mohn, C., Larøi, F., Fu, S., and Czajkowski, N. (2023). A ten-year longitudinal repeated assessment study of cognitive improvement in patients with first-episode schizophrenia and healthy controls: The oslo schizophrenia recovery (osr) study. *Schizophrenia Research*, 260:92–98.

Vasunilashorn, S., Ngo, L., Inouye, S., Libermann, T., Jones, R., Alsop, D., Guess, J., Jastrzebski, S., McElhaney, J., Kuchel, G., and Marcantonio, E. (2015). Cytokines and postoperative delirium in older patients undergoing major elective surgery. *JOURNALS OF GERONTOLOGY SERIES A-BIOLOGICAL SCIENCES AND MEDICAL SCIENCES*, 70(10):1289–1295.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer New York.

Weiler, H., Paes, B., Shah, J., and Atkinson, S. (1997). Longitudinal assessment of growth and bone mineral accretion in prematurely born infants treated for chronic lung disease with dexamethasone. *Early Human Development*, 47(3):271–286.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48(4):817–838.

Wilson, R. (1979). Analysis of longitudinal twin data. basic model and applications to physical growth measures. *Acta Geneticae Medicae et Gemellologiae*, 28(2):93–105.

Yang, C.-H., Wang, S., Wang, W.-L., Belcher, B., and Dunton, G. (2022). Day-level associations of physical activity and sedentary time in mother–child dyads across three years: a multi-wave longitudinal study using accelerometers. *Journal of Behavioral Medicine*, 45(5):702–715.

Zhao, D., Kim, M., Pastor-Barriuso, R., Chang, Y., Ryu, S., Zhang, Y., Rampal, S., Shin, H., Kim, J., Friedman, D., Guallar, E., and Cho, J. (2014). A longitudinal study of age-related changes in intraocular pressure: The kangbuk samsung health study. *Investigative Ophthalmology and Visual Science*, 55(10):6244–6250.

Zorrilla-Vaca, A., Ramirez, P., Iniesta-Donate, M., Lasala, J., Wang, X., Williams, L., Meyer, L., and Mena, G. (2022). Opioid-sparing anesthesia and patient-reported outcomes after open gynecologic surgery: a historical cohort study. *CANADIAN JOURNAL OF ANESTHESIA-JOURNAL CANADIEN D ANESTHESIE*, 69(12):1477–1492.

Analysing Matched Continuous Longitudinal Data

Supplementary Materials

A Literature review

A.1 Data sources and searches

Computerized bibliographic databases Web of science, Pubmed and Scopus were used to identify studies. These databases were searched on October 18, 2023 and without limitations regarding to the year of publication. The search criteria were restricted to containing '*longitudinal*' in the title (or abstract in Web of Science) and '*matched*' and/or '*paired*' in the abstract.

A.2 Study selection

Publications were included in this systematic review if the following inclusion criteria were met: 1) the study is longitudinal and subjects have measurements on more than two time points. 2) the data is paired, which means that there is an obvious and meaningful one-to-one correspondence between subjects. This can also be the case when there is one-to-one (propensity score) matching. 3) the response is continuous. The yield of the database search were first screened based on title and abstract and next the full text of the selected articles was screened for relevance.

A.3 Data extraction

The included studies were grouped based on the various statistical methodologies employed. These categories were

1. Difference scores
2. Comparison of subject-specific slopes
3. Paired t-tests or non-parametric alternative
4. Unpaired t-tests or non-parametric alternative
5. Linear mixed models
    (a) Without consideration for the paired nature of the data
    (b) With consideration for the paired nature of the data
6. New methodology

Table 3: Number of studies per category of statistical method.

Method	N
Paired t-tests at specific timepoints	8
Difference scores	3
Subject-specific slopes	1
Unpaired t-tests or non-parametric alternative	4
MANOVA	1
Linear mixed models	37
Without consideration for the paired nature of the data	(24)
With consideration for the paired nature of the data	(12)
New methodology	2

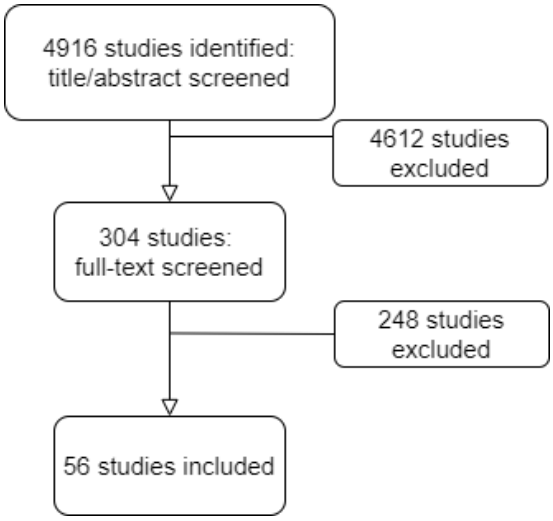

Figure 2: Flowchart of the literature search.

A.4 Results

The search strategy in the database identified 4916 potentially relevant studies (953 in Web of Science, 5 in Pubmed, and 4143 in Scopus). Based on the title and abstract, 304 articles appeared to meet the selection criteria, but after reading the full text of these studies, only 56 fulfilled the inclusion criteria (see flowchart in Fig. 2). The studies are grouped in different categories based on the used methods in Table 3. An article-by-article overview can be found in Table 4.

Authors (year)	Journal	Category
Aguilar-Mediavilla et al. (2019)	Frontiers in Psychology	LMM: Ignore pairing
Ahmed et al. (2018)	Surgery (United States)	LMM: Random effect pair
Alfieri et al. (2022)	Genes	LMM: Ignore pairing
Ancoli-Israel et al. (2014)	Supportive Care in Cancer	LMM: Ignore pairing
Andreas et al. (2006)	Journal of Consulting and Clinical Psychology	LMM: Ignore pairing
Benestad et al. (2022)	Health and Quality of Life Outcomes	LMM: Marginal
Bergström et al. (2016)	Language Learning	LMM: Ignore pairing
Beutel et al. (1996)	Journal of Psychosomatic Research	MANOVA
Border et al. (2020)	Jacc: Cardiooncology	LMM: Random effect pair
Bouwmans et al. (2015)	Chronobiology International	LMM: Unclear structure
Büttner et al. (2021)	Physical Therapy in Sport	LMM: Ignore pairing
Ceroni et al. (2012)	BMC Musculoskeletal Disorders	Paired t-test
Dall et al. (2017)	BMC Public Health	LMM: Nested
Dayal et al. (2017)	Pediatria Polska	Unpaired t-test
De Paúl and Domenech (2000)	Child Abuse and Neglect	LMM: Ignore pairing
Gardner and Boellaard (2007)	Family Relations	LMM: Ignore pairing
Gerber et al. (2016)	Journal Of The American Medical Association	LMM: Conditional
Goodman and Must (2011)	Journal Of Adolescent Health	Difference score
Gurucharri et al. (1984)	Journal of Consulting and Clinical Psychology	LMM: Ignore pairing
Hancock et al. (1993)	Paraplegia	LMM: Ignore pairing
Hands (2008)	Journal of Science and Medicine in Sport	LMM: Ignore pairing
Hull et al. (2020)	Journal of Applied Developmental Psychology	LMM: Ignore pairing
Iaffaldano et al. (2021)	Therapeutic Advances In Neurological Disorders	LMM: Ignore pairing
Isberg et al. (1993)	Scandinavian Journal of Plastic and Reconstructive Surgery and Hand Surgery	Paired t-test
Keresztes et al. (2003)	Heart & Lung	LMM: Ignore pairing
Kim (2006)	Communications in Statistics - Theory and Methods	New methodology
Kleinbub et al. (2015)	Frontiers in Psychology	LMM: Ignore pairing
Kretzschmar et al. (2016)	Osteoarthritis And Cartilage	LMM: Nested
Langer et al. (2003)	Psycho-Oncology	Paired t-test
Liu et al. (2022)	Neurology	LMM: Ignore pairing
LoCascio et al. (1998)	Calcified Tissue International	Difference score
Loher et al. (2014)	Applied Neuropsychology: Child	LMM: Ignore pairing
Luo et al. (2019)	Scientific Reports	LMM: Ignore pairing
Magnusson and Nauclér (1990)	Clinical Linguistics and Phonetics	Paired t-test
Metallinou et al. (2021)	Brain Sciences	Paired t-test
Mok et al. (2023)	Eye	Unpaired t-test
Moyle et al. (2007)	Journal of Speech Language and Hearing Research	LMM: Nested
Nickols-Richardson et al. (1999)	Journal of Bone and Mineral Research	LMM: Ignore pairing
Oertel et al. (2019)	Journal of Neuroinflammation	LMM: Nested
Oshima et al. (2020)	ERJ Open Research	LMM: Ignore pairing
Peetsma et al. (2001)	Educational Review	Paired t-test
Rebibo et al. (2013)	Surgical Endoscopy	Unpaired t-test
Roberts et al. (2018)	Obesity Surgery	Paired t-test
Ruhdorfer et al. (2015)	Osteoarthritis and Cartilage	Difference score
Schlee et al. (2021)	Journal of Clinical Medicine	Subject-specific slopes
Scholten-Peeters et al. (2020)	Journal of Headache and Pain	LMM: Ignore pairing
Shek and Dou (2020)	International Journal of Environmental Research and Public Health	LMM: Random effect pair
Shih et al. (2019)	Current Medical Research And Opinion	Unpaired t-test
Sibbel et al. (2016)	ASAIO Journal	LMM: Sandwich
Torgalsbøen et al. (2023)	Schizophrenia Research	LMM: Ignore pairing
Vasunilashorn et al. (2015)	Journals Of Gerontology Series A	Paired t-test
Weiler et al. (1997)	Early Human Development	LMM: Ignore pairing
Wilson (1979)	Acta Geneticae Medicae et Gemellologiae	New methodology
Yang et al. (2022)	Journal of Behavioral Medicine	LMM: Nested
Zhao et al. (2014)	Investigative Ophthalmology and Visual Science	LMM: Nested
Zorrilla-Vaca et al. (2022)	Canadian Journal Of Anesthesia	LMM: Ignore pairing

Table 4: Overview of the studies included in the systematic review