Journal of Clinical Epidemiology 107 (2019) 66–70

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Hybrid models were found to be very elegant to disentangle longitudinal within- and between-subject relationships

Jos W.R. Twisk[a,*], Wieke de Vente[b]

[a]Department of Epidemiology and Biostatistics, UMC Amsterdam, Amsterdam, The Netherlands
[b]Department of Educational Sciences, University of Amsterdam, Amsterdam, The Netherlands

Accepted 23 November 2018; Published online 28 November 2018

Abstract

Objectives: The interpretation of a regression coefficient obtained from a longitudinal data analysis is a combination of a within-subject part and a between-subject part. The hybrid model is used to disentangle the two components. The purpose of this article was to illustrate and discuss the use of the hybrid model in epidemiologic studies.

Study Design and Setting: In the hybrid model the between-subject part of the relationship is obtained using the individual mean value over time, whereas the within-subject part is obtained using the deviation score, that is, the differences between the observations and the individual mean value.

Results: It was shown that the regression coefficient of a standard mixed model analysis is a sort of weighted average of the between- and within-subject part of the relationship. When the outcome was continuous the separate analyses to estimate the two components of a longitudinal relationship were equal to the estimation in the hybrid model. However, for dichotomous outcome, the estimations were slightly different.

Conclusion: The hybrid model is an elegant, easy to perform method to disentangle the within- and between-subject part of a relationship in longitudinal studies.

Keywords: Longitudinal data analysis; Hybrid models; Between-subject effect; Within-subject effect; Mixed models; GEE analysis

1. Introduction

Within the field of epidemiology, there is an increasing interest in observational longitudinal studies. Regarding the analysis of longitudinal data, mixed model analysis and generalized estimating equations (GEE analysis) are the two most used methods [1,2]. Both techniques are extensions of regression analyses, and the general idea behind both methods is that an adjustment is made for the dependency of the observations within the subject. Mixed model analysis performs this adjustment by modeling the differences between the subjects either in the intercept (i.e., by adding a random intercept to the model) or in the regression coefficients for time-dependent independent variables (i.e., by adding random slopes to the model). On the other hand, GEE performs this adjustment by directly modeling the correlations between the repeated measurements within the subjects. Although linear mixed model analysis and GEE analysis show highly similar results, linear mixed model analysis is most used in longitudinal epidemiologic studies. This is probably because of the fact that mixed model analysis is slightly better when there is missing data and is slightly more flexible in the modeling of the dependency of the observations within the subject.

One of the problems with mixed model analysis is the confusing terminology. Mixed model analysis is also known as hierarchical linear modeling, multilevel analysis, random effects modeling, or random coefficient analysis; many different names for the same method. Furthermore, when mixed model analysis is used in epidemiologic studies, it is said that the regression model is divided into two parts. The fixed part contains the regression coefficients, whereas the random part contains the random intercept and/or random slope variance [3,4]. Within econometrics and sociology, for instance, regarding longitudinal studies, a distinction is made between fixed effects

Declarations of interest: none.

* Corresponding author. Department of Epidemiology & Biostatistics, UMC Amsterdam, Location VUMC, De Boelelaan 1089a, 1081 HV Amsterdam, The Netherlands. Tel.: + 31 20 44 44909; Fax: +31 20 44 44495.
E-mail address: jwr.twisk@vumc.nl (J.W.R. Twisk).

https://doi.org/10.1016/j.jclinepi.2018.11.021
0895-4356/

What is new?

Key findings

- In the hybrid model, the between subject part of the relationship is obtained by using the individual mean value over time as the independent variable.
- In the hybrid model, the within subject part of the relationship is obtained by using the deviation score as independent variable.
- The latter reflects the differences between the individual observations and the individual mean value.
- When the outcome variable is dichotomous, the results of a hybrid (logistic) mixed model analysis should be interpreted with caution.

What this adds to what is known?

- Because hybrid models are not much used in epidemiological practice, all key findings adds to what is known.

What is the implication and what should change now?

- The hybrid model is an elegant, easy to perform method that can be used to disentangle the within and between subject part of a relationship in longitudinal studies.

models, between-effects models, and random effects models. A fixed effects model is not only a model with the regression coefficients but also a model in which only the within-subject part of the relationship is estimated. In a between-effects model only the between-subject part of the relationship is estimated, whereas a random effects model is basically the same as a regular mixed model analysis in epidemiology [5–7].

In longitudinal studies the interpretation of the regression coefficient deserves specific attention, in particular, when analyzing the association between two variables that vary over time. When the independent variable is time-dependent the interpretation of the regression coefficient is twofold: a between-subject component and a within-subject component. Although the combined interpretation reflects the total longitudinal relationship between two (time-dependent) variables, in some situations, the researcher may want to disentangle the within- and between-subject interpretation. There are several models available to disentangle the two effects [2]. From these, the hybrid model seems to be the best option [3,8–12]. However, this method is not much used within epidemiologic practice.

Therefore, the purpose of the present article was to illustrate and discuss the use of the hybrid model as a possible tool to disentangle the within- and between-subject part of the relationship in longitudinal epidemiologic studies.

2. Methods

2.1. The hybrid model

When a longitudinal data analysis is performed the between-subject relationship is basically nothing more than the relationship between the mean value of the particular independent variable for each subject and the outcome (Equation 1). To obtain the within-subject part of the relationship, the independent variable must be centered around the mean value of the particular subject; this centering around the mean value is known as the deviation score (Equation 2). To obtain both the within- and the between-subject part of the relationship a combination of Equations 1 and 2 can be applied (Equation 3). The latter is known as the hybrid model.

$$Y_{it} = b_0 + \sum_{j=1}^{J} b_{Bj}\overline{X}_i + \ldots\ldots \quad (1)$$

$$Y_{it} = b_0 + \sum_{j=1}^{J} b_{Wj}\left(X_{ijt} - \overline{X}_i\right) + \ldots\ldots \quad (2)$$

$$Y_{it} = b_0 + \sum_{j=1}^{J} b_{Bj}\overline{X}_i + \sum_{j=1}^{J} b_{Wj}\left(X_{ijt} - \overline{X}_i\right) + \ldots\ldots \quad (3)$$

where Y_{it} = outcome variable for individual i at time-point t; b_0 = intercept; b_W = regression coefficient reflecting the within-subject part of the relationship; b_B = regression coefficient reflecting the between-subject part of the relationship, X_{ijt} = independent variable j for individual i at time point t, and $\overline{X}_i$ = average value of the independent variable X for individual i.

2.2. Example datasets

The use of a hybrid model will be illustrated with examples taken from the Amsterdam Growth and Health Longitudinal Study (AGHLS) [13]; an observational longitudinal study that started in 1976 with a group of adolescents from Amsterdam. Up to now, in this study, there were 10 repeated measurements performed at the ages of 13, 14, 15, 16, 21, 27, 29, 32, 36, and 42 years.

2.3. Analysis

In the first example the longitudinal relationship between cholesterol and the sum of four skinfolds (SSF) was analyzed. SSF is used as an indicator for body fatness and contains the sum of the thickness of the triceps, biceps, subscapular, and suprailiac skinfolds. SSF is a time-dependent variable; so, the overall regression coefficient reflects both the within- and between-subject part of the

relationship. In this first example data of the first six repeated measurements of the AGHLS was used, which contained 147 subjects, and in this particular dataset, there are no missing observations.

In the second example the longitudinal relationship between lung function and smoking behavior was analyzed. Lung function was operationalized with the forced expiratory volume in 1 second (FEV1), and smoking behavior was time-dependent and dichotomized into smoking and nonsmoking at each time point. In this example the last three repeated measurements of the AGHLS at 32, 36, and 42 years of age were used. The analyses were performed on an incomplete dataset and contained 290 subjects.

The last example is comparable to the first example; however, in this example, the outcome variable cholesterol was dichotomized. At each time point, the upper tertile was coded 1 (high cholesterol), and the lower two tertiles were coded 0 (low cholesterol). Again, the longitudinal relationship with SSF was analyzed.

All three examples were analyzed with mixed model analyses—Linear mixed model analyses when the outcome was continuous (example 1 and example 2) and logistic mixed model analyses when the outcome was dichotomous (example 3). In all analyses, first, a model with the individual mean score as the independent variable was analyzed (Equation 1), second, a model with the deviation score around the individual mean as independent variable was analyzed (Equation 2), and third, a hybrid model, including both the individual mean score and the deviation score as independent variables, was analyzed (Equation 3). Finally, also a regular mixed model analysis with the time-dependent independent variable was performed. In all mixed model analyses, only a random intercept was added to the model to take into account the dependency of the observations within the subject, and all analyses were performed with STATA version 14.

3. Results

Table 1 shows descriptive information regarding the datasets used in the examples of the present article.

Table 2. Results of different longitudinal data analyses[a] to disentangle the within- and between-subject relationship between cholesterol and SSF (example 1)

Model	Regression coefficient	Standard error
Between-subjects		
SSF (individual mean)	0.204	0.038
Within-subject		
SSF (deviation score)	0.181	0.021
Hybrid model		
SSF (individual mean)	0.204	0.038
SSF (deviation score)	0.181	0.021
Mixed model		
SSF	0.186	0.018

Abbreviation: SSF, sum of skinfolds.
[a] All models with only a random intercept.

Table 2 shows the results of the longitudinal analyses regarding the relationship between cholesterol and SSF. As can be seen from the results of the first three analyses, the regression coefficients obtained from the hybrid model, including the individual mean and the deviation score, are equal to the regression coefficients obtained from the two separate analyses. This can be explained by the fact that the individual mean is uncorrelated to the deviation score. Furthermore, it can be seen that the overall regression coefficient obtained from the mixed model analysis with SSF as independent variable is some sort of weighted average of the within- and between-subject part of the relationship between cholesterol and SSF.

Table 3 shows the results of the different longitudinal analyses regarding the relationship between FEV1 and smoking behavior. Most interesting part of these results is that the within- and between-subject part of the relationship between FEV1 and smoking behavior have a different sign. A positive nonsignificant between-subject relationship and a negative (highly) significant within-subject relationship. This indicates that the lung function is not much different between smokers and nonsmokers. However, when subjects started to smoke, it has an adverse effect on lung function. Furthermore, it can be seen that the inverse relationship between smoking and

Table 1. Descriptive information (mean and standard deviation) of example datasets

Examples 1 and 3 (n = 147)						
Age (y)	13	14	15	16	21	27
Cholesterol (mmol/L)	4.43 (0.67)	4.32 (0.67)	4.27 (0.71)	4.17 (0.70)	4.67 (0.78)	5.12 (0.92)
SSF (10 cm)	3.26 (1.25)	3.36 (1.35)	3.57 (1.46)	3.76 (1.50)	4.35 (1.68)	4.16 (1.61)

Example 2			
Age (y)	32 (n = 290)	36 (n = 276)	42 (n = 266)
FEV1 (L)	4.10 (0.81)	3.95 (0.79)	3.90 (0.82)
Smoking (yes/no)	49/241	61/215	50/216

Abbreviations: SSF, sum of skinfolds; FEV1, forced expiratory volume in 1 s.

Table 3. Results of different longitudinal data analyses[a] to disentangle the within- and between-subject relationship between FEV1 and smoking behavior (example 2)

Model	Regression coefficient	Standard error
Between-subject		
Smoking (individual mean)	0.202	0.125
Within-subject		
Smoking (deviation score)	−0.221	0.048
Hybrid model		
Smoking (individual mean)	0.202	0.125
Smoking (deviation score)	−0.221	0.048
Mixed model		
Smoking	−0.168	0.045

Abbreviation: FEV1, forced expiratory volume in 1 s.
[a] All models with only a random intercept.

lung function is mainly driven by the within-subject relationship.

Table 4 shows the results of the different logistic mixed model analyses to analyze the relationship between cholesterol (high vs. low) and SSF. Most surprisingly in these results is that the between- and within-subject relationships obtained from the separate logistic mixed model analyses are (slightly) different from the ones obtained from the combined hybrid logistic mixed model analysis.

4. Discussion

In the present article the hybrid model (including both the individual mean and the deviation score as independent variables) was illustrated and discussed as a possibility to disentangle the within- and between-subject part of a longitudinal relationship. It was shown that the overall regression coefficient obtained from a regular mixed model analysis is some sort of weighted average of the two separate relationships obtained from a hybrid model. The hybrid model thus offers a possibility to more precisely answer certain research questions regarding between- and within-subject relationships. However, this does not mean that a hybrid model is by definition better than a regular mixed model analysis. In most situations, one is interested in the overall longitudinal relationship, which can be obtained from a regular mixed model analysis.

Table 4. Results of different logistic mixed models analyses[a] to disentangle the within- and between-subject relationship between cholesterol (high vs. low) and SSF (example 3)

Model	Regression coefficient	Standard error
Between-subject		
SSF (individual mean)	0.947	0.198
Within-subject		
SSF (deviation score)	0.418	0.122
Hybrid model		
SSF (individual mean)	0.967	0.203
SSF (deviation score)	0.413	0.121
Mixed model		
SSF	0.561	0.106

Abbreviation: SSF, sum of skinfolds.
[a] All models with only a random intercept.

It is argued that the use of hybrid models to disentangle the within- and between-subject part of the relationship in longitudinal studies only holds when the time-dependent independent variable is not increasing or decreasing over time. When the time-dependent independent variable is changing over time, it is argued that the deviation score must not be calculated around the individual mean value but that it should be calculated around the individual regression line with time. Furthermore, when the data are unbalanced, that is, when the period and the number of repeated measurements is different for different subjects, also the between-subject part of the relationship should be calculated in a different way, that is, the between-subject part of the relationship should be captured with the intercept of an individual regression line with time when time is centered around the grand mean [3]. Although the calculation of the individual regression line and the deviation from that line makes sense, comparable results may be obtained from a hybrid model adjusted for time, which is much easier to perform. It is, however, not clear under which circumstances the different methods can be applied. In this respect, further (simulation) studies may be necessary.

Although the hybrid model is seen as an appropriate way to disentangle the within- and between-subject part of the relationship in longitudinal studies, there is some debate about its use when the outcome variable is dichotomous. In example 3, it was shown that when logistic mixed model analysis is used, the hybrid model revealed a different within- and between-subject relationship than the two separate analyses to determine the within- and between-subject relationship (see Table 4). This may seem rather strange because the individual mean and the deviation score are uncorrelated. However, this phenomenon is well known in logistic regression analysis and known as noncollapsibility [14–16]. Noncollapsibility arises from differences in the total variances between a univariable logistic model and a multivariable logistic model. Basically, the total variance is the summation of explained and unexplained variance. When an independent variable is added to a linear regression model, the unexplained variance decreases, whereas the explained variance increases with the same amount. However, in a logistic model, the unexplained variance is a fixed number. So, when an independent variable that is related to the outcome is added to a logistic model, the total variance will increase. The consequence of this noncollapsibility is that the regression coefficients of the two variables in the hybrid logistic mixed model analysis are different from the regression coefficients estimated in two separate logistic mixed model analyses. It should further be noted that the within-subject part of the relationship

can also be obtained from conditional logistic regression analysis [17].

In the present article the use of the hybrid model to disentangle the within- and between-subject relationship in longitudinal studies was illustrated using mixed model analyses. It should be noted that exactly the same holds when using GEE analyses to analyze the longitudinal relationships. Although it is known that when the outcome variable is dichotomous, the effect estimates obtained from a logistic GEE analysis will differ from the ones obtained from a logistic mixed model analysis [18]. In the present article all mixed model analyses were limited to models with only a random intercept. The use of hybrid models is, however, not limited to models with only a random intercept. As for the regular mixed model analysis, the hybrid model can easily be extended with random slopes for the time-dependent components of the model. The examples used in the present article were relatively simple without adding potential confounders, mediators, and/or effect modifiers to the models. It is obvious that the analysis of confounding, mediation, and/or effect modification in hybrid models is exactly the same as for regular mixed model or GEE analysis. Furthermore, the use of hybrid models was illustrated with examples with a continuous and dichotomous outcome. Hybrid model can, of course, be used for other outcomes as well.

It has been mentioned that the overall regression coefficient obtained from a mixed model analysis is a sort of weighted average of the between-subject and within-subject relationship. From the examples is could be seen that the weight is not always the same. It is worthwhile noting that the weight depends largely on the magnitude of the within- and between-individual variance of the independent variable [8]. It should further be realized that a hybrid model only makes sense when the independent variable is time-dependent. When the independent variable is time-independent, which is, for instance, the case for the variable treatment in a randomized controlled trial, the regression coefficient of this variable has only a between-subject interpretation. Regarding intervention studies, only in a cross-over trial or a stepped wedge trial, the effect estimate has both a within- and between-subject interpretation, and only in those situations, a hybrid model can be applied.

5. Conclusion

The hybrid model is an elegant, easy to perform method to disentangle the within- and between-subject part of a relationship in longitudinal studies.

References

[1] Fitzmaurice G, Laird N, Ware J. Applied longitudinal analysis. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2011.
[2] Twisk JWR. Applied longitudinal data analysis for epidemiology. A practical guide. 2nd ed. Cambridge: Cambridge University Press; 2013.
[3] Curran PJ, Bauer DJ. The disaggregation of within-person and between-person effects in longitudinal models of change. Annu Rev Psychol 2001;62:583−619.
[4] Twisk JWR. Applied mixed model analysis. A practical guide. Cambridge: Cambridge University Press; 2018: (in press).
[5] Hsiao C. Analysis of panel data. 2nd ed. Cambridge: Cambridge University Press; 2003.
[6] Halaby CN. Panel models in sociological research: theory into practice. Annu Rev Sociol 2004;30:507−44.
[7] Wooldridge JM. Econometric analysis of cross section and panel data. 2nd ed. Cambridge MA: MIT Press; 2010.
[8] Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. Biometrics 1998;54:638−45.
[9] Allison PD. Fixed effects regression methods for longitudinal data using SAS. Cary: SAS Institute Inc; 2005.
[10] Firebaugh G, Warner C, Massoglia M. Fixed effects, random effects, and hybrid models for causal analysis. In: Morgan SL, editor. Handbook of causal analysis for social research. Berlin: Springer; 2013.
[11] Schunk R. Within and between estimated in random-effects models: advantages and drawbacks of correlated random effects and hybrid models. Stata J 2013;13:65−76.
[12] Gunasekara FI, Richardson K, Carter K, Blakely T. Fixed effects analysis of repeated measures data. Int J Epidemiol 2014;43:264−9.
[13] Wijnstok NJ, Hoekstra T, van Mechelen W, Kemper HC, Twisk JW. Cohort profile: the Amsterdam growth and health longitudinal study. Int J Epidemiol 2013;42:422−9.
[14] Newman SC. Commonalities in the classical, collapsibility and counterfactual concepts in confounding. J Clin Epidemiol 2004;57:325−9.
[15] Greenland S, Robins JM. Identifiability, exchangeability and confounding revisited. Epidemiol Perspect Innov 2009;6:4.
[16] Hernan MA, Clayton D, Keiding N. The Simpson's paradox unravelled. Int J Epidemiol 2011;40:780−5.
[17] Neuhaus JM, CE McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. J R Stat Soc Ser B 2006;68:859−72.
[18] Twisk J, de Vente W, Apeldoorn A, de Boer M. Should we use logistic mixed model analysis for the effect estimation in a longitudinal RCT with a dichotomous outcome variable? Epidemiol Biostat Public Health 2017;14(3):e12613-1−12613-8.