European Journal of Epidemiology **19**: 769–776, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis

Jos W.R. Twisk

Department of Clinical Epidemiology and Biostatistics, EMGO-institute, Vrije Universiteit medical centre (VUmc), The Netherlands

Accepted in revised form 7 April 2004

Abstract. The analysis of data from longitudinal studies requires special techniques, which take into account the fact that the repeated measurements within one individual are correlated. In this paper, the two most commonly used techniques to analyze longitudinal data are compared: generalized estimating equations (GEE) and random coefficient analysis. Both techniques were used to analyze a longitudinal dataset with six measurements on 147 subjects. The purpose of the example was to analyze the relationship between serum cholesterol and four predictor variables, i.e., physical fitness at baseline, body fatness (measured by sum of the thickness of four skinfolds), smoking and gender. The results showed that for a continuous outcome variable, GEE and random coefficient analysis gave comparable results, i.e., GEE-analysis with an exchangeable correlation structure and random coefficient analysis with only a random intercept were the same. There was also no difference between both techniques in the analysis of a dataset with missing data, even when the missing data was highly selective on earlier observed data. For a dichotomous outcome variable, the magnitude of the regression coefficients and standard errors was higher when calculated with random coefficient analysis then when calculated with GEE-analysis. Analysis of a dataset with missing data with a dichotomous outcome variable showed unpredictable results for both GEE and random coefficient analysis. It can be concluded that for a continuous outcome variable, GEE and random coefficient analysis are comparable. Longitudinal data-analysis with dichotomous outcome variables should, however, be interpreted with caution, especially when there are missing data.

Key words: Continuous outcome variables, Dichotomous outcome variables, Generalized estimating equations, Longitudinal studies, Missing data, Random coefficient analysis

Abbreviation: GEE = Generalized estimating equations

Introduction

In the last 10 years, there has been a growing interest in longitudinal studies and in the statistical analysis of longitudinal data. Longitudinal studies are defined as studies in which the outcome variable is repeatedly measured; i.e., the outcome variable is measured on the same individual at several occasions. Therefore, observations are not independent of each other. Statistical techniques, which assume independent observations, such as linear regression analysis and logistic regression analysis, can not directly be used in longitudinal studies. For data-analyses in longitudinal studies special statistical techniques are developed, which take into account that the repeated observations of each individual, are correlated [1–3].

With traditional statistical techniques to analyze longitudinal data, such as the paired *t*-test and (M)ANOVA for repeated measurements it is possible to investigate changes in one continuous outcome variable over time and to compare the development of a continuous outcome variable over time between different groups [4, 5]. In longitudinal research, however, there are many other questions to be answered, which require sophisticated statistical techniques. In epidemiological studies, there are two such methods frequently used, i.e., generalized estimating equations (GEE) [6, 7] and random coefficient analysis [8, 9]. Both techniques are suitable for the analysis of the longitudinal relationship between a continuous outcome variable and several time-dependent and time-independent covariates. Furthermore, these techniques are suitable for the longitudinal analysis of a dichotomous outcome variable in relation to the development of other variables [9–11]. The purpose of this paper is to compare the two sophisticated statistical techniques in the analysis of a longitudinal dataset with six measurements. In the comparison an analysis with a continuous outcome variable as well as with a dichotomous outcome variable will be considered.

Materials and methods

Dataset

The dataset used in the examples is taken from the Amsterdam Growth and Health Study, an observational longitudinal study investigating the longitudinal relationship between lifestyle and health in adolescence and young adulthood [12]. In this longitudinal dataset, there are 6 measurements on 147 measurements. The research question to be answered in this example is rather simple: what is the relationship between serum cholesterol levels and four predictor variables, i.e., physical fitness at baseline, which was measured as the maximal oxygen uptake reached on a treadmill-test until exhaustion and expressed in dl min^{-1} $kg^{-2/3}$, body fatness, which was estimated by the sum of the thickness of four skinfolds, i.e., biceps, triceps, subscapular, and suprailiac, and which was expressed in cm, smoking behavior, which was measured by a questionnaire and dichotomized as smoking versus non-smoking, and gender. The predictor variables are chosen in such a way that they are either continuous or dichotomous, and either time-independent or time-dependent.

To illustrate the differences between longitudinal analysis with continuous and dichotomous outcome variables, total serum cholesterol was either expressed in mmol/liter or was dichotomised in such a way that at each measurement the upper tertile is compared to the two lowest tertiles.

Because it is often believed that the major difference between GEE and random coefficient analysis is the way missing data are treated [13, 14], in addition to the analysis of a complete dataset, two incomplete datasets were analyzed with both methods. Both incomplete datasets were derived from the full dataset. In the incomplete datasets, all subjects completed the first three measurements, but from the fourth measurements onwards, 25% of the observations were missing. In the first dataset with missing values, missing data was considered to be completely at random while in the second dataset missing data was considered to be dependent on the value of the outcome variable at the third measurement. The subjects with the highest values of serum cholesterol at $t = 3$ were assumed to be missing at the other three follow-up measurements.

The general idea behind all longitudinal data analyses is that a correction is made for the within-subject correlations, i.e., the correlated 'errors'. GEE and random coefficient analysis use different ways to take into account the dependency. In both techniques, the longitudinal relationship between a certain outcome variable and several predictor variables is estimated using all available data, including data from subjects with missing observations.

Generalized estimating equations

Within GEE, the correction for the dependency of observations is done by assuming (a priori) a certain 'working' correlation structure for the repeated measurements of the outcome variable [6, 7]. Depending on the software package used to estimate the regression coefficients, different correlation structures are available. They basically vary form an 'exchangeable' (or 'compound symmetry') correlation structure, i.e., the correlations between subsequent measurements are assumed to be the same, irrespective of the length of the interperiod, to an 'unstructured correlation structure'. In this structure no particular structure is assumed, which means that all possible correlations between repeated measurements has to be estimated.

Random coefficient analysis

Random coefficient analysis [8] is also known as multilevel analysis [9]. The basic idea behind the use of random coefficient analysis in longitudinal studies is that the regression coefficients are allowed to differ between subjects, i.e., heterogeneity across individuals is allowed. The simplest form of a random coefficient model in longitudinal studies is a model with just a random intercept, i.e., the baseline value is different for each subject. It is also possible that the intercept is not random, but that the regression coefficient with time is considered to be random (i.e., random slope). In other words, the development of a certain variable over time is allowed to vary among individuals. The most interesting possibility is the combination of a random intercept and a random slope with time. Of course it is also possible to consider the regression coefficients of the other time-dependent predictor variables to be random. However, for simplicity reasons, that situation will not be considered in this example.

Analysis

With linear GEE-analysis and random coefficient analysis, serum cholesterol as a continuous outcome variable was analyzed and with logistic GEE-analysis and random coefficient analysis, serum cholesterol as a dichotomous outcome variable was analyzed. Besides the four predictor variables, in all analyses also the linear relationship with time (added to the models as a continuous variable) was analyzed (see for formulas the appendix).

Within GEE-analyses, different correlation structures were compared with each other and within random coefficient analysis, a comparison was made between analysis with no random coefficients, with only a random intercept, and with both a random intercept and a random slope with time.

Finally, the same analyses were performed on the datasets with missing data. Both GEE- and random coefficient analyses were carried out with STATA [15].

Results

Figure 1 shows descriptive information of the time-dependent variables used in the example. The time-independent predictor variable fitness was 1.98 dl min^{-1} $kg^{-2/3}$ (SD 0.22), and there were 69 males and 79 females in the sample.

Table 1 gives an overview of the results derived from the GEE-analyses with different correlation structures, and the results of the different random coefficient analyses. Regarding the GEE-analysis, the magnitude of the regression coefficients is quite different for the two correlation structures. The standard errors, however, are almost the same. For the two models with random coefficients the magnitude of the regression coefficients was almost the same. Only for gender, the inclusion of a random slope with time influenced the regression coefficient. All regression coefficients were highly different from an analysis without any random coefficient. Regarding the standard errors for the time-independent predictor variables (fitness at baseline and gender), they were lower than the ones obtained from the analysis with random coefficients, while the standard errors of the time-dependent predictor variables (sum of skinfolds and smoking) and time were higher when calculated without random coefficients.

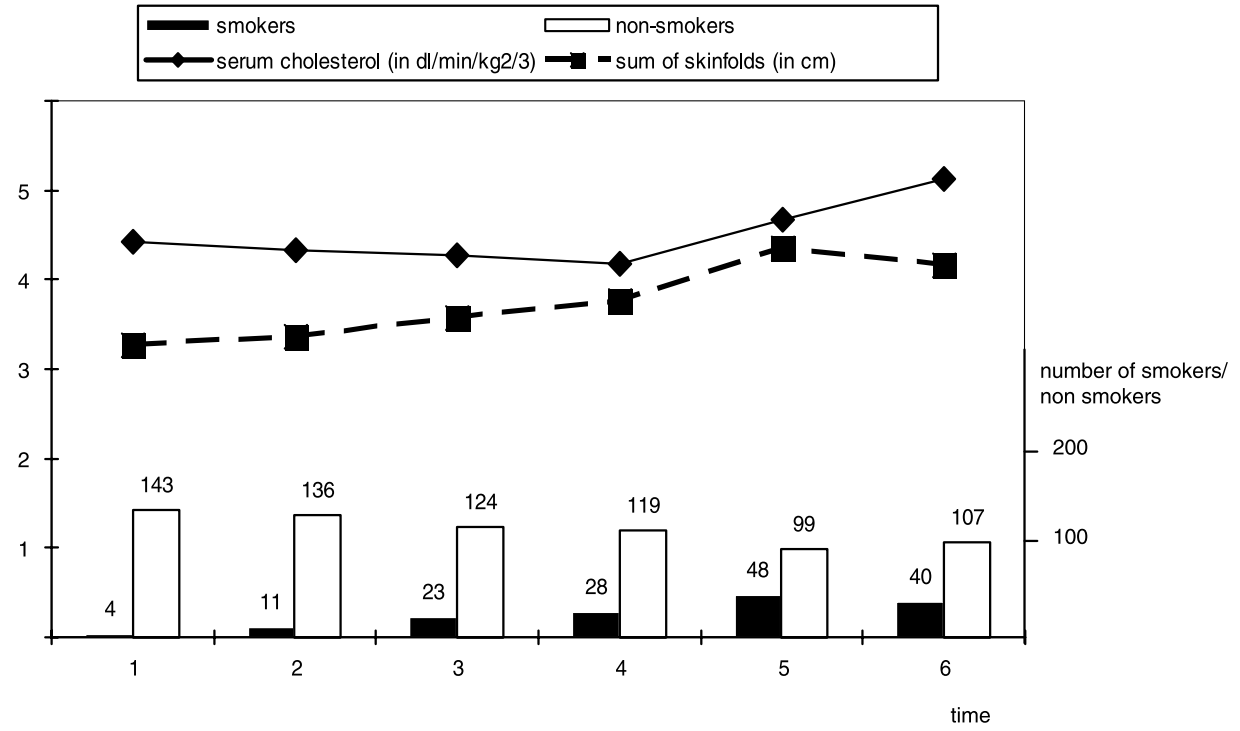

Figure 1. Development over time for the time-dependent variables used in the present example, i.e., serum cholesterol, sum of skinfolds, and smoking behavior.

Table 1. Regression coefficients and standard errors (values in parenthesis) estimated by GEE-analysis and random coefficient analysis for serum cholesterol as a continuous outcome variable

	GEE-analysis		Random coefficient analysis		
	Exchangeable correlation structure	Unstructured correlation structure	No random coefficients	Random intercept	Random intercept and slope[a]
Fitness at baseline	−0.02 (0.28)	−0.09 (0.26)	0.11 (0.15)	−0.02 (0.27)	−0.02 (0.27)
Sum of skinfolds	0.11 (0.02)	0.08 (0.02)	0.17 (0.02)	0.11 (0.02)	0.11 (0.02)
Smoking	−0.11 (0.06)	−0.11 (0.06)	−0.02 (0.07)	−0.11 (0.06)	−0.12 (0.06)
Gender	0.10 (0.13)	0.08 (0.13)	0.05 (0.06)	0.10 (0.12)	0.04 (0.12)
Time	0.11 (0.01)	0.15 (0.02)	0.09 (0.02)	0.11 (0.01)	0.11 (0.01)

[a]Only a random slope with time is considered.

For the analysis with a dichotomous outcome variable, the situation is a bit different (Table 2). For GEE, both the regression coefficient and standard errors were almost the same for an exchangeable and unstructured correlation structure. For the magnitude of the standard errors, the same pattern was found as has been found for the analysis of a continuous outcome variable. For the random coefficient analysis, there was a remarkable difference in the magnitude of both the regression coefficients and the standard errors of all analyses. A comparison between the 'naive' analysis (i.e., the analysis without random coefficients) and the analysis with random coefficients shows that the standard errors from the 'naive' analysis were much lower for all four predictor variables as well as for time.

Table 2. Regression coefficients and standard errors (values in parenthesis) estimated by GEE-analysis and random coefficient analysis for serum cholesterol as a dichotomous outcome variable (i.e., the upper tertile compared to the two lowest tertiles)

	GEE-analysis		Random coefficient analysis		
	Exchangeable correlation structure	Unstructured correlation structure	No random coefficients	Random intercept	Random intercept and slope[a]
Fitness at baseline	0.22 (0.76)	0.22 (0.76)	0.56 (0.43)	0.88 (1.81)	0.33 (1.63)
Sum of skinfolds	0.34 (0.06)	0.33 (0.06)	0.47 (0.06)	0.70 (0.14)	0.72 (0.14)
Smoking	−0.15 (0.20)	−0.12 (0.19)	−0.13 (0.21)	−0.25 (0.36)	−0.23 (0.36)
Gender	0.08 (0.38)	0.09 (0.37)	0.05 (0.19)	0.34 (0.71)	0.46 (0.70)
Time	−0.08 (0.04)	−0.08 (0.04)	−0.11 (0.05)	−0.16 (0.07)	−0.07 (0.10)

[a]Only a random slope with time is considered.

Table 3. Regression coefficients and standard errors (values in parenthesis) derived from GEE-analysis and random coefficient analysis[a], investigating the longitudinal relationships between serum cholesterol (both continuous and dichotomous) and several predictor variables. Full dataset, and two incomplete datasets

	Fitness at baseline	Sum of skinfolds	Smoking	Gender	Time
Serum cholesterol (continuous)					
Full dataset					
GEE-analysis	−0.02 (0.28)	0.11 (0.02)	−0.11 (0.06)	0.10 (0.13)	0.11 (0.01)
Random coefficient analysis	−0.02 (0.27)	0.11 (0.02)	−0.12 (0.06)	0.04 (0.12)	0.11 (0.01)
Missing completely at random					
GEE-analysis	0.05 (0.26)	0.14 (0.02)	−0.10 (0.07)	0.05 (0.13)	0.10 (0.02)
Random coefficient analysis	0.06 (0.27)	0.14 (0.02)	−0.11 (0.06)	0.02 (0.12)	0.10 (0.01)
Missing dependent on earlier observations[b]					
GEE-analysis	0.05 (0.27)	0.15 (0.03)	−0.07 (0.07)	0.05 (0.13)	0.07 (0.01)
Random coefficient analysis	0.05 (0.27)	0.15 (0.02)	−0.08 (0.06)	0.05 (0.12)	0.07 (0.01)
Serum cholesterol (dichotomous)					
Full dataset					
GEE-analysis	0.22 (0.76)	0.34 (0.06)	−0.15 (0.20)	0.08 (0.38)	−0.08 (0.04)
Random coefficient analysis	0.33 (1.63)	0.72 (0.14)	−0.23 (0.36)	0.46 (0.70)	−0.07 (0.10)
Missing completely at random					
GEE-analysis	0.48 (0.79)	0.40 (0.07)	−0.02 (0.34)	−0.04 (0.19)	−0.08 (0.04)
Random coefficient analysis	0.78 (1.80)	0.87 (0.16)	0.07 (0.73)	−0.03 (0.40)	−0.12 (0.10)
Missing dependent on earlier observations[b]					
GEE-analysis	0.37 (0.76)	0.40 (0.07)	0.03 (0.33)	−0.05 (0.22)	−0.12 (0.04)
Random coefficient analysis	0.36 (1.90)	0.86 (0.17)	0.35 (0.80)	−0.02 (0.42)	−0.16 (0.11)

[a]GEE-analysis with an exchangeable correlation structure; random coefficient analysis with a random intercept and a random slope with time.
[b]Subjects (25%) with the highest values for serum cholesterol at $t = 3$ were assumed to be missing for the last three follow-up measurements.

Table 3 shows the results of the analysis on the incomplete datasets. To simplify, only the results of one GEE-analysis (i.e., with an exchangeable correlation structure) and one random coefficient analysis (i.e., with both a random intercept and a random slope with time) are given. For the continuous outcome variable, incompleteness of the dataset only had marginal influence on the results of the longitudinal data analysis. Surprisingly, the results for the GEE-analysis and the random coefficient analysis were almost equal for the two missing data patterns.

For the dichotomous outcome variable, for both GEE- and random coefficient analysis, the differences between the complete and the incomplete datasets were huge. This holds for both the regression coefficients and standard errors.

Discussion

The purpose of this paper was to compare the results of a GEE-analysis and a random coefficient analysis on a longitudinal dataset with six measurements. Furthermore, different possibilities within the two techniques were compared with each other.

In the literature it is assumed that GEE-analysis is robust against a wrong choice for a correlation structure, i.e., it does not matter which correlation structure is chosen, the results of the longitudinal analysis will be more or less the same [16, 17]. However, when the results of analysis with different working correlation structures are compared to each other, especially for the analysis with a continuous outcome variable, the magnitude of the regression coefficients are different. It is therefore important to realize which correlation structure should be chosen for the analysis. Although the unstructured correlation structure is always the best, also the simplicity of the correlation structure has to be taken into account. The number of parameters (in this case correlation coefficients) which needs to be estimated differs for the various working correlation structures. In the example dataset with six repeated measurements, for instance, for an exchangeable structure only one correlation coefficient has to be estimated, while for the unstructured correlation structure, 15 correlation coefficients must be estimated. As a result, the power of the statistical analysis is influenced by the choice for a certain structure. The best option is therefore to choose the simplest structure which fits the data well. The first step in choosing a certain correlation structure can be to investigate the within-person correlation coefficients for the outcome variable (Table 4). Based on these coefficients, an exchangeable correlation structure seems to be the simplest appropriate choice in this particular situation. It should be kept in mind that when analyzing covariates, the correlation structure can change (i.e., the choice of the correlation structure should better be based conditionally on the covariates).

Table 4. Within-person correlation coefficients for serum cholesterol

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	–	0.76	0.70	0.67	0.64	0.59
t_2		–	0.77	0.78	0.67	0.59
t_3			–	0.85	0.71	0.63
t_4				–	0.74	0.65
t_5					–	0.69
t_6						–

For random coefficient analysis, one has to choose which coefficients have to be assumed random. This choice is easier than the choice for a working correlation structure in GEE-analysis. This is due to the fact that most standard software, which can be used for random coefficient analysis, provides −2 log likelihood values of each model, which can be used to evaluate different models. In the presented example, a model with both a random intercept and a random slope with time was found to be the most appropriate.

Interpretation of the regression coefficients

The interpretation of the magnitude of the regression coefficients obtained from either GEE-analysis or random coefficient analysis is not straightforward. Basically the obtained regression coefficient is a 'pooled' coefficient of a within-subject and a between-subjects relationship. This has the following implications for the interpretation of the regression coefficients: suppose that for a particular subject the serum cholesterol concentration is relatively high at each of the repeated measurements and does not change much over time. Suppose further that for that particular subject the sum of skinfolds is also relatively high at each of the repeated measurements. This indicates a longitudinal 'between-subjects' relationship between serum cholesterol and the sum of skinfolds. Suppose that for another subject the serum cholesterol concentration increases rapidly along the longitudinal period, and suppose that for the same subject this pattern is also found for the sum of skinfolds. This indicates a 'within-subject' relationship between serum cholesterol and the sum of skinfolds. Both relationships are part of the overall longitudinal relationship, so both should be taken into account in the analysis of the longitudinal relationship. The regression coefficient estimated with either GEE-analysis or random coefficient analysis 'combines' the two possible relationships into one regression coefficient.

Comparison between GEE-analysis and random coefficient analysis

Both GEE- and random coefficient analyses are highly suitable to analyze longitudinal data, because

in both methods a correction is made for the dependency of the observations within one individual. The question than arises: which of the two methods is better? Unfortunately, no clear answer can be given. For continuous outcome variables, GEE-analysis with an exchangeable correlation structure is the same as a random coefficient analysis with only a random intercept (see Table 1). The correction for the dependency of observations with an exchangeable 'working correlation' structure is the same as allowing individuals to have random intercepts. When the dependency of observations is slightly more complicated, GEE-analysis with a different correlation structure can be used or random coefficient analysis with additional random regression coefficients for other variables (e.g., time). Although random coefficient analysis is slightly more flexible, it should be realized that 'regular' random coefficient analysis is limited by the fact that the random regression coefficients are assumed to be normally distributed.

For dichotomous outcome variables, the situation is more complex. It is important to realize that the regression coefficient calculated with GEE-analysis is the average value of the individual regression lines. Therefore, the regression coefficients estimated with GEE-analysis are called 'population averaged' [3, 6]. The regression coefficients calculated with random coefficient analysis can be seen as the 'average individual' or 'subject specific'. For the linear longitudinal regression analysis, both the GEE- and the random coefficient approach leads to exactly the same results; i.e., the 'population average' coefficient is equal to the 'subject specific' coefficient. For the logistic longitudinal regression analysis, both approaches lead to different results. This has to do with the fact that in logistic regression analysis the intercept has a different interpretation than in linear regression analysis. The regression coefficients calculated with a logistic GEE-analysis will always be lower than the coefficients calculated with a comparable random coefficient analysis [18–20]. This was also seen in the results reported in Table 2.

When a dichotomous outcome variable is analyzed in a longitudinal study, should GEE-analysis or random coefficient analysis be used? If one is performing a population study and one is interested in the relationship between a dichotomous outcome variable and several other predictor variables, GEE-analysis will probably provide the most 'valid' answer. However, if one is interested in the individual development over time of a dichotomous outcome variable, random coefficient analysis will probably provide the most 'valid' results. It should, however, also be noted that random coefficient analyses with a dichotomous outcome variable are not fully developed yet. Different software packages give different results and within one software package there are (mostly) more than one possibility to estimate the coefficients and unfortunately, the different estimation procedures often lead to totally different results [21]. In other words, although in theory random coefficient analysis can be suitable in some situations, in practice one should be very careful in using this technique in the longitudinal analysis of a dichotomous outcome variable.

How important is correcting for the dependency of observations?

It is interesting to evaluate the importance of correcting for the dependency of the observations in longitudinal studies. To do so, the results of the random coefficient analysis can be compared to the random coefficient analysis without a random coefficient (a 'naive' analysis) which assumes that all 882 observations (i.e., 147×6) are independent. Although also the magnitude of the regression coefficients differed between the 'naive' analysis and the analysis corrected for the dependency of observations, the major differences are observed in the magnitude of the standard errors. In general, ignoring the dependency of the observations leads to an underestimation of the standard errors of the time-independent predictor variables and to an overestimation of the standard errors of the time-dependent predictor variables [6]. For the time-independent predictor variables in the 'naive' analysis it is assumed that each measurement within a particular individual provides 100% new information, while part of the information was already available in earlier measurements of that individual. A part, which is reflected in the within-person correlation coefficient. Depending on the magnitude of that coefficient each repeated measurement within one individual provides less than 100% of new information. This leads in the corrected analysis to bigger standard errors. For the time-dependent predictor variables on the other hand, both GEE-analysis and random coefficient analysis use the fact that the same individuals are measured over time. This leads to lower standard errors.

Influence of missing data

For the continuous outcome variable the regression coefficients and standard errors calculated with the different methods on the incomplete datasets were only slightly different. Also the type of the missing data pattern did not influence the differences between GEE-analysis and random coefficient analysis. Furthermore, the differences found in the dataset with random missing data were not bigger than the differences observed for the dataset with selective missing data. So, in this particular situation, both GEE and random coefficient analysis were 'valid' in datasets with missing observations, even when the missing data is (highly) selective. This is rather surprisingly, because in the literature it is always argued that one of the

biggest differences between GEE-analysis and random coefficient analysis is found in the analysis of incomplete datasets [13, 14]. It is argued that GEE-analysis requires missing data to be completely at random, while random coefficient analysis is less restrictive in its requirements of missing data patterns. For the dichotomous outcome variable, the analyses of incomplete datasets led to remarkably different results than the analysis of the complete dataset. In general, the influence of missing data in the analysis of a dichotomous outcome variable was rather unpredictable. A possible solution to avoid this is to use imputation techniques when missing observations occur [22, 23]. However, in another study it was shown that the use of imputation techniques can also lead to unpredictable results in the longitudinal analysis of a dichotomous outcome variable [21].

Conclusions

For continuous outcome variables, GEE and random coefficient analysis are highly comparable and both methods lead to 'valid' results when applied to incomplete datasets. For dichotomous outcome variables, the regression coefficients calculated with GEE are always lower than the coefficients calculated with random coefficient analysis. Applied to incomplete datasets, the results are (highly) different than the results obtained from a complete dataset. They are highly unpredictable, and should therefore be interpreted cautiously.

Appendix

For the longitudinal analysis with a continuous outcome variable, the following statistical model was used:

$$Y_{it} = \beta_0 + \beta_1 t + \beta_{2j} \sum_{j=1}^{J} X_{ijt} + \beta_{3m} \sum_{m=1}^{M} X_{im} + \varepsilon_{it} ,$$

where Y_{it} is the outcome variable for subject i at time t, β_0 the intercept, β_1 the regression coefficient for time, t the time, β_{2j} the regression coefficient for time-dependent predictor variable j, X_{ijt} the time-dependent predictor variable j for subject i at time t, J the number of time-dependent predictor variables, β_{3m} the regression coefficient for time-independent predictor variable m, X_{im} the time-independent predictor variable m for subject i, M the number of time-independent variables, and ε_{it} is the 'error' for subject i at time t.

For GEE-analysis, the model is extended with a correction for a 'working correlation structure':

$$Y_{it} = \beta_0 + \beta_1 t + \beta_{2j} \sum_{j=1}^{J} X_{ijt} + \beta_{3m} \sum_{m=1}^{M} X_{im} + [\text{corr}] + \varepsilon_{it} ,$$

where [corr] is the working correlation matrix, while for random coefficient analysis the model is extended with possible random regression coefficients:

$$Y_{it} = \beta_{0i} + \beta_{1i} t + \beta_{2ij} \sum_{j=1}^{J} X_{ijt} + \beta_{3m} \sum_{m=1}^{M} X_{im} + \varepsilon_{it} ,$$

where β_{0i} is the random intercept, β_{1i} is the random regression coefficient for *time*, and β_{2ij} the random regression coefficient for time-dependent predictor variable j. It should be noted that in the present example only the intercept and the regression coefficient for time are considered to be random.

For the longitudinal analysis with a dichotomous outcome variable, a the following model was used:

$$\text{logit}(Y_{it}) = \beta_0 + \beta_1 t + \beta_{2j} \sum_{j=1}^{J} X_{ijt} + \beta_{3m} \sum_{m=1}^{M} X_{im} .$$

For both GEE-analysis and random coefficient analysis, the same extensions are used as mentioned before.

References

1. Zeger SL, Liang K-Y. An overview of methods for the analysis of longitudinal data. Stat Med 1992; 11: 1825–1839.
2. Hand DJ, Crowder MJ. Practical longitudinal data analysis. London: Chapman and Hall, 1996.
3. Diggle PJ, Liang K-Y, Zeger SL. Analysis of longitudinal data. New York: Oxford University Press Inc., 1994.
4. Morrisson DF. Multivariate statistical methods. New York: McGraw-Hill, 1976.
5. Crowder MJ, Hand DJ. Analysis of repeated measures. London: Chapman and Hall, 1990.
6. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42: 121–130.
7. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986; 73: 45–51.
8. Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics 1982; 38: 963–974.
9. Goldstein H. Multilevel statistical models, 2nd edn. London: Edward Arnold, 1995.
10. Vonesh EF, Carter RL. Mixed effect nonlinear regression for unbalanced repeated measures. Biometrics 1992; 48: 1–17.
11. Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. Biometrika 1991; 78: 153–160.
12. Kemper HCG (ed.), The Amsterdam Growth Study: A longitudinal analysis of health, fitness and lifestyle. HK Sport Science Monograph Series, Vol. 6. Champaign Il: Human Kinetics Publishers Inc., 1995.
13. Albert PS. Longitudinal data analysis (repeated measures) in clinical trials. Stat Med 1999; 18: 1707–1732.
14. Omar RZ, Wright EM, Turner RM, et al. Analyzing repeated measurements data: A practical comparison of methods. Stat Med 1999; 18: 1587–1603.

15. Stata Corporation. Stata statistical software: Release 6. College Station, Texas, USA, 1999.
16. Liang K-Y, Zeger SL. Regression analysis for correlated data. Annu Rev Publ Health 1993; 14: 43–68.
17. Twisk JWR. Different statistical models to analyze epidemiological observational longitudinal data: An example from the Amsterdam Growth and Health Study. Int J Sports Med 1997; 18(Suppl 3): S216–S224.
18. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev 1991; 59: 25–36.
19. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject specific approaches for analyzing repeated measures binary outcomes. Am J Epidemiol 1998; 147: 694–703.
20. Crouchley R, Davies RB. A comparison of GEE and random coefficient models for distinguishing heterogeneity, nonstationarity and state dependence in a collection of short binary event series. Stat Model 2001; 1: 271–285.
21. Twisk JWR. Applied longitudinal data analysis for epidemiology. A practical guide. Cambridge: Cambridge University Press, 2003.
22. Twisk JWR, Vente W de. Attrition in longitudinal studies. How to deal with missing data. J Clin Epidemiol 2002; 55: 329–337.
23. Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley, 1987.

Address for correspondence: Department of Clinical Epidemiology and Biostatistics, EMGO-institute, Vrije Universiteit medical centre (VUmc), Vd Boechorststraat 7, 1081 BT Amsterdam, The Netherlands
Phone: +31-20-4448409; Fax: +31-20-4448181
E-mail: jwr.Twisk@vumc.nl