Ziegler, Kastner, Chang-Claude:

Analysis of Pregnancy and Other Factors on Detection of Human Papilloma Virus (HPV) Infection Using Weighted Estimating Equations for Follow-Up Data

Sonderforschungsbereich 386, Paper 201 (2000)

Online unter: http://epub.ub.uni-muenchen.de/

Analysis of Pregnancy and Other Factors on Detection of Human Papilloma Virus (HPV) Infection Using Weighted Estimating Equations for Follow–Up Data

Andreas Ziegler [1] Christian Kastner [2] Jenny Chang-Claude [3]

May 20, 2000

Abstract

Generalised estimating equations have been well established to draw inference for the marginal mean from follow-up data. Many studies suffer from missing data that may result in biased parameter estimates if the data are not missing completely at random. Robins and coworkers proposed to use weighted estimating equations (WEE) in estimating the mean structure if drop-out occurs missing at random. We illustrate the differences between the WEE and the commonly applied available case analysis in a simulation study. We apply the WEE and re-analyse data on pregnancy and HPV infection. We estimate the response probabilities and demonstrate that the data are not missing completely at random. Upon use of the WEE, we are able to show that pregnant women have an increased odds for an HPV infection compared with study subjects after delivery ($p = 0.027$). We conclude that the WEE are useful in analysing follow-up data with drop-outs.

Keywords: Correlated Data Analysis, Generalised Estimating Equations, Horvitz-Thompson Estimation, Marginal Models, Missing Data, Weighted Estimating Equations

1 INTRODUCTION

Several approaches for the analysis of follow-up data have been well developed in the last decade. For example, the Generalised Estimating Equations (GEE; Liang and Zeger, 1986) can be applied to investigate the marginal effect of an intervention or potential risk factors. The GEE have been implemented in some statistical software packages (Ziegler, Kastner and Blettner, 1998) and are widely used in practice. The available procedures generally require complete data. Many studies suffer, however, from missing or incomplete data. In this situation, either all available cases or the complete cases are used by most computer programs. The GEE approach may yield biased estimates if data are not missing completely at random (Laird, 1988; Liang and Zeger, 1986; Robins, Rotnitzky and Zhao, 1995).

In the literature, two different semiparametric approaches have been proposed to deal with the problem of missing data. Both concepts underlying these methods are intuitive and appealing. Imputation methods impute missing data. They have been discussed e.g. by Paik (1997) and Xie and Paik (1997) for follow-up studies. By contrast, weighting methods discard the incomplete data but weight observations inversely proportional to their observation probability (Paik, 1997). Thus, the weighting estimating equations (WEE) generally follow the classical Horvitz-Thompson approach. They have been extensively discussed in recent years (Robins, 1994; Robins et al., 1995; Robins and Rotnitzky, 1995; Rotnitzky and Robins, 1995; Zhao, Lipsitz and Lew, 1996) but have rarely been applied in practice. This may be due to the fact that they were not available in accessible form with computer software (Carlin, Wolfe, Coffey and Patton, 1999). WEE have been proposed for both forms of item non-response that are commonly distinguished: non-response in explanatory variables, i.e. in the x (Robins, 1994; Zhao et al., 1996), and non-response in dependent variables, i.e. in the y (Robins et al., 1995; Robins and Rotnitzky, 1995;

1 Medical Centre for Methodology and Health Research, Institute of Medical Biometry and Epidemiology, Philipps–University of Marburg, Bunsenstr. 3, 35033 Marburg, Germany, ziegler@mailer.uni-marburg.de

2 Institute of Statistics, LMU München, Ludwigstr. 33, 80539 München, Germany, kchris@stat.uni-muenchen.de

3 German Cancer Research Centre, Im Neuenheimer Feld 320, 62219 Heidelberg, Germany, j.chang-claude@dkfz.heidelberg.de

Rotnitzky and Robins, 1995). According to the structure of our real data on human papilloma virus (HPV) infection, we restrict our attention to missing data in dependent variables. Explanatory variables of interest are assumed to be completely observed. We furthermore consider only WEE for longitudinal data with monotone missing patterns; that is, once a subject leaves the study, it will never return. In the framework of clinical trials, an individual fulfilling this criterion is usually termed a drop-out.

Our interest is to consistently estimate the marginal mean; the association between the responses is treated as nuisance. For the following reasons, we shall use the Independence Estimating Equations (IEE) for parameter estimation. Firstly, the IEE can be considered as a special case of the GEE in which the association is not modeled in addition to the association structure (Ziegler, Kastner and Blettner, 1998). Thus, estimation is less time consuming. Secondly, the IEE are as efficient as the GEE in various situations (Fitzmaurice, 1995; Mancl and Leroux, 1996). For example, if only baseline covariates are included so that all explanatory variables are constant within one observational unit i.e. cluster, IEE are as efficient as GEE, if the true correlation is equal between any two sample time points. Thirdly, in the case of negative correlations, a GEE estimator need not exist (Hanley, Negassa and deB. Edwardes, 2000). The IEE estimator, however, can be used in this situation. Finally, the GEE underly two important implicit assumptions that are required for the validity of this method: The association needs to be correctly specified. Otherwise, the association parameters involved might be subject to uncertainty of definition which can result in a breakdown of the asymptotic properties of the estimators (Crowder, 1995). Furthermore, the mean structure needs to be correctly specified as a function of all—probably time point specific—explanatory variables. Otherwise, the resulting estimator might be biased (Pepe and Anderson, 1994). These two implicit assumptions do not apply to the IEE.

The aim of this paper is a re-analysis of longitudinal data from a study on the effects of pregnancy and other factors on detection of HPV infection (Chang-Claude, Schneider, Smith, Blettner, Wahrendorf and Turek, 1996) using both the WEE approach and the "classical" complete case analysis using the IEE. Both approaches are implemented in the program MAREG (Kastner, Fieger and Heumann, 1997).

In the next section we introduce the data. In Section 3 the IEE are derived assuming complete observations. We discuss missing data mechanisms for dependent variables in Section 4. Section 5 deals with approaches to estimate the response probabilities. These estimates are used in the WEE of Robins et al. (1995) which are introduced in section 6 assuming a monotone missing data pattern. The difference between the IEE and the WEE in the presence of missing response data is illustrated in Section 7 by simulations. The analysis of the HPV data is presented in Section 8.

2 HUMAN PAPILLOMA VIRUS INFECTION DATA

HPV infection of the lower female genital tract has been clinically observed to appear more frequently during pregnancy and show a high regression rate after delivery (Barber, Werdel, Symbula, Williams, Burkett and Taylor, 1992; Garry and Jones, 1985; Schneider, Hotz and Gissman, 1987). Epidemiological studies have established that infection with certain genital "high risk" and "low risk" HPV types is associated with the development of specific cancers (Koutsky, Holmes, Critchlow, Stevens, Paavonen and Beckmann, 1992; zur Hausen, 1991; Gissmann, Wolnik, Ikenberg, Koldovsky, Schnurch and zur Hausen, 1983).

One aim of the study conducted by Chang-Claude et al. (1996) was to clarify previous conflicting data concerning the detection of HPV infection in the pregnant vs. the non–pregnant state while adjusting for other known risk factors of cervical cancer. The longitudinal study used a cohort design with (internal) control group to compare HPV detection over time in pregnant women compared to a group of age-frequency matched non–pregnant controls. The study design, the subjects and serum samples have previously been described in detail (Chang-Claude et al., 1996). Briefly, all subjects were followed to evaluate changes in HPV status. Study subjects were identified at the University Womens' Clinic in Ulm, Germany, between October 1987 and May 1990. Women who had a history of reproductive cancer or other disorders associated with HPV infection were excluded from the study. The investigators aimed to carry out 5 follow-ups at 3–month intervals after enrollment. At the first visit a self-administered questionnaire was completed by all participants including questions on reproductive history, gynecological diseases, contraceptive methods, sexual behavior, urogenital disease of the partner, smoking, marital status, and education. Infection with HPV and Vira-Pap smear were determined at each time point. Abnormal Vira-Pap smear was strongly associated with HPV infection, thus an indicator for HPV infection. At follow-up

visits, a shorter version of the original questionnaire was filled in. It included questions on pregnancy status at interim, contraception, hormone use, gynecological diseases, sexual behavior, urogenital disease of the partner, and smoking behavior in the past three months.

Table I shows the observed monotone response pattern for the participating individuals. 155 of the 267 individuals were followed-up at all time points; for 36 individuals baseline values were available only. 27.7% of the completely followed-up women were pregnant at baseline, while 31.8% of the 267 subjects were pregnant at baseline.

Table I: Observed monotone response patterns for the human papilloma virus infection data. Observation = 1 if individual was observed, and 0 otherwise.

Follow-up					
1	2	3	4	5	Number
1	1	1	1	1	155
1	1	1	1	0	19
1	1	1	0	0	21
1	1	0	0	0	15
1	0	0	0	0	21
0	0	0	0	0	36

Upon use of univariate logistic regression and autoregressive logistic regression, Chang-Claude et al. (1996) were able to identify age (in years) as strong determinant of HPV infection. They also showed an association of the lifetime number of sexual partners (≤ 3, 4–6, ≥ 7) and the number of cigarettes smoked daily (0, $1-19$ and ≥ 20) with a $p > 0.05$. In addition, the odds for HPV infection postpartum tended to be lower than the odds for non-pregnant women (Odds Ratio (OR) 0.68; 95% confidence interval (CI) $0.26-1.80$), although this was not statistically significant. The chance for an HPV infection was not statistically significant higher compared with non-pregnant study subjects (OR 1.22; 95% CI $0.49-3.05$). However, the final model included some other factors like current condom (yes = 1 and 0, otherwise) use that were not significant at the 5% test-level.

In our re-analysis of the HPV data (section 8), we treated HPV infection in analogy to Chang-Claude et al. (1996). Thus, HPV infection was modeled as a binary random variable being 1 if an HPV infection with any of the HPV types 16, 18, 31, 33, and 35 could unequivocally be determined and 0 if the test results were negative. Furthermore, a binary Vira-Pap smear surrogate variable was created with 1 if a clear–cut negative result was obtained and 0 otherwise.

3 INDEPENDENCE ESTIMATING EQUATIONS

Let $\mathbf{y}_i$ be a vector of responses from n clusters with T observations for the ith cluster. For each response y_{it}, several covariates $\mathbf{x}_{it}$ are available which can be summarized to the matrix $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})'$. The method can be easily extended to unequal cluster sizes T_i. The pairs $(\mathbf{y}_i, \mathbf{X}_i)$ are assumed to be independently identically distributed. We focus on marginal models so that we do not consider models including state dependence or duration dependence. The association within a cluster needs, however, to be taken into account.

For independent observations, the well-known Generalized Linear Model (GLM) allows flexibility in modeling mean and variance structures. For binary variables (HPV infection present or absent), the relation between the response y_{it} and $\mathbf{x}_{it}$ may be described e.g. via the logit link so that $E(y_{it}|\mathbf{x}_{it}) = \mu_{it} = \text{logit}^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta})$. Furthermore, the variance function is usually chosen as the binomial variance function $\mu_{it}(1-\mu_{it})$. $\boldsymbol{\beta}$ is the $p \times 1$ parameter vector of interest which is estimated by solving the IEE

$$\mathbf{u}(\beta) = \sum_{i=1}^{n} \mathbf{D}_i' \mathbf{V}_i^{-1} \boldsymbol{\epsilon}_i = \mathbf{D}'\mathbf{V}^{-1}\boldsymbol{\epsilon} = \mathbf{0}. \tag{1}$$

Here, $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}'$ is the diagonal matrix of first derivatives and $\mathbf{V}_i$ is the diagonal matrix of the variances from the GLM. $\boldsymbol{\epsilon}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ is the ordinary residual with $\boldsymbol{\mu}_i$ being the vector of the mean structure. $\mathbf{D}$ and $\boldsymbol{\epsilon}$ are the stacked $\mathbf{D}_i$ matrices and $\boldsymbol{\epsilon}_i$ vectors, respectively. Finally, $\mathbf{V}$ is the block diagonal matrix of $\mathbf{V}_i$.

An analytic solution of (1) exists for the linear model. In general, (1) have to be solved iteratively by a FISHER-scoring algorithm or iterative weighted least squares (IWLS). If the observations are indeed independent and suitable regularity conditions hold, the estimator $\hat{\boldsymbol{\beta}}_{IEE}$ is consistent and asymptotically normal with variance matrix

$$Cov(\hat{\boldsymbol{\beta}}_{IEE}) = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}. \tag{2}$$

For correlated observations, however, the true variance matrix $Cov(\mathbf{y}_i|\mathbf{X}_i) = \boldsymbol{\Omega}_i \neq \mathbf{V}_i$ is not diagonal. Thus, the use of (2) generally leads to biased variance estimates. Therefore, Zeger, Liang and Self (1985) proposed to use the robust estimator because it is a strongly consistent estimator of $V(\hat{\boldsymbol{\beta}}_{IEE})$ even if the dependent variables within a cluster are correlated. The robust variance estimator is also known as the sandwich information matrix or the HUBER-estimator (s. e.g. Ziegler, Kastner and Blettner, 1998).

4 MISSING DATA MECHANISMS

As stated in the Introduction, we assume that the explanatory variables of interest $\mathbf{X}_i$ are observed completely; dependent variables y_{it} may be missing. In different applications the missing data mechanism can be explained by additional variables that are not of interest for the mean structure. Suppose that surrogate variables $\mathbf{z}_{it}$ are observed in addition to y_{it}, if y_{it} is observed.

Let $\mathbf{w}_{i0} = (\text{vec}(\mathbf{X}_i)', y_{i0}, \mathbf{z}_{i0}')'$ be comprised of covariates $\mathbf{X}_i$ and the observations of y_{i0} and $\mathbf{z}_{i0}$ prior to follow–up. We assume that y_{i0} and $\mathbf{z}_{i0}$ are observed completely. If $\mathbf{X}_i$ are not baseline covariates, they are assumed to be either fixed or independent of the response variables. Furthermore, we set $\mathbf{w}_{it} = (y_{it}, \mathbf{z}_{it}')'$ for $t = 1, \ldots, T$. Bars are used to indicate variables including the whole history except the current observation, thus $\bar{\mathbf{w}}_{it} = (\mathbf{w}_{i0}', \mathbf{w}_{i1}', \ldots, \mathbf{w}_{i(t-1)}')'$.

In the following, r_{it} denotes the missing data indicator, such that $r_{it} = 1$, if the pair $(y_{it}, \mathbf{z}_{it})$ is observed and 0 otherwise. We assume a monotone missing data pattern so that $r_{i(t+1)} = 0$, if $r_{it} = 0$ for any t. Thus, $r_{iT} = 1$ indicates that the data of cluster i are completely observed. We assume that the data are missing at random (MAR) in the sense of Laird (1988). This implies that the response probability at time t only depends on observations prior to t. We do, however, not assume that the data are missing completely at random (MCAR; Laird, 1988). This would imply that the probability for a response at time t may depend on the explanatory variables $\mathbf{X}_i$ but not on the history $\mathbf{w}_{it}$ observed up to t.

5 ESTIMATION OF RESPONSE PROBABILITIES

In the WEE method, observations are weighted by their inverse observation probability. In order to guarantee the existence of the estimates, we assume that the probability λ_{it} to remain in the study is greater than 0 for each study subject and each point:

$$\lambda_{it} = P(r_{it} = 1 | r_{i(t-1)} = 1, \bar{\mathbf{w}}_{it}) > \delta > 0\,. \tag{3}$$

The response probabilities $\lambda_{it} = \lambda_{it}(\boldsymbol{\gamma}|r_{i(t-1)}, \bar{\mathbf{w}}_{it})$ may depend on an additional parameter $\boldsymbol{\gamma}$ that is modeled as a function of the history up to t. If the response probabilities $\lambda_{it}(\boldsymbol{\gamma})$ are unknown, $\boldsymbol{\gamma}$ needs to be estimated. A natural choice to model the response probability as a function of $\boldsymbol{\gamma}$ is the logit link. If no observation is missing at a specific time point t, λ_{it} need not be estimated. In this situation, $\lambda_{it} = 1$. If at least one observation is missing and present, respectively, for every time point, an estimate $\hat{\boldsymbol{\gamma}}$ can be obtained by maximizing the partial likelihood function $L(\boldsymbol{\gamma}) = \prod_{i=1}^{n} \prod_{t=1}^{T} L_{it}(\boldsymbol{\gamma})$, where

$$L_{it}(\boldsymbol{\gamma}) = \begin{cases} \lambda_{it}(\boldsymbol{\gamma})^{r_{it}}[1 - \lambda_{it}(\boldsymbol{\gamma})]^{1-r_{it}} & \text{if} \quad r_{i(t-1)} = 1 \\ 1 & \text{if} \quad r_{i(t-1)} = 0 \end{cases}$$

is the core of a GLM for binary data if $r_{i(t-1)} = 1$. Note that cluster i does not contribute to the partial likelihood function at time t if $r_{i(t-1)} = 0$. In our application with HPV infection data, we used the logit link to establish the relationship between the response probability λ_{it} and $\boldsymbol{\gamma}$. We assume that the model for λ_{it} is correctly specified.

If the data are MAR, the product of the response probabilities including time t, $\pi_{it} = \pi_{it}(\boldsymbol{\gamma}) = \lambda_{i1}(\boldsymbol{\gamma}) \cdot \ldots \cdot \lambda_{it}(\boldsymbol{\gamma})$ may be interpreted as the conditional probability of observing cluster i at time t

given the entirely observed history $\bar{\mathbf{w}}_{it}$. In order to formulate the WEE, the conditional probabilities π_{it} together with their observational status r_{it} are collected in a diagonal matrix $\mathbf{\Pi} = \mathbf{\Pi}(\boldsymbol{\gamma})$ with elements r_{it}/π_{it}.

6 WEIGHTED ESTIMATING EQUATIONS

Both the assumption that the data are MAR and the positivity of the response probabilities from eq. (3) are the fundamental prerequisites to identify the marginal mean $E(y_{it}|\mathbf{X}_i)$. They imply

$$E(y_{it}r_{it}/\pi_{it}|\mathbf{X}_i) = E(y_{it}|\mathbf{X}_i)\,. \tag{4}$$

This is the key result required to formulate the WEE. It states that—given the explanatory variables—the mean individual time point specific response is equal to the mean individual time point specific response weighted by their inverse observation probability, if the observational status is taken into account.

The validity of eq. (4) can be shown easily for data that are MCAR. Here, the observational status r_{it} is independent from the response y_{it} so that the left hand side of (4) can be factorised to

$$E(y_{it}r_{it}/\pi_{it}|\mathbf{X}_i) = E(y_{it}|\mathbf{X}_i)E(r_{it}|\mathbf{X}_i)/\pi_{it}\,.$$

Finally, $E(r_{it}|\mathbf{X}_i) = P(r_{it} = 1|\mathbf{X}_i) = \pi_{it}$ because the missing data pattern is monotone. The more complex case of data that are MAR is discussed in detail by Robins et al. (1995).

The fundamental difference between the WEE and the IEE can be easily seen from (4): The WEE rest upon the left hand side of (4), while the IEE are based on the right hand side of (4). Correspondingly, the residuals $\boldsymbol{\epsilon}$ of the WEE are weighted by $\mathbf{\Pi}(\boldsymbol{\gamma})$, the inverse observation probability multiplied by the actual observation status. This weighting implies that data with missing response do not contribute to the WEE. Using the notation of section 2, an estimator $\hat{\boldsymbol{\beta}}_{WEE}$ is the solution of the WEE

$$\mathbf{u}(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}) = \mathbf{D}'\mathbf{V}^{-1}\mathbf{\Pi}(\hat{\boldsymbol{\gamma}})\boldsymbol{\epsilon} = \mathbf{0}\,. \tag{5}$$

Robins et al. (1995, Appendix A) have shown that $\hat{\boldsymbol{\beta}}_{WEE}$ is asymptotically normal with mean $\boldsymbol{\beta}$ under suitable regularity conditions. Its variance matrix can be strongly consistently estimated with variance matrix given in the Appendix. The reader should note that the usually applied robust variance matrix need not yield positive estimates of the variance matrix.

Robins and Rotnitzky (1995) have shown that prior knowledge concerning the response probabilities does not provide additional information, if (i) the mean structure μ_{it} is correctly specified, (ii) the data are MAR and (iii) the response probabilities λ_{it} are greater than 0. This result also implies that the the WEE may also be applied without loss of power even if the data are MCAR. This will be shown by a simple simulation study in the next section.

7 SIMULATION STUDY

In a simulation study using binary dependent variables we compared the properties of the WEE estimator with the IEE estimator that used the information of all observations (available case; AV) and the IEE estimator that used the information of all clusters without missing data (complete cluster; CC). The IEE estimator of the simulated data before generation of missings (full data estimator; FD) was used as reference.

The simulation proceeded as follows. First, the complete data set without missing observations was generated. Second, observations were deleted from the complete simulated data set using pre-specified missing data mechanisms. Several approaches for simulating correlated binary data have been proposed (Park, Park and Shin, 1996; Gange, Linton, Scott, DeMets and Klein, 1995; Lee, Scott and Soo, 1993; Emrich and Piedmonte, 1991). We did not fix the marginal moments in our approach but based our simulations on the theoretical work of Fitzmaurice and Laird (1993) which requires specification of conditional log-odds ratios. Given the conditional log-odds ratios, the joint multinomial distribution of the response patterns given the explanatory variables can be computed. The joint distribution is then used to determine the marginal moments. Random numbers were generated using the DRAND48 generator, which is supplied by SunOS 5.5 as a C-library function (SunOS, 1995, man Pages(3C)). The

seed was set using the system time for each experiment. Both the WEE and the IEE were solved by MAREG (Kastner et al., 1997).

We simulated 100 clusters with $t = 0, \ldots, 3$ each. $t = 0$ was used as baseline and was assumed to be always observed. We generated two different marginal models. In both models, the response depended on an intercept and a dummy-coded treatment variable. However, in the first model the intercept and the treatment effect were time-constant, while in the second model the treatment effect is time-varying. Furthermore, we use two different models for the dropout-process. In the first model the response probability at time t depended on a constant so that the data were MCAR. In the second model, the response probability at time t depended on a constant and the response at time $t-1$. This implies that the data were MAR. In both cases the parameters were chosen so that the dropout rate was about 10%, 20% and 30% at $t = 1, 2$ and 3, respectively. Missing data were generated using the logistic function. Finally, two different degrees of association were used. The first yielded a small correlation of about 0.1, while the second resulted in a high correlation of about 0.5.

Table II shows mean and standard deviations (in brackets) of the estimated parameters of the model with time-constant treatment effect. The theoretical parameters were $\beta_{\text{intercept}} = -0.5$ and $\beta_{\text{treatment}} = 1$. The drop-out process was modeled assuming $\gamma_{\text{intercept}} = 2$, if the data were MCAR. For data that were MAR, we used $\gamma_{\text{intercept}} = 1$ and $\gamma_{y_{(t-1)}} = 3$ as parameters. Table III displays the results for the model with time-varying treatment effects. Here we assumed $\beta_{\text{intercept}} = -0.5$, and $\beta_{\text{treatment}} = 0, 0.5, 1$ and 1 at $t = 0, 1, 2$ and 3, respectively. The same parameters as in the cluster-constant case were chosen for modeling the drop-out process.

Table II: Parameter estimates and robust standard errors (in parenthesis) from the simulations of the marginal model with cluster-constant covariates. FD: full data, WEE: weighted estimating equations, AV: available case, CC: complete clusters.

		$\rho = 0.1$				$\rho = 0.5$			
Dropout	Parameter	FD	WEE	AV	CC	FD	WEE	AV	CC
MCAR	intercept	-0.509	−0.509	−0.509	−0.506	−0.497	−0.495	−0.496	−0.500
		(0.162)	(0.174)	(0.172)	(0.193)	(0.237)	(0.252)	(0.250)	(0.289)
	treatment	1.017	1.017	1.016	1.011	0.999	0.998	0.997	1.006
		(0.229)	(0.251)	(0.250)	(0.279)	(0.329)	(0.351)	(0.347)	(0.403)
MAR	intercept	-0.511	−0.509	−0.475	−0.217	−0.500	−0.502	−0.313	0.059
		(0.170)	(0.199)	(0.186)	(0.218)	(0.237)	(0.263)	(0.250)	(0.291)
	treatment	1.010	1.007	0.987	0.965	1.013	1.014	0.975	0.936
		(0.230)	(0.262)	(0.247)	(0.287)	(0.321)	(0.370)	(0.337)	(0.401)

Both tables demonstrate that all approaches yielded consistent parameter estimates as long as the data are MCAR. In this situation, the WEE and the AV yielded virtually identical parameter estimates and standard errors. As expected, the FD estimator had the lowest standard errors across all models, while the CC estimator had the largest standard errors. Therefore, the CC approach is not recommended for applications even if the data are MCAR.

If the data were MAR, the IEE estimator using the AV or the CC approach yielded biased parameter estimates, while the WEE remained consistent. The bias increased with the correlation of the responses. Furthermore, the AV estimator and the CC estimator led to quite contrary results for some models. For example, the mean parameter estimates of $\beta_{\text{treatment}_3}$ were 1.258 and 0.895 for the AV and the CC approach, respectively, with a single time-dependent treatment effect and high correlation of $\varrho = 0.5$. Here, the AV approach resulted in an overestimation, while the CC approach yielded an underestimation of the true treatment effect.

Our results are comparable to those obtained by Robins et al. (1995) and Robins and Rotnitzky (1995) for continuous dependent variables. The bias was, however, not so pronounced in our simulations. This finding can be explained by the naturally lower variation of binary variables compared with continuous data.

Table III: Parameter estimates and robust standard errors (in parenthesis) from the simulations of the marginal model with time-varying covariates. FD: full data, WEE: weighted estimating equations, AV: available case, CC: complete clusters.

		$\rho = 0.1$				$\rho = 0.5$			
Dropout	Parameter	FD	WEE	AV	CC	FD	WEE	AV	CC
MCAR	intercept	-0.498	−0.496	−0.494	−0.494	−0.499	−0.498	−0.498	−0.496
		(0.165)	(0.181)	(0.181)	(0.205)	(0.232)	(0.251)	(0.249)	(0.284)
	treatment_0	-0.012	−0.017	−0.018	−0.021	−0.010	−0.012	−0.011	−0.022
		(0.336)	(0.345)	(0.345)	(0.427)	(0.385)	(0.396)	(0.394)	(0.467)
	treatment_1	0.499	0.498	0.497	0.496	0.503	0.498	0.498	0.501
		(0.349)	(0.374)	(0.375)	(0.421)	(0.376)	(0.401)	(0.399)	(0.458)
	treatment_2	1.005	1.007	1.006	1.009	1.013	1.024	1.025	1.018
		(0.348)	(0.393)	(0.393)	(0.429)	(0.379)	(0.433)	(0.432)	(0.469)
	treatment_3	0.998	1.004	1.003	1.002	1.001	1.009	1.009	1.007
		(0.332)	(0.398)	(0.398)	(0.409)	(0.377)	(0.452)	(0.451)	(0.470)
MAR	intercept	-0.508	−0.505	−0.473	−0.213	−0.499	−0.496	−0.311	0.052
		(0.174)	(0.206)	(0.192)	(0.222)	(0.237)	(0.264)	(0.253)	(0.298)
	treatment_0	-0.007	−0.009	−0.042	0.021	−0.011	−0.013	−0.198	−0.067
		(0.340)	(0.362)	(0.353)	(0.433)	(0.388)	(0.402)	(0.397)	(0.487)
	treatment_1	0.511	0.496	0.479	0.544	0.493	0.496	0.401	0.468
		(0.332)	(0.378)	(0.370)	(0.448)	(0.376)	(0.422)	(0.413)	(0.486)
	treatment_2	1.007	1.004	1.009	1.065	1.011	1.023	1.097	1.037
		(0.343)	(0.428)	(0.417)	(0.476)	(0.397)	(0.476)	(0.465)	(0.534)
	treatment_3	1.028	1.042	1.075	0.816	1.014	1.040	1.258	0.895
		(0.351)	(0.458)	(0.440)	(0.457)	(0.389)	(0.509)	(0.486)	(0.513)

8 RESULTS OF HUMAN PAPILLOMA VIRUS INFECTION DATA

In the first step of our analysis, we modeled the response probabilities for an infection with HPV using the data presented in section 2. The relationship between a response and possible covariates was established with the logit link. To evaluate the significance of the parameter estimates, we used model based standard errors since the WEE require a correct specification of the response probabilities. We hypothesized that the probability of taking part in an examination increases with the age of a study subject. We furthermore hypothesized that the probability is lower after delivery due to less free time of the study subject. It probably also depends on the HPV status lag 1, i.e. prior to the current observation, and the Vira-Pap status surrogate variable lag 1. Thus, if the study subject did not receive a clear-cut negative result at $t-1$, the probability to take part in the examination at t should be high.

Table IV displays the results for the corresponding weight model. Age lag 1 (OR for a 10 year age difference 1.553; 95% CI 1.135 – 2.125) and reproductive status lag 1 revealed p-values < 0.05 with regression coefficients in the expected direction (overall p for reproductive status 0.005; after delivery p = 0.003). Thus, the probability to take part in the examination was higher for older women. It was lower for study subjects after delivery (postpartum) compared with pregnant women and non-pregnant individuals (reference category). Unexpectedly, HPV infection status lag 1 had a p-value > 0.05. It remained, however, in the model since additional parameters in the model of the response probabilities do not decrease the efficiency of the parameter estimates of the mean structure (Robins and Rotnitzky, 1995, Theorem 1). Interestingly, the probability to take part in the examination depended on the Vira-Pap status surrogate variable lag 1 ($p = 0.047$). As expected, it was lower for an individual that either had a clear-cut negative HPV result at the last examination (OR 0.502; 95% CI 0.253 – 0.996). Since the Vira-Pap status had an influence on the response probability, the assumption that the data were MCAR was violated.

In the second step of our analysis, we modeled the parameters of interest for the mean structure using the data of section 2 and possible interactions between these variables. Our aim was to detect

Table IV: Odds ratios, 95% confidence intervals and p-values of the weighting model for the weighted estimating equation analysis of the human papilloma virus infection data

Variable	Odds ratio	Confidence interval	P-value
Intercept	2.090	0.739 – 5.905	0.164
Negative Vira-Pap status	0.502	0.253 – 0.996	0.049
HPV	0.490	0.156 – 1.541	0.222
Age (10 year difference)	1.553	1.135 – 2.125	0.005
Reproductive status			0.005
(pregnant)	1.150	0.690 – 1.919	0.591
(postpartum)	0.423	0.242 – 0.739	0.003

covariates that were potentially significant for an HPV infection. Therefore, we decided to include only those variables in the final model revealing p-values < 0.05. The reproductive status was coded as in the weighting model. Since there was no difference ($p > 0.05$) between non smokers and study subjects smoking maximal 19 cigarettes per day, these groups were combined. Thus, daily smoking of cigarettes was dichotomised for the analysis with categories 1 = "heavy smoker (more than 19 cigarettes per day)" and 0, otherwise.

Table V displays the final model for the HPV infection data. It was derived using backward selection for the WEE. Table V additionally shows the results for the available case analysis (AV). The complete cluster analysis (CC) did not converge, most likely due to the relatively low number of complete clusters. In accordance with the results of (Chang-Claude et al., 1996), age was a predictor for an HPV infection for both WEE and AV (OR 0.600 for WEE; 0.543 for AV). Thus, the odds for an HPV infection was lower for old individuals. The final model also included smoking for both estimation methods with a higher odds for an HPV infection for heavy smokers. The reproductive status revealed a p-value < 0.05 in the WEE analysis only ($p = 0.031$ for postpartum vs. non-pregnant). It was not significant at the 5% test-level ($p = 0.087$) upon use of the AV analysis. This may be due to the fact that the data are not MCAR. Pregnant women had no significantly increased risk for an HPV infection compared with study subjects that were non-pregnant ($p = 0.688$ for WEE). However, the odds for an HPV infection for pregnant women compared with those postpartum was significantly higher ($p = 0.027$ for WEE; $p = 0.083$ for AV).

Table V: Odds ratios (OR), 95% confidence intervals (CI) and p-values of the weighted estimating equations (WEE) and the available case analysis (AV) for the human papilloma virus (HPV) infection data

	WEE			AV		
Variable	OR	CI	P-value	OR	CI	P-value
Intercept	0.156	0.043 – 0.565	0.005	0.235	0.058 – 0.950	0.042
Age (10 year difference)	0.600	0.406 – 0.889	0.013	0.543	0.353 – 0.836	0.005
Heavy smoker	3.235	1.112 – 9.414	0.031	3.025	1.056 – 8.667	0.039
Reproductive status			0.081			0.214
(pregnant)	1.176	0.533 – 2.596	0.688	1.053	0.473 – 2.344	0.899
(postpartum)	0.139	0.023 – 0.831	0.031	0.165	0.021 – 1.298	0.087

The influence of one cluster on the parameter estimates was checked with the robust COOK-statistic (Ziegler and Grömping, 1998) which is given in the Appendix. Two study participants had a markedly influence on the parameter estimates. The robust COOK-statistic of individuals 124 and 877 were 1.64 and 1.77, respectively. After elimination of these subjects, neither the daily number of smoked cigarettes nor the reproductive status revealed p-values < 0.05. These two study participants were the only individuals that had a positive HPV status at almost all investigations. Therefore, they had a high influence on the parameter estimates.

9 DISCUSSION

A variety of methods have been established for the analysis of longitudinal data with non-normal dependent variables. They have primarily been developed for complete data. Many studies, however, suffer

from missing response data. Therefore, there is a growing interest to adequately deal with missing data. Basically three different approaches can be distinguished. The first group consists of maximum likelihood methods for the estimation of the marginal mean (Diggle and Kenward, 1994; Fitzmaurice and Laird, 1993; Fitzmaurice, Laird and Lipsitz, 1994; Fitzmaurice, Clifford and Heath, 1996). However, their consistency relies on the correct specification of the joint distribution of the clustered responses. This may be crucial if the aim is to consistently estimate the marginal mean so that second—the correlation—and higher order moments of the responses are considered as nuisance.

Two well-known alternatives are imputation methods (Heyting, Tolboom and Essers, 1992; Paik, 1997; Xie and Paik, 1997) and weighting methods (Flanders and Greenland, 1991; Zhao, Prentice and Self, 1992; Robins, 1994; Robins et al., 1995; Robins and Rotnitzky, 1995; Rotnitzky and Robins, 1995). They are different, in general. However, imputation and weighting yield identical results under specific circumstances (Little, 1986; Paik, 1997). The weighting methods are more attractive than the imputation methods because the weight—the inverse observation probability—is not a function of the marginal mean parameters, as pointed out by Paik (1997). Furthermore, the WEE have optimality properties so that they result in semiparametric efficient parameter estimates under suitable regularity conditions (Robins and Rotnitzky, 1995). The consistency of the estimates, however, relies on the correct specification of the response probabilities and the marginal mean.

Though the WEE are appealing, they have been criticized by Carlin and co-workers (Carlin et al., 1999). These authors stated that the WEE were unavailable in an accessible form with current computer software. This obviously limits the application of any method in practice. Recently, the WEE for monotone missing data patterns have been implemented in the computer package MAREG (Kastner et al., 1997) which can be retrieved from the Web. We have compared the WEE with the CC and the AV which are usually applied in practice (s. Section 7). We showed for the simulated models that the WEE, the CC and the AV yielded virtually identical results as long as the data were MCAR. Both the CC and the AV estimators were biased when the responses were MAR but not MCAR. The biased increased with the correlation of the responses. Only the WEE estimator was consistent in this situation.

We employed the WEE to HPV infection data with monotone missing responses. To our knowledge, it was the first application of the WEE to real data with missing binary responses. First of all, we showed by modeling of response probabilities that the data were not MCAR. Thus, the AV and the CC estimator need not be consistent. Interestingly, the CC data were not even estimable, most likely due to the low number of complete clusters. The WEE and the AV estimator yielded different results: the risk factor of primary interest—the reproductive status—was not significant at the 5% test-level in the final model of the marginal mean, while it revealed a p-value < 0.05 in the WEE model. Since the data were not MCAR, we consider the WEE estimates to be the more reliable. We are, however, aware of the fact that the reproductive status was only marginally significant after extensive model building. Furthermore, we performed a re-analysis of the HPV infection data. Specifically, the WEE method was not even mentioned in the study protocol. Nevertheless, the seemingly conflicting results concerning the risk for an HPV infection by pregnancy could be explained as follows. If the sample of non-pregnant women mainly consists of those women that had never been pregnant, the risk for an HPV infection of pregnant women need not be significant. However, if it mainly consists of women after delivery, pregnancy could well show an increased risk for an HPV infection. Finally, the robust Cook-statistic showed that the significance of the reproductive status relied on the inclusion of only two study subjects who had an HPV infection at almost all investigations. This strong influence of two clusters is due to the fact that an HPV infection is a relatively rare event.

Summing up, the WEE is an appropriate method to deal with missing responses for longitudinal data. They have been implemented for monotone missing data patterns and are available to the public. To be applicable in a broad range of applications, however, an extension to arbitrary missing data patterns for the responses is required.

APPENDIX

The variance estimator of $\hat{\boldsymbol{\beta}}_{WEE}$ for the WEE has been proposed by Robins et al. (1995) and can be obtained by applying the chain rule together with a standard Taylor series expansion. Using the notation from sections 3 and 6, it may strongly consistently estimated by

$$\widehat{V(\hat{\boldsymbol{\beta}}_{WEE})} = \mathbf{A}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^{-1}\mathbf{C}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})\mathbf{A}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})^{-1}, \tag{6}$$

where

$$\mathbf{A}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) = \hat{\mathbf{D}}'\hat{\mathbf{V}}^{-1}\boldsymbol{\Pi}(\hat{\boldsymbol{\gamma}})\hat{\mathbf{D}} \quad \text{and} \quad \mathbf{C}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \widehat{\text{resid}}_i \; \widehat{\text{resid}}_i' \; .$$

$-\mathbf{A}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ is the estimated Fisher information matrix and symmetric because $\hat{\mathbf{V}}$ and $\boldsymbol{\Pi}(\hat{\boldsymbol{\gamma}})$ are diagonal matrices. $\mathbf{C}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ is similar to the estimated outer product gradient (OPG) of the usual robust variance estimator. However, it is more complex since the WEE (5) depend on an additionally estimated nuisance parameter $\hat{\boldsymbol{\gamma}}$. Therefore, $\mathbf{u}_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$, being the usual term for the estimated OPG, needs to be corrected by $\mathbf{u}_i(\hat{\boldsymbol{\gamma}})$. Thus, $\widehat{\text{resid}}_i = \widehat{\text{resid}}[\mathbf{u}_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}), \mathbf{u}_i(\hat{\boldsymbol{\gamma}})]$ is the estimated residual from the linear regression of $\mathbf{u}_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ on $\mathbf{u}_i(\hat{\boldsymbol{\gamma}})$:

$$\widehat{\text{resid}}_i = \mathbf{u}_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}) - \Big(\sum_{i=1}^{n} \mathbf{u}_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\mathbf{u}_i(\hat{\boldsymbol{\gamma}})'\Big)\Big(\sum_{i=1}^{n} \mathbf{u}_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})\mathbf{u}_i(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})'\Big)^{-1}\mathbf{u}_i(\hat{\boldsymbol{\gamma}}).$$

Model checking can be performed using the robust COOK-statistic (Ziegler, Blettner, Kastner and Chang-Claude, 1998). For the ith cluster it is given by

$$\Big(\hat{\boldsymbol{\beta}}_{WEE(-i)} - \hat{\boldsymbol{\beta}}_{WEE}\Big)' \, V(\widehat{\hat{\boldsymbol{\beta}}_{WEE}}) \, \Big(\hat{\boldsymbol{\beta}}_{WEE(-i)} - \hat{\boldsymbol{\beta}}_{WEE}\Big) \; ,$$

where $\hat{\boldsymbol{\beta}}_{WEE(-i)}$ is the fully-iterated deletion-one WEE estimator.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft.

REFERENCES

Barber, S. R., Werdel, J., Symbula, M., Williams, J., Burkett, B. A. and Taylor, P. T. (1992). Serore-activity to HPV-16 proteins in women with early cervical neoplasia, *Cancer Immunol Immunother* **35**: 33–38.

Carlin, J. B., Wolfe, R., Coffey, C. and Patton, G. C. (1999). Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods: Prevalence and incidence of smoking in an adolescent cohort, *Statistics in Medicine* **18**: 2655–2679.

Chang-Claude, J., Schneider, A., Smith, E., Blettner, M., Wahrendorf, J. and Turek, L. (1996). Longitudinal study of the effects of pregnancy and other factors on detection of HPV, *Gynecologic Oncology* **60**: 355–362.

Crowder, M. J. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measurements, *Biometrika* **82**: 407–410.

Diggle, P. J. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis, *Applied Statistics* **43**: 49–94.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates, *The American Statistician* **45**: 302–304.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data., *Biometrics* **51**: 309–317.

Fitzmaurice, G. M., Clifford, P. and Heath, A. F. (1996). Logistic regression models for binary panel data with attrition, *Journal of the Royal Statistical Society, Series A* **159**: 249–263.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses, *Biometrika* **80**: 141–151.

Fitzmaurice, G. M., Laird, N. M. and Lipsitz, S. R. (1994). Analysing incomplete longitudinal binary responses: A likelihood-based approach, *Biometrics* **50**: 601–612.

Flanders, W. D. and Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs, *Statistics in Medicine* **10**: 739–747.

Gange, S. J., Linton, K. L. P., Scott, A. J., DeMets, D. L. and Klein, R. (1995). A comparison of methods for correlated ordinal measures with ophtalmic applications, *Statistics in Medicine* **14**: 1961–1974.

Garry, R. and Jones, R. (1985). Relationship between cervical condylomata, pregnancy and subclinical papillomavirus infection, *J. Reprod. Med.* **30**: 393–399.

Gissmann, L., Wolnik, L., Ikenberg, H., Koldovsky, U., Schnurch, H. G. and zur Hausen, H. (1983). Human papillomaviruses 6 and 11 DNA sequences in genital and larygeal papillomas and in some cervical cancers, *Proc. Natl. Acad. Sci. USA* **80**: 560–563.

Hanley, J. A., Negassa, A. and deB. Edwardes, M. D. (2000). GEE analysis of negatively correlated binary responses: a caution, *Statistics in Medicine* **19**: 715–722.

Heyting, A., Tolboom, J. T. B. M. and Essers, J. G. A. (1992). Statistical handling of drop-outs in longitudinal clinical trials, *Statistics in Medicine* **11**: 2043–2061.

Kastner, C., Fieger, A. and Heumann, C. (1997). MAREG and WinMAREG—a tool for marginal regression models, *Computational Statistics and Data Analysis* **24**: 235–241. URL: (http://www.stat.uni-muenchen.de/∼andreas/mareg/winmareg.html)

Koutsky, L. A., Holmes, K. K., Critchlow, C. W., Stevens, C. E., Paavonen, J. and Beckmann, A. M. (1992). A cohort study of the risk of cervical intreapithelial neoplasia grade 2 or 3 in relation to papillomavirus infection, *New England Journal of Medicine* **327**: 1272–1278.

Laird, N. M. (1988). Missing data in longitudinal studies, *Statistics in Medicine* **7**: 305–315.

Lee, A. J., Scott, A. J. and Soo, S. (1993). Comparing Liang-Zeger estimates with maximum likelihood in bivariate logistic regression, *Statistical Computation and Simulation* **44**: 133–148.

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

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means, *International Statistical Review* **54**: 139–158.

Little, R. J. A. and Schenker, N. (1995). Missing data, *in* G. Arminger, C. C. Clogg and M. E. Sobel (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Plenum, New York, pp. 39–75.

Mancl, L. A. and Leroux, B. G. (1996). Efficiency of regression estimates for clustered data, *Biometrics* **52**: 500–511.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random, *Journal of the American Statistical Association* **92**: 1320–1329.

Park, C. G., Park, T. and Shin, D. W. (1996). A simple method for generating correlated binary variates, *Journal of the American Statistical Association* **50**: 306–310.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data, *Communications in Statistics, Part B—Simulation and Computation* **23**: 939–951.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models, *Communications in Statistics, Part A—Theory and Methods* **23**: 2379–2412.

Robins, J. M. and Rotnitzky, A. G. (1995). Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association* **90**: 122–129.

Robins, J. M., Rotnitzky, A. G. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**: 106–120.

Rotnitzky, A. G. and Robins, J. M. (1995). Semiparametric estimation of models for means and covariances in the presence of missing data, *Scandinavian Journal of Statistics* **22**: 323–333.

Schneider, A., Hotz, M. and Gissman, L. (1987). Increased prevalence of human papillomaviruses in the lower genital tract of pregnant women, *International Journal of Cancer* **40**: 198–201.

SunOS (1995). *SunOS Reference Manual*, Sun Microsystems, Mountain View.

Xie, F. and Paik, M. C. (1997). Multiple imputation methods for the missing covariates in generalized estimating equation, *Biometrics* **53**: 1538–1546.

Zeger, S. L., Liang, K.-Y. and Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates, *Biometrika* **72**: 31–38.

Zhao, L. P., Lipsitz, S. R. and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations, *Biometrics* **52**: 1165–1182.

Zhao, L. P., Prentice, R. L. and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model, *Journal of the Royal Statistical Society, Series B* **54**: 805–811.

Ziegler, A., Blettner, M., Kastner, C. and Chang-Claude, J. (1998). Identifying influential families using regression dianostics for generalized estimating equations, *Genetic Epidemiology* **15**: 341–353.

Ziegler, A. and Grömping, U. (1998). The generalised estimating equations: A comparison of procedures available in commercial statistical software packages, *Biometrical Journal* **40**: 245–260.

Ziegler, A., Kastner, C. and Blettner, M. (1998). The generalised estimating equations: An annotated bibliography, *Biometrical Journal* **40**: 115–139.

zur Hausen, H. (1991). Viruses in human cancers, *Science* **254**: 1167–1173.