American Journal of Epidemiology

Vol. 147, No. 7
Printed in U.S.A.

Comparison of Population-Averaged and Subject-Specific Approaches for Analyzing Repeated Binary Outcomes

Frank B. Hu,[1] Jack Goldberg,[2] Donald Hedeker,[2,3] Brian R. Flay,[3] and Mary Ann Pentz[4]

Several approaches have been proposed to model binary outcomes that arise from longitudinal studies. Most of the approaches can be grouped into two classes: the population-averaged and subject-specific approaches. The generalized estimating equations (GEE) method is commonly used to estimate population-averaged effects, while random-effects logistic models can be used to estimate subject-specific effects. However, it is not clear to many epidemiologists how these two methods relate to one another or how these methods relate to more traditional stratified analysis and standard logistic models. The authors address these issues in the context of a longitudinal smoking prevention trial, the Midwestern Prevention Project. In particular, the authors compare results from stratified analysis, standard logistic models, conditional logistic models, the GEE models, and random-effects models by analyzing a binary outcome from two and seven repeated measurements, respectively. In the comparison, the authors focus on the interpretation of both time-varying and time-invariant covariates under different models. Implications of these methods for epidemiologic research are discussed. *Am J Epidemiol* 1998;147:694–703.

generalized estimating equations; longitudinal studies; random-effects regression models; repeated measurement

Repeated measures designs or longitudinal studies occupy an important role in clinical and epidemiologic research. In a clinical trial, for example, patients may be randomly assigned to different treatment conditions and repeatedly classified in terms of presence or absence of clinical improvement, side effects, or specific symptoms. Most adolescent smoking prevention trials involve a baseline measure of smoking behaviors and related factors, followed by an intervention program, an immediate posttest, and several follow-up measurements. This kind of design offers the opportunity to study the time course of change and the long-term effects of the treatment or intervention (1). It also offers increased statistical power and robustness for model selection (2).

Received for publication February 18, 1997, and accepted for publication September 20, 1997.

Abbreviations: CI, confidence interval; GEE, generalized estimating equations.

[1] Department of Nutrition, Harvard School of Public Health, Boston, MA.
[2] Epidemiology and Biostatistics Division, School of Public Health, University of Illinois at Chicago, Chicago, IL.
[3] Prevention Research Center, University of Illinois at Chicago, Chicago, IL.
[4] Department of Prevention Medicine, University of Southern California, Los Angeles, CA.
Reprint requests to Dr. Frank B. Hu, Dept. of Nutrition, Harvard School of Public Health, 655 Huntington Ave., Boston, MA 02115.

Analyses of repeated measures data need to accommodate the statistical dependence among the repeated observations within subjects. For normally distributed data, variants of the random-effects model of Laird and Ware (3) are commonly used, and software for estimating this class of models is now widely available, such as SAS Proc Mixed (4), BMDP5V (5), HLM (6), and MLn (7).

Recently, several models have been proposed to model binary outcomes that arise from repeated measures designs. Most of the models can be grouped into two classes (8, 9): the "subject-specific" and the "population-averaged" approaches. Random-effects logistic models (10, 11) are commonly used to estimate subject-specific effects, while the generalized estimating equations (GEE) method of Liang and Zeger (12) is usually used to provide population-averaged effects.

While both the GEE and random-effects approaches are extensions of models for independent observations to time-dependent data, they address the problem of time-dependency differently. Also, the regression coefficients or odds ratios obtained from the two approaches are numerically different, as are their interpretations (8, 13, 14). Although comparisons of the two approaches have appeared in the statistical literature (9, 15–18), there has been little research on such comparisons that is accessible to public health re-

searchers. In addition, researchers in substantive areas are often unclear as to how these models relate to statistical methods or models with which they are already familiar (e.g., stratified analysis or the standard logistic model). The purposes of this paper include 1) describing and comparing the main features of random-effects logistic models and GEE logistic models in the context of population-averaged and subject-specific approaches; 2) showing the connection between these models and conventional epidemiologic methods, such as stratified analysis and standard logistic models; and 3) illustrating the use and interpretation of random-effects and GEE logistic models in comparison with each other and in comparison with usual logistic models by re-analyzing a longitudinal smoking prevention dataset with multiple timepoints.

The remainder of this paper is organized as follows. We first discuss the main features of population-averaged and subject-specific approaches for binary data. Then we describe the dataset from a smoking and drug abuse prevention study, the Midwestern Prevention Project (MPP) (19). To compare models, we first analyze the data considering only two timepoints (baseline and one-year follow-up). Finally, we present and compare results from standard logistic, GEE, and random-effects models by analyzing data from seven timepoints. We discuss the implications of these methods for epidemiologic research. In the Appendix, we provide details of the programs we used to fit the GEE and random-intercept logistic models in this paper.

THE POPULATION-AVERAGED APPROACH

To estimate treatment effects in a longitudinal prevention trial, we consider the following model:

$$\text{logit}(\mu_{ij}) = \log\frac{\Pr(Y_{ij} = 1)}{1 - \Pr(Y_{ij} = 1)} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_i \quad (1)$$

$$\text{Var}(Y_{ij}) = \mu_{ij}(1 - \mu_{ij}), \quad (2)$$

where Y_{ij} denotes a binary outcome (i.e., smoking, 0 = no and 1 = yes) for subject i at time j, $\mu_{ij} = E(Y_{ij})$ denotes the expectation or the mean of the response, x_i denotes the treatment group (0 = control, 1 = treatment) for subject i, and t_{ij} denotes the time corresponding to the jth measurement for subject i. Equation 1 assumes a linear relation between the log odds of response and both time and treatment group; equation 2 indicates that the variance of the binary response is a known function of its mean. In this model, $\exp(\beta_1)$ is the odds of an event at time j divided by the odds at time $j - 1$, controlling for treatment group, while $\exp(\beta_2)$ is the odds of the event among subjects in the treatment group divided by the odds among subjects in the control group, controlling for time. Because both $\exp(\beta_1)$ and $\exp(\beta_2)$ are ratios of subpopulation risk, they are referred to as population-averaged effects. In other words, the estimate of time effects (β_1) does not distinguish between observations belonging to the same or different subjects.

Traditional epidemiologic methods such as stratified analysis (i.e., Mantel-Haenszel method (20)) and standard logistic models for independent binary outcomes (21) are essentially population-averaged approaches. However, these methods are usually not appropriate for correlated binary outcomes arising from longitudinal studies due to the dependency of the repeated measurements. For longitudinal data, Liang and Zeger (12) proposed the generalized estimating equations (GEE) approach, which is an extension of generalized linear models (GLM) (22), to estimate the population-averaged estimates while accounting for the dependency between the repeated measurements. Specifically, the dependency or correlation between repeated measures is taken into account by robust estimation of the variances of the regression coefficients. In fact, the GEE approach treats the time dependency as a nuisance, and a "working" correlation matrix for the vector of repeated observations from each subject is specified to account for the dependency among the repeated observations. The "working correlation" is assumed to be the same for all subjects, reflecting *average dependence* among the repeated observations over subjects. Several "working" correlation structures can be specified, including independent, exchangeable, autoregressive, and unstructured. An independent working correlation assumes zero correlations between repeated observations. An exchangeable working correlation assumes uniform correlations across time. An autoregressive working correlation assumes that observations are only related to their own past values through first or higher order autoregressive (AR) process. An unstructured working correlation assumes unconstrained pairwise correlations. Liang and Zeger (12) show that under the assumption of missing completely at random (MCAR, discussed below), the GEE approach provides consistent estimators of the regression coefficients and of their robust variances even if the assumed working correlation is misspecified.

Estimation of the standard logistic model is equivalent to GEE estimation with an independent working correlation structure. With repeated binary outcomes, the standard logistic model yields the same population-averaged estimates as the GEE. However, the standard errors from the standard logistic models are biased because the independence assumption is violated. The

biases are dependent on whether the covariates vary with time (23, 24). Regression models ignoring the time dependency tend to overestimate the standard errors of time-varying covariates and underestimate the standard errors of time-invariant covariates.

THE SUBJECT-SPECIFIC APPROACH

In contrast to the population-averaged approach, the subject-specific approach can distinguish observations belonging to the same or different subjects. Random-effect models are commonly used to estimate subject-specific effects. Two methods can be used to estimate the subject-specific effects in the random-effects models: maximum likelihood and conditional likelihood procedures (25).

For repeated binary responses, logistic regression models (10, 11), probit regression models (26), and log-linear type models (27) have been proposed. A simple random-effects logistic model for estimating treatment effects in a longitudinal prevention trial can be written as:

$$\text{logit } P(Y_{ij} = 1|\upsilon_i) = (\beta_0 + \upsilon_i) + \beta_1 t_{ij} + \beta_2 x_i, \quad (3)$$

where υ_i is the random subject deviation.

This model is a generalization of the standard logistic model in which the intercept (deviation), υ_i, is allowed to vary with subjects. Thus, the model is often called a random-intercept logistic model. The random effect is usually assumed to be distributed as N (0, σ_υ^2). The fixed part β_0 represents the log odds of the response ($Y_{ij} = 1$) for a control subject (i.e., $x = 0$) at baseline (i.e., $t = 0$) with random effect $\upsilon_i = 0$.

By including a random-intercept in the model, the interdependencies among the repeated observations within subjects are explicitly taken into account. The variance-covariance structure in a random-intercept logistic model is analogous to the "compound symmetry" form assumed in a mixed-model analysis of variance (ANOVA) and the "exchangeable working correlation" in the GEE model described previously. As with the standard logistic model, likelihood ratio tests and/or Wald tests can be used to assess the treatment or time effects. The test in the random-effects model, however, is performed while accounting for the interdependency among the repeated observations within subjects.

The interpretation of e^β for a binary independent variable in a random-effects logistic model is somewhat different from that in a standard logistic model (9). With the standard logistic model, the baseline risk is simply the proportion of positive responses in the control group at baseline (e^{β_0}), while in the random-intercept logistic model, the baseline risk is assumed to follow a distribution ($e^{\beta_0+\upsilon_i}$). Therefore, the corresponding change in absolute risk with and without the covariate varies from one subject to another, depending on the baseline rate. Consequently, the odds ratios estimated from a random-effects logistic model additionally adjust for heterogeneity of the subjects, which can be considered to be due to unmeasured variables such as genetic predisposition and unobserved influences of social environmental factors. In the example from the smoking prevention project that follows, we can think of the random intercept as a subject's propensity to smoke (across the study timepoints) that is independent of the effects of the model covariates. In a random intercept model, this propensity is assumed to be constant across time. However, by including additional random subject-effects into the model, this propensity can vary across time. The unobserved propensity to smoke can reflect subjects' different genetic predispositions and/or unmeasured social environmental influences particular to a given subject. For this reason, the random effects are sometimes thought of as an omitted subject-varying covariate (9). Further discussion on the philosophical detail concerning use of random effects models can be found in Longford (28).

Because the estimated effects are adjusted for individual differences, they are often termed "subject-specific" effects. Therefore, the odds ratios estimated from the random-effects models should be interpreted in terms of the change due to the covariates for a single individual (or, more specifically, individuals with the same level on the random subject effect υ_i) even if the variable is indeed a between-subjects factor such as treatment group (8). Thus, the random-effects model is most useful when inference about individual differences is of major interest.

The interpretation of a within-subject factor such as time is further complicated by missing data. For a balanced design (i.e., no missing data), the time effects are purely within-subject effects. With missing data, however, the time effects include both between-subject and within-subject components. Generally this does not impose a problem for interpretation if certain missing data assumptions are met. As Laird (29) points out, random-effects models using maximum likelihood estimation provide valid inferences in the presence of ignorable non-response. By ignorable non-response, it is meant that the probability of missingness is dependent on measured covariates and/or previously observed values of the outcome. In this situation, the time trend for subjects with missing observations is estimated by "borrowing" strength from subjects with the same or similar characteristics.

While maximum marginal likelihood methods can

be used to estimate the parameters of the general random-effects logistic model (30), for the random-intercept model, some of the parameters can be estimated using conditional likelihood methods (25). To obtain a consistent estimate of the vector of coefficients β, the conditional likelihood is given by:

$$\Pi_i(1 + \exp(-D_i'\beta))^{-1}, \quad (4)$$

where D_i is the within-subject difference in terms of the covariates. Through the conditional likelihood approach, the β_{0+v_i} parameters in equation 3 are removed by conditioning on their sufficient statistics in the likelihood. In addition, this method only uses data from observations that are discordant on both response and the covariates (e.g., time). As a result, it cannot be used to estimate the effects of a time-invarying covariate such as treatment condition. However, the model can estimate the interaction effects between time and treatment condition. For data with only two timepoints, the estimated odds ratio of time reduces to the ratio of discordant pairs (31). Because conditional likelihood is unaffected by the sampling scheme (i.e., retrospective vs. prospective sampling), it can be used in studies of either case-control or cohort forms (32). The random-effects approach, however, may be biased under case-control sampling (32).

EXAMPLE: THE MIDWESTERN PREVENTION PROJECT

The data of interest were collected in a longitudinal smoking and drug abuse prevention trial, the Midwestern Prevention Project. The design of this project and the intervention methods have been described in detail elsewhere (19, 33). Briefly, the project is a longitudinal school- and community-based trial of prevention of cigarette smoking and drug abuse in adolescents. The data used in this illustration focus on a longitudinal panel of youth from eight schools in the Kansas City Standard Metropolitan Statistical Area who were tracked individually from 1984 to 1992. Subjects were sixth and seventh grade students who were evaluated in a baseline measurement in Fall 1984 (n = 1,607). The subsample included in this study was 1,002 individuals who were randomly selected to be followed up to adulthood (502 in the treatment group and 500 in the control group, with assignment by school). About half of the sample were males and 82.2 percent were whites; 40 percent were in grade 6 and 60 percent were in grade 7 when the program started. Self-administered questionnaires were used to assess students' smoking behaviors and related factors. The questionnaires were filled out by students in classrooms prior to intervention in Fall 1984, again at 6 months, and at one year. Subsequent assessment was conducted annually until 1992. For self-administered surveys, carbon monoxide measure of cigarette smoking, using the MiniCO Indicator (Catalyst Research Corporation, Owings Mills, Maryland), was used to increase the accuracy of self-reported smoking data.

For the purpose of illustration, only one item of several measurements of smoking behaviors is considered, that is, smoking in the month prior to the survey (coded as 0 = no, 1 = yes; it is simply called "monthly smoking" in the remainder of the paper). This study focuses on the data from the first seven timepoints of measurements (from baseline to 5-year follow-up) and, for simplicity, analyzes data only for those students with complete data across all seven waves (n = 682). Also, for the sake of simplicity, we ignore the intraclass correlations due to classrooms and schools (see references 34 and 35).

Table 1 treats the seven waves of measurement as repeated cross-sectional surveys and displays the prevalence of monthly smoking in the treatment and control groups. In general, the prevalence of smoking increased across time for both groups. However, the trend is not the same for the two groups. For the first five timepoints, the rate of increase is greater for the control group than for the treatment group. For the last two timepoints, the prevalence leveled off in the control group but continued to rise in the treatment group.

Illustrative analyses using population-averaged approaches

To illustrate the connections between GEE and stratified analysis, we first analyze smoking data from only two timepoints (baseline and one year). Table 2 shows the data laid out in contingency tables stratified by time and treatment group. The Mantel-Haenszel odds ratio for the effects of treatment group on smoking is 0.96 (95 percent confidence interval (CI) 0.68–1.36), indicating no significant differences in smoking between the two groups after controlling for time. In this calculation, we simply ignore the lack of independence in subjects' responses at the two timepoints. The

TABLE 1. Percent prevalence of monthly smoking in the last month across time by treatment group: the Midwestern Prevention Project, 1984–1992

Timepoint	Control group (n = 318)	Treatment group (n = 364)
Baseline (T1)	5.7	9.6
6 months (T2)	6.9	10.2
1 year (T3)	15.7	11.0
2 years (T4)	25.5	15.1
3 years (T5)	31.1	20.3
4 years (T6)	33.0	30.2
5 years (T7)	34.9	35.4

TABLE 2. Contingency tables for smoking data at baseline and one year: the Midwestern Prevention Project, 1984–1992

	Smoking		
	No	Yes	Total
I. Smoking and treatment group stratified by time			
Baseline			
Control	300	18	318
Treatment	329	35	364
Total	629	53	682
OR* = 1.77			
95% CI* 0.99–3.18			
One year			
Control	268	50	318
Treatment	324	40	364
Total	592	90	682
OR = 0.66			
95% CI 0.42–1.03			
OR_{MH}* = 0.96			
95% CI 0.68–1.36			
II. Smoking and time stratified by treatment group			
Control group			
Baseline	300	18	318
One year	268	50	318
Total	568	68	636
OR = 3.11			
95% CI 1.81–5.35			
Treatment group			
Baseline	329	35	364
One year	324	40	364
Total	653	75	728
OR = 1.16			
95% CI 0.72–1.87			
OR_{MH} = 1.80			
95% CI 1.27–2.56			

* OR, odds ratio; CI, confidence interval; OR_{MH}, Mantel-Haenszel odds ratio.

Mantel-Haenszel odds ratio comparing one year with baseline smoking is 1.80 (95 percent CI 1.27–2.58), indicating a significant increase in smoking from baseline to one year after controlling for treatment group. The Breslow-Day test for interaction is significant ($\chi^2 = 6.95$, degrees of freedom (df) = 1, $p < 0.01$), suggesting that the time effects are not the same for the treatment and control groups.

We estimated two logistic models (table 3), one with only main effects of time and treatment group and the other including a time by group interaction term so that the results can be directly compared to those from the stratified analyses described above. As expected, results from the logistic models are nearly identical to those from the stratified analyses. The interaction effect of treatment group and time is significantly less than zero, suggesting that the increase in smoking is greater in the control group than in the treatment group. In the standard logistic regression analyses, again, the time dependency is simply ignored; the repeated observations from the same individual are treated as independent observations.

Table 3 also lists results from GEE logistic models. While the estimates from the GEE models are similar to those from the standard logistic models, the standard errors are different. The standard error for the treatment group is smaller in the standard logistic model, whereas the standard errors for time and the interaction are smaller in the GEE models. Thus, the test statistics (Wald χ^2) for time and the interaction term from the GEE models are noticeably larger than those from the standard logistic models.

Illustrative analyses using subject-specific approaches

As discussed above, the random-effects approach is closely related to the matched-pair method in epidemiology, such as the conditional likelihood approach. To illustrate this point, we lay out the joint distributions of the responses for all subjects, by treatment and group (table 4). This setup is analogous to a matched study in which the repeated observations for each

TABLE 3. Standard logistic regression and GEE* logistic regression analyzing the effects of treatment group and time on smoking (baseline and one-year data): the Midwestern Prevention Project, 1984–1992

Models	β	SE*	Wald χ^2	Odds ratio	95% CI*
I. Standard logistic models†					
Treatment group (0 = control, 1 = treatment)	−0.04	0.18	0.06	0.96	0.68–1.36
Time (0 = baseline, 1 = one year)	0.59	0.18	10.5	1.80	1.26–2.58
Test for interaction	−0.99	0.38	6.83		
II. GEE logistic models†,‡,§					
Treatment group (0 = control, 1 = treatment)	−0.09	0.20	0.20	0.91	0.61–1.36
Time (0 = baseline, 1 = one year)	0.59	0.15	15.5	1.80	1.34–2.43
Test for interaction	−0.99	0.31	10.2		

* GEE, generalized estimating equations; SE, standard error; CI, confidence interval.
† The effects of treatment group and time are estimated from the main effects model.
‡ Exchangeable, independent, and unspecified "working" correlations give near-identical estimates and SEs.
§ SEs for GEE models are robust SEs.

TABLE 4. Matched pair contingency tables for smoking data at baseline and one year: the Midwestern Prevention Project, 1984–1992

Smoking at baseline	Smoking at one year		
	No	Yes	Total
I. Joint distribution of smoking at baseline and one year			
No	566	63	629
Yes	26	27	53
Total	592	90	682
McNemar test: OR_{time}* = 2.42 95% CI 1.53–3.82			
II. Smoking at baseline and one year stratified by treatment group			
Control group			
No	262	38	300
Yes	6	12	18
Total	268	50	318
OR_{time} = 6.33 95% CI 2.68–14.97			
Treatment group			
No	304	25	329
Yes	20	15	35
Total	324	40	364
OR_{time} = 1.25 95% CI 0.69–2.25			
Test for treatment × time interaction: χ^2=10.21, df*=1, $p < 0.01$			

* OR_{time}, matched odds ratio for time; CI, confidence interval; df, degrees of freedom.

subject at two timepoints can be thought of as a matched pair. Because each "pair" is matched on treatment group, the group effects cannot be examined using the matched-pair method. However, because time is always discordant for the two observations in each subject, the effects of time and the interaction of time and group can be examined using the matched-pair method.

The matched odds ratio for time is 2.42 and its 95 percent confidence interval excludes one, suggesting a significant increase in smoking from baseline to one year. The interaction between time and group is tested by using the "discordant pairs" for the control and treatment groups in a single table and calculating the Pearson chi-square (36). The test statistic is highly significant ($\chi^2 = 10.21$, df = 1, $p < 0.01$), suggesting that the effects of time are not the same for treatment and control groups ($OR_{time} = 6.33$ for the control group and $OR_{time} = 1.25$ for the treatment group).

Table 5 compares the results from the conditional logistic models and random-intercept logistic models by analyzing the data from baseline and one year. The effects of time are identical between the two methods. The random-effects logistic model is used to estimate the effects of the treatment, controlling for time. The estimated odds ratio for treatment group is 0.92 (95 percent CI 0.50–2.01), which can be compared with 0.91 (95 percent CI 0.61–1.36) from the GEE model. With respect to the interaction of treatment group and time, both the estimate (absolute value) and standard error are greater in the conditional logistic model than in the random-effect model. But the test statistics (Wald χ^2) are relatively close to each other and to that from the GEE logistic model.

Analyses with all seven timepoints

The standard logistic, GEE, and random-effects models are now used to evaluate group differences across all seven timepoints from the Midwestern Prevention Project. In all models, the dependent variables are the repeated measures of smoking from the 6-month to 5-year follow-ups (from T2 to T7). The independent variables include a linear time effect (from 1 to 6), a quadratic time effect ($time^2$), treatment group, and interactions between group and the time effects. Sex, race, grade at baseline, and baseline smoking are entered as controlling variables.

We estimated two models for each set of analyses, one with only main effects of sex, race, grade, baseline

TABLE 5. Conditional logistic regression and random-effects logistic regression analyzing the effects of treatment and time on smoking (baseline and one-year data)

Models	β	SE*	Wald χ^2	Odds ratio	95% CI*
I. Conditional logistic models†					
Time (0 = baseline, 1 = one year)	0.89	0.23	14.4	2.42	1.53–3.82
Test for interaction treatment group × time	−1.62	0.53	9.30		
II. Random-intercept logistic model†,‡					
Treatment group (0 = control, 1 = treatment)	−0.08	0.31	0.06	0.92	0.50–2.01
Time (0 = baseline, 1 = one year)	0.89	0.24	13.7	2.42	1.52–3.87
Test for interaction	−1.53	0.49	9.55		

* SE, standard error; CI, confidence interval.
† The effects of treatment group and time are estimated from the main effects model.
‡ Random-effect standard deviations were estimated as $\sigma_v = 2.25$ (SE = 0.32) in the main effects model, and $\sigma_v = 2.38$ (SE = 0.35) in the model with interaction.

smoking, time, and treatment group and the other that added time^2, time by group, and time^2 by group.

Table 6 shows the estimates, standard errors, and Wald χ^2 from all models. In all main-effects models, the linear time effect is significantly larger than zero, indicating an overall increasing trend of smoking across time. The effect of treatment group is significantly less than zero, indicating increased smoking in the control group compared with the treatment group after controlling for time and other covariates. In addition, the race effect is significant, indicating that whites are more likely to smoke than nonwhites. Those who smoked at baseline are more likely to smoke at later timepoints.

In the models including interaction terms, the quadratic time effects are significantly less than zero, indicating that in the control group (i.e., group = 0), the rate of increase in smoking has leveled off. The group by linear effects of time are significantly smaller than zero, suggesting that the increasing rate is greater in the control group than in the treatment group. The group by quadratic time effects, however, are significantly larger than zero. Taken together, there is a deceleration in the increasing trend of monthly smoking in the control group (estimate of the quadratic trend is -0.16 from the random-intercepts model) and a slight acceleration in the treatment group (estimate of the quadratic trend is $-0.16 + 0.20 = 0.04$ from the random-intercepts model).

In general, the parameter estimates from the standard logistic model and the GEE model are relatively close, whereas the Wald χ^2 from the random-effects model and the GEE model are relatively close. As expected, the standard errors for time-invariant covariates such as sex, race, and treatment group are smaller in the standard logistic models, while the standard errors for time-varying covariates such as linear trend and the interaction terms are generally smaller in the GEE models. Both the estimates and standard errors from the random-effects model are larger than those from the GEE models, although the test statistics are relatively close.

DISCUSSION

In this paper, we have illustrated how GEE models relate to stratified analyses and standard logistic models and how random-effects models relate to the matched-pair method in epidemiology. Understanding these connections might be helpful in making a choice of methods for correlated outcomes as well as in interpreting estimated effects from different models. The GEE approach models the marginal distributions and treats the longitudinal data as though they were cross-sectional. The interpretation of estimated regression coefficients and odds ratios is in line with the Mantel-Haenszel method and the standard logistic models. The dependence between repeated observations is taken into account by robust variance estimation. When the number of subjects is large and missing data are not an issue (i.e., no missing or MCAR, discussed below), the estimated regression coefficients from the standard logistic model should be very similar to the estimates from the GEE method (37). However, the standard errors from the standard logistic models are biased. The biases are dependent on whether the covariates vary with time. Consistent with the statistical literature (23, 24), our results show that regression models ignoring the time dependency tend to overestimate the standard errors of time-varying covariates and underestimate the standard errors of time-invariant covariates.

TABLE 6. Logistic models estimating intervention effects on the trends of smoking across all follow-up timepoints: the Midwestern Prevention Project, 1984–1992

Covariate	Standard logistic model			GEE† logistic model			Random-intercept logistic model‡		
	Estimate	SE†	Wald χ^2	Estimate	Robust SE	Wald χ^2	Estimate	SE	Wald χ^2
Main effects model									
Sex (0 = female, 1 = male)	0.15	0.08	3.37	0.17	0.13	1.79	0.24	0.19	1.56
Race (0 = nonwhite, 1 = white)	1.13***	0.14	67.6	1.24***	0.24	26.51	1.72***	0.28	36.48
Grade (0 = 6th, 1 = 7th)	−0.01	0.10	0.01	0.03	0.16	0.04	0.01	0.23	0
Baseline smoking (0 = no, 1 = yes)	1.92***	0.13	216.68	1.92***	0.22	76.51	3.00***	0.34	76.56
Time (1–6)	0.37***	0.03	213.92	0.37***	0.02	216.85	0.55***	0.03	419.02
Group (0 = control, 1 = treatment)	−0.47***	0.10	22.96	−0.43***	0.16	7.42	−0.67***	0.23	8.53
Interaction terms added									
Time^2	−0.11***	0.02	20.27	−0.11***	0.02	30.14	−0.16***	0.03	26.32
Group × time	−0.93***	0.26	12.54	−0.94***	0.22	18.26	−1.40***	0.32	19.01
Group × time^2	0.13***	0.03	14.57	0.13***	0.03	21.56	0.20***	0.04	21.07

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

† GEE, generalized estimating equations; SE, standard error.

‡ Random-effect standard deviations were estimated as $\sigma_u = 1.96$ (SE = 0.12) in the main effects model, and $\sigma_u = 1.99$ (SE = 0.13) in the model with interaction terms.

Although the random-effects approach naturally relates to the conditional logistic model with respect to the interpretation of the estimated regression coefficients, these two approaches are distinct in handling the random-intercept. While the random-effects approach using maximum marginal likelihood estimation explicitly models and estimates the random subject effects (i.e., the degree of heterogeneity across subjects in the probability of response, not attributable to covariates in the model), the conditional likelihood approach treats random subject effects as "nuisance" and they are conditioned out of the likelihood function. In addition, the estimated covariate effects in the more general random-effects logistic model are based on both within-subject and between-subject comparisons, while the conditional likelihood approach is based entirely on within-subject comparisons and thus provides no information about covariates which do not vary over time. Under a restricted condition (no missing data and no between-subject components), the conditional likelihood approach is as efficient as the random-effects approach (25). As such, the conditional likelihood approach is appropriate for longitudinal analyses if only within-subject effects are of interest.

For the purposes of illustration, we only analyzed subjects with complete data. In reality, missing data are inevitable in longitudinal studies. In order for the models to provide valid estimates in the presence of missing data, certain assumptions have to be met (38). The most stringent assumption is data missing completely at random (MCAR). Under this assumption, the missingness does not depend on any individual characteristics. In other words, missing subjects can be considered as a simple random sample of all subjects. However, MCAR can also be satisfied if missingness is explained by model covariates, e.g., treatment conditions, age, sex, and race. This special case of MCAR is what Little (39) refers to as covariate-dependent missing. The GEE model discussed in this paper requires this assumption. The assumption for random-effects models is less restrictive, namely, the missing data are assumed to be missing at random (MAR), which conditionally allows missingness to depend on an individual's previously observed values of the dependent variable. This assumption is more realistic in practice.

The absolute values of the estimates from the random-effects models are generally larger than those from GEE models (9). The discrepancies in the estimates between the two approaches are largely dependent on the correlation between the repeated measures. However, the "marginal" coefficients of fitted random-effects models are very similar to those from a GEE model. In fact, a subject-specific estimate can be converted to a marginal or population-averaged coefficient through the following equation (8):

$$\beta_{PA} \simeq \beta_{SS}/\sqrt{(1 + 0.346\sigma_v^2)}, \quad (5)$$

where β_{PA} is a population-averaged coefficient, β_{SS} is a subject-specific estimate, and σ_v^2 is the variance of the random effect. This relation can be shown using the estimates obtained from the random-effects and GEE logistic models listed in table 6.

A relevant question is which approach is more appropriate (i.e., GEE or random-effects). GEE models are desirable when the research focus is on differences in population-averaged response, while random-effects models are appropriate when the research focus is on the change in individuals' responses (8, 25). In our example, if the differences in the increasing trends of smoking between treatment and control groups are of interest, then random-effects models estimating the changes in individuals' smoking behavior across time are more appropriate. The random-effect approach can be also useful in modeling co-twin control studies, where the major interest is the within-pair comparisons across exposure discordant pairs (40). In this situation, the random-intercept reflects the dependency within each twin pair over and above the influence of other model terms. The heterogeneity in these pair-varying intercepts may be due to unmeasured shared genetic and environmental factors common to the twins within a pair. On the other hand, the GEE approach is more desirable when the objective is to make inference about group differences. In the smoking prevention example, if we want to estimate the averaged treatment effect, regardless of individual change over time, then the population-averaged parameter is of more interest. An advantage of the GEE approach is that it can provide robust variance estimation, whereas the random-effects may be sensitive to different assumptions about the variance and covariance correlation structure, which are usually difficult to validate (8).

In epidemiology, longitudinal studies have been used in many situations, such as clinical and prevention trials, prospective studies of exposure-disease relations, and repeated health services utilization surveys. In these situations, different types of measures over time may emerge, such as repeated measures of continuous variables (e.g., weight in pounds (kg), blood pressure in mmHg), repeated measures of binary variables (e.g., obesity, hospitalization), repeated measures of ordinal variables (e.g., frequency or severity of a symptom), and other measures that are irreversible over time (e.g., death). Software for the analysis of repeated measures of continuous variables is now

widely available, such as Proc Mixed in SAS and BMDP5V. Survival analyses are widely used to analyze the timing of "absorbing" events such as death or onset of a disease in cohort studies.

The methods discussed in this paper are suitable for the analyses of repeated measures of binary variables in epidemiology. In many rich data bases of longitudinal studies in epidemiology, such as the Framingham Study (41) or the Nurses' Health Study (42), repeated measures of discrete outcomes are common. Traditional analyses of these longitudinal studies have often been restricted to data obtained from baseline and one other timepoint. The statistical models discussed in this paper, while more complex than the traditional approaches, use all available data and can produce more efficient estimates (2, 43).

ACKNOWLEDGMENTS

Data collection for this project was supported by National Institute of Drug Abuse (NIDA) grant no. 03976 and data analysis was supported by NIDA grant no. 06307. This research was conducted when Dr. Hu was at Prevention Research Center, University of Illinois at Chicago.

The authors thank Drs. Dick Campbell and Paul Levy for many helpful suggestions. They also thank two anonymous referees for helpful comments.

REFERENCES

1. Bock RD. Multivariate statistical methods in behavioral research. New York: McGraw Hill, 1975.
2. Zeger SL, Liang K-Y. An overview of methods for the analysis of longitudinal data. Stat Med 1992;11:1825–39.
3. Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics 1982;38:963–74.
4. SAS. "The mixed procedure." In: SAS/STAT software: changes and enhancements, release 6.07. (SAS technical report no. P-229). Cary, NC: SAS Institute, Inc, 1992:chapter 16.
5. Dixon WJ. BMDP statistical software. Berkeley, CA: University of California Press, 1986.
6. Bryk AS, Raudenbush SW, Congdon RT. Hierarchical linear modeling with the HLM/2L and HLM/3L programs. Chicago, IL: Scientific Software, Inc, 1994.
7. Woodhouse G. A guide to MLn for new users. London: Multilevel Models Project, Institute of Education, University of London, 1995.
8. Zeger SL, Liang K-Y, Albert PS. Methods for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44:1049–60.
9. 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.
10. Stiratelli R, Laird NM, Ware JH. Random-effects models for serial observations with binary response. Biometrics 1984;40: 961–71.
11. Wong GY, Mason WM. The hierarchical logistic regression model for multilevel analysis. J Am Stat Assoc 1985;80: 513–24.
12. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
13. Galbraith JI. The interpretation a regression coefficient. Biometrics 1991;47:1593–6.
14. Zeger SL, Liang K-Y, Albert PS. Response: The interpretation of a regression coefficient. Biometrics 1991;47:1596–7.
15. Park T. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. Stat Med 1993;12:1723–32.
16. Ten Have TR, Landis JR, Weaver SL. Association models for periodontal disease progression: a comparison of methods for clustered binary data. Stat Med 1995;14:423–9.
17. Newhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. Stat Methods Med Res 1992;1:249–73.
18. Pendergast JF, Gange SJ, Newton MA, et al. A survey of methods for analyzing clustered binary response data. Int Stat Rev 1996;64:89–118.
19. Pentz MA, Dwyer JH, MacKinnon DP, et al. A multicommunity trial for primary prevention of adolescent drug use. JAMA 1989;261:3259–66.
20. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. JNCI 1959;22: 719–48.
21. Hosmer DW Jr, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc, 1989.
22. McCullagh P, Nelder JA. Generalized linear models. 2nd ed. New York: Chapman and Hall, 1989.
23. Dunlop DD. Regression for longitudinal data: a bridge from least squares regression. Am Stat 1994;48:299–303.
24. Fitzmaurice GM, Laird NM, Rotnitzky AG. Regression models for discrete longitudinal responses. Stat Sci 1993;8: 284–309.
25. Diggle P, Liang K-Y, Zeger SL. Analysis of longitudinal data. New York: Oxford University Press, 1994.
26. Gibbons RD, Bock RD. Trend in correlated proportions. Psychometrika 1987;52:113–24.
27. Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika 1991;78:45–51.
28. Longford NT. Random coefficient models. New York: Oxford University Press, 1993.
29. Laird NM. Missing data in longitudinal studies. Stat Med 1988;7:305–15.
30. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. Biometrics 1994;50:933–44.
31. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–7.
32. Neuhaus JM, Jewell NP. The effect of retrospective sampling on binary regression models for clustered data. Biometrics 1990;46:977–90.
33. Pentz MA, MacKinnon, Flay BR, et al. Primary prevention of chronic diseases in adolescence: effects of the Midwestern Prevention Project on tobacco use. Am J Epidemiol 1989;130: 713–24.
34. Siddiqui O, Hedeker D, Flay BR, et al. Intraclass correlation estimates in a school-based smoking prevention study: outcome and mediating variables, by sex and ethnicity. Am J Epidemiol 1996;144:425–33.
35. Murray DM, Rooney BL, Hannan PJ, et al. Intraclass correlation among common measures of adolescent smoking: estimates, correlates, and applications in smoking prevention studies. Am J Epidemiol 1994;140:1038–50.
36. Breslow NE, Day NE. Statistical methods in cancer research. Vol 1. The analysis of case-control studies. (IARC Scientific Publications no. 32). Lyon: International Agency for Research on Cancer, 1980.
37. Zeger SL. Commentary. Stat Med 1988;7:161–8.
38. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley, 1987.
39. Little RJA. Modeling the drop-out mechanism in repeated-measures studies. J Am Stat Assoc 1995;90:1112–21.
40. Hu FB, Goldberg J, Hedeker D, et al. Modeling ordinal

responses from co-twin control studies. Stat Med. In press.
41. Dawber TR. The Framingham Study. Cambridge, MA: Harvard University Press, 1980.
42. Willett WC, Green A, Stampfer MJ, et al. Relative and absolute excess risks of coronary heart disease among women who smoke cigarettes. N Engl J Med 1987;317:1303–9.
43. Dwyer J, Feinleib M. Introduction to statistical models for longitudinal observation. In: Dwyer JH, Feinleib M, Lippert P, et al, eds. Statistical models for longitudinal studies of health. New York: Oxford University Press, 1992:3–48.
44. Karim MR, Zeger SL. GEE: a SAS macro for longitudinal data analysis. (Technical report no. 674). Baltimore, MD: Department of Biostatistics, John Hopkins University, 1988.
45. Davis SD. A computer program for regression of repeated measures using generalized estimating equations. Comput Methods Programs Biomed 1993;40:15–31.
46. SAS Institute Inc. SAS/STAT software: the GENMOD procedure, release 6.09. (SAS technical report no. P-243). Cary, NC: SAS Institute Inc, 1993.
47. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. Comput Methods Programs Biomed 1996;49:157–76.
48. Statistics and Epidemiologic Research Corp. EGRET reference manual. Seattle, WA: SERC, 1992.
49. SAS Institute Inc. SAS/STAT software: the PHREG procedure. Cary, NC: SAS Institute Inc, 1993.

APPENDIX

The GEE logistic models in this paper were fit using a SAS macro written by Karim and Zeger (44). The solution to the GEE involves an iterative solution that alternates between quasi-likelihood methods (using iteratively reweighed least-squares) for estimating parameter estimates and a robust method for estimating correlation as a function of the parameter estimates. Davis (45) has also published a FORTRAN 77 program to fit the GEE models. Besides fitting logistic models, both the Karim and Zeger program and the Davis program can be used to fit linear models for continuous outcomes and Poisson models for count data. The specification of the link and variance functions is the same as in fitting the generalized linear models (GLM) by GENMOD procedure in SAS 6.08 (46). The GENMOD procedure in SAS 6.12 can now fit GEE models.

The random-intercept logistic model in this paper was fitted by a FORTRAN program called MIXOR written by Hedeker and Gibbons (47). MIXOR can allow for multiple random-effects (e.g., random intercept and random slope) and includes both logistic and probit response functions. Because the likelihood does not exist in a closed form for the general random-effect model, Gauss-Hermite quadrature method is used to perform the integration numerically by summing over a specified number of quadrature points. Usually, the more points used, the more accurate the approximation, but the more time it takes. In the present paper, we used 20 quadrature points for model estimation. In addition, MIXOR can be used to fit random-effects models for ordinal response data (30).

The random-intercepts model described in this paper can also be fit by EGRET (48). The random-intercept model is referred to as the logistic-normal model by EGRET. As with MIXOR, Gauss-Hermite quadrature is used to integrate over the random-effects distribution.

Besides the logistic-normal model, EGRET can also fit beta-binomial logistic models and logistic-binomial models. When using the same number of quadrature points (20 points), MIXOR and EGRET produce almost identical results for the data described above. The conditional logistic models in this paper were estimated using the SAS PHREG procedure (49). EGRET can also fit conditional logistic models.