Published in final edited form as:
Biometrics. 2010 June ; 66(2): 365–373. doi:10.1111/j.1541-0420.2009.01306.x.

Longitudinal studies of binary response data following case-control and stratified case-control sampling: design and analysis

Jonathan S. Schildcrout and
Departments of Biostatistics and Anesthesiology, Vanderbilt University School of Medicine, 1161 21st Avenue South, S-2323 Medical Center North, Nashville, TN 37232, USA, jonathan.schildcrout@vanderbilt.edu

Paul J. Rathouz
Department of Health Studies, University of Chicago, 5841 South Maryland Ave, MC 2007, Chicago, Illinois 60637, USA, prathouz@uchicago.edu

SUMMARY

We discuss design and analysis of longitudinal studies after case-control sampling, wherein interest is in the relationship between a longitudinal binary response that is related to the sampling (case-control) variable, and a set of covariates. We propose a semiparametric modelling framework based on a marginal longitudinal binary response model and an ancillary model for subjects' case-control status. In this approach, the analyst must posit the population prevalence of being a case, which is then used to compute an offset term in the ancillary model. Parameter estimates from this model are used to compute offsets for the longitudinal response model. Examining the impact of population prevalence and ancillary model misspecification, we show that time-invariant covariate parameter estimates, other than the intercept, are reasonably robust, but intercept and time-varying covariate parameter estimates can be sensitive to such misspecification. We study design and analysis issues impacting study efficiency, namely: choice of sampling variable and the strength of its relationship to the response, sample stratification, choice of working covariance weighting, and degree of flexibility of the ancillary model. The research is motivated by a longitudinal study following case-control sampling of the time course of ADHD symptoms.

Keywords

Bias; binary data; efficiency; Generalized Estimating Equations; longitudinal data; logistic regression; outcome dependent sampling

1. Introduction

The Attention Deficit Hyperactivity Disorder (ADHD) Study (Lahey, 1998; Hartung et al., 2002) is a longitudinal study on 255 children that seeks to identify risk and prognostic factors in early childhood for ADHD symptoms, diagnoses, and functional outcomes across

Supplementary Materials
Web Appendices A and B referenced in Sections 3 and 4.2, respectively, are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

childhood, adolescence and early adulthood. In the paper we model ADHD prevalence as a function of time and baseline predictors in the first eight waves of data (including baseline). One hundred thirty-eight children who were referred to one of two participating clinics due to parent or teacher suspicion of ADHD symptom exhibition were enrolled in the study, as was a demographically and socioeconomically similar group of 117 non-referred children. All participants were followed over seven annual visits after baseline. Assessment of ADHD symptoms was made at each visit using the Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM-IV; American Psychatric Association, 1994) criteria, and these assessments were used to generate at each wave a diagnosis of ADHD in the previous six months. While participant referral was a strong predictor of ADHD symptom level, particularly at the first (baseline) visit, the relationship was not deterministic, and some referred subjects did not meet criteria for ADHD at baseline. Conversely, non-referred participants exhibited symptoms and at times met diagnostic criteria for ADHD.

Because referred and non-referred participants are at high and low risk, respectively, for expressing symptoms during followup, the ADHD study design allows researchers to observe substantial response variation and thereby to potentially estimate many target regression effects efficiently. Because the sampling scheme is biased, however, standard longitudinal data analysis methods do not apply. In this manuscript, we discuss analytical strategies and design considerations when such “case-control” (e.g., referred and non-referred) sampling is followed by longitudinal followup on a binary response related to case-control status at baseline. Similar biased sampling in longitudinal studies has been used elsewhere (e.g., Lahey et al., 1999), and the methods described herein would apply in those settings as well.

To formalize the problem, assume interest lies in the longitudinal marginal relationship $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ where i indexes subjects in a population, $\boldsymbol{Y}_i$ is a binary vector of responses on the ith subject, and $\boldsymbol{X}_i$ is a design matrix containing predictor and adjustment variables of interest. Subjects are sampled from the population into the study with probability that depends on a univariate case-control or sampling variable Z_i which is related to $\boldsymbol{Y}_i$, or possibly on $(Z_i, \boldsymbol{X}_{1i})$, where $\boldsymbol{X}_{1i}$ is contained in $\boldsymbol{X}_i$. The analytic goal is to make inferences on the marginal mean $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$. Though case-control sampling is in the general sense a stratified design, for the purpose of this paper, we refer to sampling based on $(Z_i, \boldsymbol{X}_{1i})$ as *stratified sampling* (dropping the case-control designation for ease of exposition only), and we refer to sampling based only on Z_i as *case-control sampling*.

The case-control design we consider is a specific instance of what is more generally termed outcome dependent sampling (ODS). The majority of such ODS designs, including those pertaining to longitudinal and correlated data, require explicit acknowledgment of non-equal probability of participant ascertainment in the analysis. Neuhaus and Jewell (1990) and Qaqish et al. (1997) discuss the implications of ignoring the ODS design with cluster-based sampling for correlated data, and Neuhaus and Jewell propose subject-specific conditional logistic regression models when sampling is based upon binary response vector sums. Similarly, Schildcrout and Heagerty (2008) describe sampling based on the presence/absence of binary response series variation and propose conditional maximum likelihood analyses for marginal models. Case-control family studies, an alternative design, sample on a single component (the proband) of a cluster rather than on a summary of the entire cluster-level response vector. Whittemore (1995), Zhao et al (1998), and Neuhaus, Scott, and Wild (2002) approach case-control family studies via marginal models; Neuhaus, Scott and Wild (2006) have more recently developed methods using subject-specific models. Our design is linked to the case-control family study; however, we sample on an ancillary variate that is related but not equal to the index response. Whereas Neuhaus et al. (2006) discuss a ‘stochastic’ sampling design like ours, they propose likelihood-based estimation. In contrast, we develop a semiparametric estimation strategy for the marginal model E($\boldsymbol{Y}_i \mid \boldsymbol{X}_i$) using generalized estimating equations (GEE; Liang and Zeger, 1986). Advantages to likelihood-based estimation over GEE are well known (e.g., model selection and missing data); however, we believe it is important to develop an estimation strategy for this design using methods that, unlike parametric approaches, are insensitive to dependence model misspecification. Our approach can be implemented using standard GEE software and we provide a macro for doing so in Stata (StatCorp, 2007) on the second author's (PJR) website (http://health.bsd.uchicago.edu/rathouz/Software).

This manuscript is organized as follows. In section 2, we describe modeling assumptions that must be made for valid inferences and a general strategy that can be used for estimation with this study design. We detail a semiparametric estimation approach to parameter estimation and inference under logistic regression models for Z_i and $\boldsymbol{Y}_i$ in section 3. Section 4 reports on simulation studies conducted for the purpose of examining the finite sample operating characteristics of the proposed estimator, focusing on the impact on bias of model misspecification and on statistical efficiency of design and estimation strategies. We return to the ADHD study in section 5 and describe an analysis of those data. Finally, we provide concluding remarks and a discussion in section 6.

2. Sampling and modeling assumptions

Consider a target population wherein each subject i in the population admits ($\boldsymbol{Y}_i$, $\boldsymbol{t}_i$, $\boldsymbol{X}_i$). Here, $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$ is a longitudinal series of binary outcomes such as annual ADHD diagnosis, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{in_i})'$ is a $n_i \times p$ matrix of covariates predicting $\boldsymbol{Y}_i$, and $\boldsymbol{t}_i = (t_{i1}, \dots, t_{in_i})'$ is a vector of observation times which may also be contained in $\boldsymbol{X}_i$. For example, in the ADHD study, each row j of $\boldsymbol{X}_i$ may contain a vector of baseline (e.g., gender and ethnicity) and time-varying (e.g., wave, age, other psychiatric diagnoses, or interactions between baseline predictors and time) predictors for ADHD diagnosis Y_{ij} at time t_{ij}. For purposes of exposition, we assume that the number n_i and values $\boldsymbol{t}_i$ of observation times are fixed by design. In practice, n_i and $\boldsymbol{t}_i$ can vary either functionally or stochastically depending on baseline predictors contained in $\boldsymbol{X}_i$, so long as they are independent of $\boldsymbol{Y}_i$ given such baseline predictors.

We assume that interest lies in the marginal probability that $Y_{ij} = 1$ given $\boldsymbol{X}_i$ in the target population,

$$\mu_{P_{ij}} = \Pr(Y_{ij} = 1 \mid \boldsymbol{X}_i) = g^{-1}(\beta_0 + \boldsymbol{x}_{ij}'\beta_1) \tag{1}$$

(subscript P for target population), where $g(\cdot)$ is a link function mapping (0, 1) to the real line, and, generally, $\boldsymbol{\beta}_1$ is the parameter of interest. Note that (1) implicitly contains the "reproducibility" or "no interference" assumption that

$$\Pr(Y_{ij} = 1 \mid \boldsymbol{X}_i) = \Pr(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}),$$

i.e., that predictors available in $\boldsymbol{X}_i$ provide no additional predictive value for Y_{ij} over and above the information available in $\boldsymbol{x}_{ij}$. In the ADHD study, for example, this assumption would be easily satisfied if $\boldsymbol{x}_{ij}$ contains only baseline and non-stochastic predictors such as time or age. In situations wherein $\boldsymbol{x}_{ij}$ contains stochastic predictors such as other mental health diagnosees as time t_{ij}, the assumptions needs to be more carefully examined.

In randomly drawing a sample from the target population, let S_i be an indicator variable for the ith subject in the population being selected into the sample, and assume that $S_i \amalg S_{i'}$, $i \neq i'$. Under simple random sampling, or sampling that is related to $\boldsymbol{X}_i$ but not to $\boldsymbol{Y}_i$, models and inferences for $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)'$ in μ_{Pij} can be carried out ignoring the sampling process. A typical approach would be to fit mean model (1) along with a working correlation model, corr$(Y_{ij}, Y_{ik}|\boldsymbol{X}_i)$, $j \in \{1, \ldots, n_i\}$, $k \in \{1, \ldots, n_i\}$, $j \neq k$, using GEE.

Now, consider a more challenging design wherein sampling S_i is related in some way to $\boldsymbol{Y}_i$ and possibly $\boldsymbol{X}_i$. In this setting, the sample is no longer representative of the target population. Rather, it represents a pseudo-population that is a reweighted version of the target population, where weights vary as a function of $\boldsymbol{Y}_i$, or if sampling also depends upon $\boldsymbol{X}_i$, as a function of $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$. To proceed, define the sampling probability $\rho_{ij}(y, \boldsymbol{X}_i) \equiv \Pr(S_i = 1|Y_{ij} = y, \boldsymbol{X}_i)$. That is, $\rho_{ij}(y, \boldsymbol{X}_i)$ is defined to equal the probability of being sampled conditional on the entire design matrix $\boldsymbol{X}_i$, but only on the j^{th} response Y_{ij}; even though S_i may depend on the entire vector $\boldsymbol{Y}_i$, for reasons that will become evident, we focus here only on this marginal probability. We have included subscripts ij on $\rho_{ij}(y, \boldsymbol{X}_i)$ to indicate that this probability could vary with one or more of the observation number j, time t_{ij}, and design matrix $\boldsymbol{X}_i$. Then, conditional on being sampled (i.e., $S_i = 1$), standard Bayes' Theorem calculations applied to target population model (1) yield the following pseudo-population marginal odds model,

$$\frac{\Pr(Y_{ij}=1|\boldsymbol{X}_i, S_i=1)}{\Pr(Y_{ij}=0|\boldsymbol{X}_i, S_i=1)}=\frac{\mu_{Sij}}{1-\mu_{Sij}}=\frac{\mu_{Pij}}{1-\mu_{Pij}}\frac{\rho_{ij}(1, \boldsymbol{X}_i)}{\rho_{ij}(0, \boldsymbol{X}_i)} \tag{2}$$

(subscript S indicating pseudo-population $\underline{s}$ample).

When $\rho_{ij}(1, \boldsymbol{X}_i)/\rho_{ij}(0, \boldsymbol{X}_i)$ is known, we may use (2) to make inferences about target population parameters through parameter estimation for the pseudo-population model. Estimation with GEE would require specification of the mean model for μ_{Sij} given by (2) and a working correlation model for corr$(Y_{ij}, Y_{ik}|\boldsymbol{X}_i, S_i = 1)$ in the pseudo-population.

Here, we consider the circumstance where the sampling fraction, $\rho_{ij}(1, \boldsymbol{X}_i)/\rho_{ij}(0, \boldsymbol{X}_i)$ is unknown. We assume that sampling depends upon $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ only indirectly through a binary case-control or sampling variable Z_i, or, when a stratified sampling scheme is implemented, only indirectly through $(Z_i, \boldsymbol{X}_{1i})$, where $\boldsymbol{X}_i = (\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i})$, and $\boldsymbol{X}_{1i}$ contains a subset of the information, generally available at baseline, in $\boldsymbol{X}_i$. In the ADHD study, Z_i indicates referral status, and $\boldsymbol{X}_{1i}$ contains subject's gender. In other designs it may also include measures such as baseline age, neighborhood or community variables available at the time of enrollment, etc. Stratifed sampling may be utilized in order to improve estimation efficiency on the coefficients for $\boldsymbol{X}_{1i}$ as well as covariates in $\boldsymbol{X}_{2i}$ that are related to $\boldsymbol{X}_{1i}$. Formally, we assume that, without stratification, $S_i \amalg (\boldsymbol{Y}_i, \boldsymbol{X}_i)|Z_i$, while for the stratified design,

$$S_i \coprod (\boldsymbol{Y}_i, \boldsymbol{X}_i)|(Z_i, \boldsymbol{X}_{1i}). \tag{3}$$

In the ADHD study, (3) indicates that, given referral status and gender, selection is independent of baseline or subsequent ADHD diagnoses and of other predictor variables. In subsequent exposition, we focus on the stratified sampling design. Our development is easily adapted to the simpler, unstratified design if that is of interest.

Let $\pi(z, \boldsymbol{X}_{1i}) = \Pr(S_i = 1 \mid Z_i = z, \boldsymbol{X}_{1i})$, $z = 0, 1$. Then, if (3) holds, knowledge of $\pi(1, \boldsymbol{X}_{1i})/\pi(0, \boldsymbol{X}_{1i})$ permits estimation of the ratio $\rho_{ij}(1, \boldsymbol{X}_i)/\rho_{ij}(0, \boldsymbol{X}_i)$. This in turn permits inferences about parameters **β** in model (1) via relationship (2). To see this, define $\lambda_{Pij}(y, \boldsymbol{X}_i) = \Pr(Z_i = 1|Y_{ij} = y, \boldsymbol{X}_i)$, $y = 0, 1$. Note that this quantity may at first appear counterintuitive, since Z_i occurs prior to Y_{ij} in time. Nevertheless, this "reverse" conditional probability certainly exists and can be modeled. This model is ancillary to model of interest (1) and is specified in order to render identifiable parameters in model (1). Utilization of this intermediary model to identify parameters in the target model follows directly from Lee, McMurchy, and Scott (1997) and Neuhaus et al. (2006). Owing to the reverse time sequence and to the fact that the conditioning statistic Y_{ij} varies with j, we will tend to choose flexible specifications for $\lambda_{Pij}(y, \boldsymbol{X}_i)$. Note also that, as with $\rho_{ij}(y, \boldsymbol{X}_i)$, $\lambda_{Pij}(y, \boldsymbol{X}_i)$ is conditional on the entire design matrix $\boldsymbol{X}_i$, but only on the j^{th} response Y_{ij}. Similarly to (2), Bayes' Theorem calculations yield an odds model for Z_i in the pseudo-population, viz,

$$\frac{\Pr(Z_i=1|Y_{ij}=y, \boldsymbol{X}_i, S_i=1)}{\Pr(Z_i=0|Y_{ij}=y, \boldsymbol{X}_i, S_i=1)}=\frac{\lambda_{Sij}(y, \boldsymbol{X}_i)}{1-\lambda_{Sij}(y, \boldsymbol{X}_i)}=\frac{\lambda_{Pij}(y, \boldsymbol{X}_i)}{1-\lambda_{Pij}(y, \boldsymbol{X}_i)}\frac{\pi(1, \boldsymbol{X}_{1i})}{\pi(0, \boldsymbol{X}_{1i})}, \tag{4}$$

$y = 0, 1$, where $\lambda_{Sij}(y, \boldsymbol{X}_i) = \Pr(Z_i = 1|Y_{ij} = y, \boldsymbol{X}_i, S_i = 1)$. Additionally, due to (3),

$$\rho_{ij}(y, \boldsymbol{X}_i)=\pi(0, \boldsymbol{X}_{1i})\{1-\lambda_{Pij}(y, \boldsymbol{X}_i)\}+\pi(1, \boldsymbol{X}_{1i})\lambda_{Pij}(y, \boldsymbol{X}_i)\ ,\ y=0, 1,$$

from which we can write the ratio

$$\frac{\rho_{ij}(1, \boldsymbol{X}_i)}{\rho_{ij}(0, \boldsymbol{X}_i)}=\frac{1-\lambda_{Pij}(1, \boldsymbol{X}_i)+\{\pi(1, \boldsymbol{X}_{1i})/\pi(0, \boldsymbol{X}_{1i})\}\lambda_{Pij}(1, \boldsymbol{X}_i)}{1-\lambda_{Pij}(0, \boldsymbol{X}_i)+\{\pi(1, \boldsymbol{X}_{1i})/\pi(0, \boldsymbol{X}_{1i})\}\lambda_{Pij}(0, \boldsymbol{X}_i)}. \tag{5}$$

In Section 3, relationship (4) and sampling ratio $\pi(1, \boldsymbol{X}_{1i})/\pi(0, \boldsymbol{X}_{1i})$ will be used to specify and fit a model for $\lambda_{Pij}(y, \boldsymbol{X}_i)$ using data from the pseudo-population. This will lead to estimates of $\rho_{ij}(1, \boldsymbol{X}_i)/\rho_{ij}(0, \boldsymbol{X}_i)$ using (5) which can be used in (2) to make **β**-inferences.

3. Implementation with logistic regression and GEE

Here, we present a specific approach to the program outlined in section 2, beginning with model specification and estimation for $\lambda_{Pij}(y, \boldsymbol{X}_i)$. Suppose $\lambda_{Pij}(y, \boldsymbol{X}_i)$ is modeled as a logistic regression of Z_i of the form

$$\lambda_{Pij}(y, \boldsymbol{X}_i)=\text{logit}^{-1}(\boldsymbol{w}_{1,ij}^{\prime}\gamma_1+y\times\boldsymbol{w}_{2,ij}^{\prime}\gamma_2), \tag{6}$$

where $\boldsymbol{w}_{1,ij}$ and $\boldsymbol{w}_{2,ij}$ are functions of $(\boldsymbol{X}_i, \boldsymbol{t}_i)$. We assume that $\boldsymbol{w}_{1,ij}$ and $\boldsymbol{w}_{2,ij}$ are sufficiently rich so that

$$\Pr(Z_i=1|Y_{ij}=y, \boldsymbol{w}_{1,ij}, \boldsymbol{w}_{2,ij})=\Pr(Z_i=1|Y_{ij}=y, \boldsymbol{X}_i). \tag{7}$$

We separately denote $\boldsymbol{x}_{ij}$, $\boldsymbol{w}_{1,ij}$ and $\boldsymbol{w}_{2,ij}$ because, even though they may be overlapping in their information content, they may take on different functional forms and because, in order to compute (5), any interactions with y in (6) need to be made explicit. In most applications, we would expect to maximize model flexibility in ancillary model (6), but to be more parsimonious in our specification of model of interest (1). For example, while $\boldsymbol{t}_i$, may be included as a linear term in $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$, the relationship between Z_i and both $\boldsymbol{t}_i$ and the interaction between $\boldsymbol{t}_i$ and y may be non-linear. Later, in the ADHD study example, we allow $\boldsymbol{t}_i$ to be a series of time-specific indicator variables in $\boldsymbol{w}_{1,ij}$, a piecewise linear spline function in $\boldsymbol{w}_{2,ij}$, and a simple linear term in $\boldsymbol{x}_{ij}$.

Via (4), (6) induces a model for λ_{Sij}, i.e.,

$$\lambda_{Sij}(y, \boldsymbol{X}_i)=\text{logit}^{-1}\left(\boldsymbol{w}'_{1,ij}\gamma_1+y \times \boldsymbol{w}'_{2,ij}\gamma_2+\log\{\pi(1, \boldsymbol{X}_{1i})/\pi(0, \boldsymbol{X}_{1i})\}\right). \tag{8}$$

Model (8) can be fitted to the data from a case-control sample using standard logistic regression software, including $\log\{\pi(1, \boldsymbol{X}_{1i})/\pi(0, \boldsymbol{X}_{1i})\}$ as an offset term in the linear predictor. Specifically, setting $\lambda_{Sij} = \lambda_{Sij}(Y_{ij}, \boldsymbol{X}_i)$ and $\gamma=(\gamma'_1, \gamma'_2)'$, γ is estimated by solving the logistic regression score $\sum_i \boldsymbol{T}_i(\boldsymbol{\gamma}) = \boldsymbol{0}$, yielding $\hat{\boldsymbol{\gamma}}$, where

$$\boldsymbol{T}_i(\gamma)=\sum_{j=1}^{n_i}\begin{pmatrix} \boldsymbol{w}_{1,ij} \\ Y_{ij} \times \boldsymbol{w}_{2,ij} \end{pmatrix}(Z_i-\lambda_{Sij}). \tag{9}$$

$\boldsymbol{T}_i(\boldsymbol{\gamma})$ nominally treats the j^{th} term in (9) as independent of the other $n_i - 1$ terms in the sum. This independence does not, of course, hold, as all terms share the same response variable Z_i. Nevertheless, $\boldsymbol{T}_i(\boldsymbol{\gamma})$ is unbiased and so in general will yield consistent estimators for $\boldsymbol{\gamma}$. Therefore, $\boldsymbol{\gamma}$ can be estimated using any logistic regression GEE software program which permits offset terms and allows for the independence correlation structure.

Turning to model specification for μ_{Pij}, let $g(\cdot)$ be the logit function. Then (1) implies $\mu_{Pij}=\text{logit}^{-1}(\beta_0+\boldsymbol{x}'_{ij}\beta_1)$, and, from (2),

$$\mu_{Sij}=\text{logit}^{-1}\left(\beta_0+\boldsymbol{x}'_{ij}\beta_1+B_{ij}\right), \tag{10}$$

wherein the bias-correction term $B_{ij} = B_{ij}(\boldsymbol{X}_i) = \log\{\rho_{ij}(1, \boldsymbol{X}_i)/\rho_{ij}(0, \boldsymbol{X}_i)\}$ appears as an offset. By (5) and (6), B_{ij} is a function of $\boldsymbol{\gamma}$, and so is estimable by plugging in $\hat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$ to obtain $\hat{B}_{ij}$. With $\hat{B}_{ij}$, the sampled data can then be analyzed using marginal model (10). This mean model can be complemented with a working correlation model

$$c_{Sijk}(\alpha)=\text{corr}(Y_{ij}, Y_{ik}|\boldsymbol{X}_i, S_i=1;\alpha) \tag{11}$$

in the sampled pseudo-population, governed by parameter $\boldsymbol{\alpha}$, though $\hat{\boldsymbol{\alpha}}$ cannot be applied to inferences regarding the target population. The model specified via (10) and (11) with $\hat{B}_{ij}$ replacing B_{ij} can then be estimated directly using any standard GEE software program. If the working correlation model is a reasonable approximation to the true correlation structure in the sampled pseudo-population, it should result in an increase in statistical efficiency for $\boldsymbol{\beta}$ estimation under (10) relative to, say, estimation under the independence working correlation model (Liang et al, 1992;Fitzmaurice, 1995;Mancl and Leroux, 1996;Schildcrout and Heagerty, 2005). Standard errors for $\hat{\boldsymbol{\beta}}$ will not however be correct under this approach, since they must account for the uncertainty in estimation of $\boldsymbol{\gamma}$.

Standard errors can be calculated via a corrected version of the sandwich estimator (Liang and Zeger, 1986). Note that $\boldsymbol{\beta}$ is estimated by solving the GEE logistic regression estimating equation $\sum_i \boldsymbol{U}_i(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}) = 0$ for β, yielding $\hat{\boldsymbol{\beta}}$, where

$$\boldsymbol{U}_i(\beta,\gamma)=\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i-\mu_{Si}).$$

Here, as in the usual GEE setup, $\mu_{Si}=(\mu_{Si1},\ldots,\mu_{Sin_i})'$, $\boldsymbol{D}_i'=(1_{n_i},\boldsymbol{X}_i)'\boldsymbol{A}_i$, $\boldsymbol{A}_i=\mathrm{diag}\{\mu_{Sij}(1-\mu_{Sij})\}_{j=1}^{n_i}$, $\boldsymbol{V}_i=\boldsymbol{A}_i^{1/2}\boldsymbol{C}_i\boldsymbol{A}_i^{1/2}$, and $\boldsymbol{C}_i$ is the $n_i \times n_i$ matrix with element (j, k) given by (11). Correlation parameter $\boldsymbol{\alpha}$ is estimated iteratively with $\boldsymbol{\beta}$, but owing to the orthogonality of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in $\boldsymbol{U}_i$, estimation of $\boldsymbol{\alpha}$ has no asymptotic impact on the validity of the standard errors of $\hat{\boldsymbol{\beta}}$ (Liang and Zeger, 1986). Robust standard errors for $\hat{\boldsymbol{\beta}}$ are developed by viewing $(\hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\beta}}')'$ as the solution to the “stacked” estimating equation

$$\sum_i \begin{pmatrix} \boldsymbol{T}_i(\gamma) \\ \boldsymbol{U}_i(\beta,\gamma) \end{pmatrix}=\boldsymbol{0}. \tag{12}$$

The asymptotic variance of $(\hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\beta}}')'$ is then given as

$$\mathrm{AVar}(\widehat{\gamma}',\widehat{\beta}')'=\widehat{\boldsymbol{I}}^{-1}\widehat{\boldsymbol{Q}}\,\widehat{\boldsymbol{I}}^{-1\prime}, \tag{13}$$

where the ^’s indicate that $(\boldsymbol{\gamma}', \boldsymbol{\beta}')'$ has been replaced by $(\hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{\beta}}')'$,

$$\boldsymbol{Q}=\sum_i \begin{pmatrix} \boldsymbol{T}_i(\gamma) \\ \boldsymbol{U}_i(\beta,\gamma) \end{pmatrix}^{\otimes 2}, \quad \text{and } \boldsymbol{I}=\begin{pmatrix} \boldsymbol{I}_{TT} & \boldsymbol{0} \\ \boldsymbol{I}_{UT} & \boldsymbol{I}_{UU} \end{pmatrix}. \tag{14}$$

In (14),

$$\boldsymbol{I}_{TT}=\sum_i\sum_{j=1}^{n_i}\mathrm{E}\left(-\frac{\partial \boldsymbol{T}_i}{\partial \gamma'}\right)=\sum_i \begin{pmatrix} \boldsymbol{w}_{1,ij} \\ Y_{ij}\times \boldsymbol{w}_{2,ij} \end{pmatrix}^{\otimes 2}\{\lambda_{Sij}(1-\lambda_{Sij})\}^{-1},$$

the upper right quadrant of $\boldsymbol{I}$ is 0 because $\mathrm{E}(-\partial \boldsymbol{T}_i/\partial \boldsymbol{\beta}') = 0$,

$$\boldsymbol{I}_{UT}=\sum_i \mathrm{E}\left(-\frac{\partial \boldsymbol{U}_i}{\partial \gamma'}\right)=\sum_i \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{A}_i\left(\frac{\partial \boldsymbol{B}_i}{\partial \gamma'}\right),$$

and

$$\boldsymbol{I}_{UU}=\sum_i \mathrm{E}\left(-\frac{\partial \boldsymbol{U}_i}{\partial \beta'}\right)=\sum_i \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i.$$

In $\boldsymbol{I}_{UT}$, $\boldsymbol{B}_i = (B_{i1},\ldots, B_{in_i})'$ and

$$\left(\frac{\partial \boldsymbol{B}_i}{\partial \gamma'}\right)'=-\begin{pmatrix} \boldsymbol{W}_{1i}' \\ \boldsymbol{W}_{2i}' \end{pmatrix}F_i(1)+\begin{pmatrix} \boldsymbol{W}_{1i}' \\ 0 \end{pmatrix}F_i(0), \tag{15}$$

where

$$F_i(y)=\text{diag}\left\{\lambda_{Pij}(y,\boldsymbol{X}_i)\{1-\lambda_{Pij}(y,\boldsymbol{X}_i)\}\frac{1-\{\pi(1,\boldsymbol{X}_{1i})/\pi(0,\boldsymbol{X}_{1i})\}}{1-\lambda_{Pij}(y,\boldsymbol{X}_i)+\{\pi(1,\boldsymbol{X}_{1i})/\pi(0,\boldsymbol{X}_{1i})\}\lambda_{Pij}(y,\boldsymbol{X}_i)}\right\}_{j=1}^{n_i},$$

$\boldsymbol{W}_{1i} = (\boldsymbol{w}_{1i1}, \ldots, \boldsymbol{w}_{1in_i})'$ and $\boldsymbol{W}_{2i} = (\boldsymbol{w}_{2i1}, \ldots, \boldsymbol{w}_{2in_i})'$ (see Web Appendix A available online at http://www.biometrics.tibs.org).

4. Finite sample operating characteristics of estimators

In the previous section we outlined a strategy to estimate population model parameters as well as uncertainty estimation in a longitudinal study following (stratified) case-control sampling. We now explore via Monte-Carlo simulation, the impact that misspecification of $\pi(1, \boldsymbol{X}_{1i})/\pi(0, \boldsymbol{X}_{1i})$ and $\lambda_{Sij}(y, \boldsymbol{X}_i)$ can have on inferential validity, and the effect that design and estimation strategy can have on estimation efficiency.

4.1 Population model

The population model we consider is a marginalized transition and latent variable model (Schildcrout and Heagerty, 2007) which is given by:

$$\text{logit}(\mu_{Pij}) = \beta_0+\beta_t t_{ij}+\beta_{x_1}x_{1i}+\beta_{x_2}x_{2i} \quad (16)$$

$$\text{logit}(\mu^c_{Pij})=\Delta_{ij}+\gamma Y_{ij-1}+b_i \quad (17)$$

Equation (16) is the marginal mean model for Y_{ij} which captures the impact of target covariates on the average response, and equation (17) is the conditional mean model for $(Y_{ij}|Y_{i,j-1}, b_i)$ that captures within-subject response dependence. The conditional mean model introduces two sources of dependence among repeated measurements within an individual. Subject-to-subject heterogeneity in predisposition for a positive response ($Y_{ij} = 1$) is introduced by the random intercept b_i, and serial dependence is introduced by the transition term, Y_{ij-1} with coefficient γ. The marginal and conditional mean models, along with the distributional assumption, $b_i \sim N(0, \sigma^2_b)$, complete the multivariate distribution of $[\boldsymbol{Y}_i \mid \boldsymbol{X}_i]$ and allow us to generate data for the population. The value, Δ_{ij}, linking μ_{Pij} and μ^c_{Pij} has been described in a number of earlier manuscripts (e.g., Azzalini, 1994; Heagerty, 1999 and 2002; Schildcrout and Heagerty, 2007). For this simulation, $\boldsymbol{x}_{ij} = (t_{ij}, x_{1i}, x_{2i})'$, $\beta_0 = -2.75$, $\boldsymbol{\beta}_1 = (\beta_t, \beta_{x_1}, \beta_{x_2}) = (0.25, 0.75, 0.75)$, $\sigma_b = 2.5$ and $\gamma = 1$. Covariates x_{1i} and x_{2i} are binary and time-invariant with $\Pr(x_{1i} = 1) = 0.2$ and $\Pr(x_{2i} = 1 \mid X_{1i}) = 0.25 + 0.1X_{1i}$, and t_{ij} is a time covariate with $t_{ij} = j, j \in \{1, \ldots, n_i\}$. We assume a missing completely at random dropout mechanism, where the last follow-up time n_i is uniformly distributed between three and eight. The large σ_b value is intended to reflect the substantial between-subject heterogeneity in ADHD diagnoses; many children never exhibit symptoms or meet diagnostic criteria, while others do so often. This model induces a marginal prevalence at t_{i1} of $\Pr(Y_{i1} = 1) \approx 0.119$.

The sampling covariate Z_i is binary, and we generated its value for each subject using the model, $\lambda_{Pij} = \text{logit}^{-1}(\gamma_0 + \gamma_1 Y_{i1})$ with $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ fixed at two sets of values: (−5, 10) and (−3, 3). We will refer to the former as strong $Z_i \sim Y_{i1}$ dependence ($\text{Str}_{z\sim y}$) and the latter as weak $Z_i \sim Y_{i1}$ dependence (Wk$_{z\sim y}$). Note that with Str$_{z\sim y}$, Z_i is effectively equal to the first response value Y_{i1}. Pr(Z_i = 1) equals 0.124 and 0.101, respectively under Str$_{z\sim y}$ and Wk$_{z\sim y}$.

4.2 Sampling from the population

For the purpose of sampling from the population, we consider both unstratified and stratified approaches, denoted respectively by $S(z)$ and $S(z, x_1)$. In $S(z)$, the sampling probability depends only on Z_i, i.e., $\pi(z, \boldsymbol{X}_{1i}) = \pi(z)$. In $S(z, x_1)$, it depends both on Z_i and subject-level covariate x_{i1}, i.e., $\pi(z, \boldsymbol{X}_{1i}) = \pi(z, x_{i1})$. In a population of size N with N_z, $z \in \{(0, 1)\}$ and N_{z,x_1}, $(z, x_1) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ members in each sampling stratum, we sample each subject according to their stratum membership with probability $\pi(z) = 200/N_z$ for $S(z)$ and $\pi(z, x_1) = 100/N_{z,x_1}$ for $S(z, x_1)$. The number of sampled subjects n_{z,x_1} in each stratum (z, x_1), follows a binomial distribution so that the expected number of controls (Z_i = 0) and cases (Z_i = 1) is equal to 200. For a detailed description of how data were generated, see the Web Appendix B available online at http://www.biometrics.tibs.org.

4.3 Analysis models and model misspecification

In model fitting, for each replicate, we focus on the impact of two forms of model misspecification: 1) misspecified sampling ratio $\pi(1)/\pi(0)$ (for $S(z)$) or $\pi(1, x_1)/\pi(0, x_1)$ (for $S(z, x_1)$), and 2) misspecified $(\boldsymbol{w}_{1,ij}, \boldsymbol{w}_{2,ij})$ in (8) which is equivalent to violation of assumption (7); analysis models are summarized in table 1. In the former case, misspecification occurs during estimation, when we assume the value the sampling ratio is one-half of, one-fifth of, and twice its true value. In the latter case, we consider four functional forms for $(\boldsymbol{w}_{1,ij}, \boldsymbol{w}_{2,ij})$. In order of increasing flexibility, the linear predictors in (8) are given as follows: (lp-1), $\boldsymbol{w}_{1,ij} = \boldsymbol{w}_{2,ij} = (1, x_{1i}, x_{2i})'$; (lp-2), $\boldsymbol{w}_{1,ij} = \boldsymbol{w}_{2,ij} = (1, x_{1i}, x_{2i}, t_{ij})'$; (lp-3), $\boldsymbol{w}_{1,ij} = (1, x_{1i}, x_{2i}, x_{1i} \times x_{2i}, t_{ij}, x_{1i} \times t_{ij}, x_{2i} \times t_{ij})'$ and $\boldsymbol{w}_{2,ij} = (1, x_{1i}, x_{2i}, t_{ij})'$; (lp-4), (lp-3) with the main effect t_{ij} replaced in $\boldsymbol{w}_{1,ij}$ by time-specific indicator variables, and t_{ij} in interaction terms in $\boldsymbol{w}_{1,ij}$ and in $\boldsymbol{w}_{2,ij}$ by a piecewise linear function with a knot at t = 3. In all cases, GEE with exchangeable working covariance weighting (GEE_E) was used for $\boldsymbol{\beta}$ estimation with (12), and Wald-based 95% confidence intervals were computed based on standard errors estimated using (13). Finally, we consider a modeling approach that ignores the study design, i.e., is a standard GEE model for $E(\boldsymbol{Y}_i|\boldsymbol{X}_i)$ with exchangeable correlation matrix. We used the exchangeable working correlation model because, noting that σ = 2.5 and γ = 1, the random intercept is the dominating source of response dependence in these data, and although the true dependence structure also has a serial component, very few standard GEE software packages permit estimation that acknowledges both sources.

4.4 Results: Inferential Validity

We now discuss inferential validity in the presence of possible misspecification of sampling ratio $\pi(1)/\pi(0)$ or $\pi(1, x_1)/\pi(0, x_1)$ and of model (8). Results are displayed in table 2 for the four combinations of $S(z)$ and $S(z, x_1)$, and Str$_{z\sim y}$ and Wk$_{z\sim y}$. As expected, approach 1, which has the most flexible model for Z_i and uses the correctly-specified sampling ratio, does very well in terms of both bias and coverage probability. As with case-control studies, sampling ratio misspecification (approaches 2, 3, 4 and 8) led to biased estimates of the intercept, which is often not a large concern. It also led to biased estimates of the time-varying covariate parameter β_t; this is not unexpected because the slope over time summarizes the change in time-specific intercepts, which are biased due to sampling ratio misspecification. Biases in estimates of time-invariant covariate coefficients were generally modest and less than ten percent, except with $S(z, x_1)$ where estimation approach 8 was severly biased for the stratification variable coefficient, β_{x_1}. Severe inflexibility in the model for Z_i (approach 5), also led to large biases in the intercept and time parameters, but with more flexible approaches such as 6 or 7, these biases were substantially reduced. In most cases, when bias was low or zero for a given parameter estimator, 95% confidence intervals were accurate.

4.5 Results: Estimation Efficiency

Regarding the impact of the design and estimation strategy on parameter estimation efficiency, we consider four specific contrasts (table 3): 1) Str$_{z\sim y}$ versus Wk$_{z\sim y}$ (with $S(z)$ and GEE_E) 2) $S(z, x_1)$ versus $S(z)$ (with Str$_{z\sim y}$ and GEE_E), 3) GEE_E versus independence weighted GEE (GEE_I; with Str$_{z\sim y}$ and $S(z)$), and 4) estimation strategy 1 versus others (using Str$_{z\sim y}$, $S(z)$ and GEE_E). We only consider efficiency for parameter value combinations that yielded approximately valid inferences in the last section (coverage percentages ≥ 92).

Efficiency gains for Str$_{z\sim y}$ over Wk$_{z\sim y}$ were pronounced. Wk$_{z\sim y}$ variances were up to 42% larger for $\hat{\beta}_{x_1}$ and $\hat{\beta}_{x_2}$ and 26% larger $\hat{\beta}_t$. With $S(z, x_1)$, efficiency improvements over $S(z)$ were observed for $\hat{\beta}_{x_1}$ and $\hat{\beta}_{x_2}$ by values as high as 38%. No efficiency improvements were observed for $\hat{\beta}_t$ because stratification variable x_{1i} was unrelated to t_{ij}.

GEE_E improved estimation efficiency only modestly over GEE_I which may be expected for time-invariant covariate estimates; however, we anticipated larger efficiency gains for time-varying covariate coefficient, β_t. We speculate that this is due to the dependence of $\boldsymbol{\beta}$ estimates on $\boldsymbol{\gamma}$ estimates; i.e., $\boldsymbol{I}_{UT}$ from (14) was non-zero. Finally, with the exception of estimation approach 4, approach 1 was no more efficient—and sometimes less so—than the other approaches for the time-invariant covariate coefficients. That it was as efficient as 5, 6, and 7, we gather that estimation of more parameters in 1 does not have a major impact on uncertainty in $\hat{\boldsymbol{\beta}}$. That it was more efficient than 4, and less efficient than 2, 3, and 8, can possibly be explained by the impact of ‘weighting’ of cases relative to controls. By increasing the assumed values of $\pi(1, x_1)/\pi(0, x_1)$, we are effectively giving greater weight to cases, and differential weighting of subjects is known to impact estimation efficiency.

4.6 Summary

To summarize, substantial effort should be made to ascertain reasonable approximations of $\pi(1, x_1)/\pi(0, x_1)$, unless interest is only in time-fixed covariate coefficients. Topic specific experts should be involved in this process, and sensitivity analyses should be conducted over a range of reasonable values in order to examine the extent to which inference would change based on mild to moderate misspecification. Similarly, if time-varying covariate coefficients are of interest, it is important to build a sufficient model for (8). The crucial elements of this model include a flexible functional form of t_{ij}, Y_{ij}, and their interaction. Finally, estimation efficiency can be improved by choosing study designs with stratification and strong relationships between Z_i and $\boldsymbol{Y}_i$.

5. Application to natural history studies of childhood mental health disorders

Participants of the ADHD study were sampled on the basis of whether ($Z_i = 1$) or not ($Z_i = 0$) they were referred to one of the two participating clinics. The study was matched on gender, G_i, and can be thought of as stratified since the probability of being sampled depended upon the pair, (Z_i, G_i). Patient referral was strongly related to ADHD symptom diagnosis at baseline as $\Pr(Y_{i1} = 1 \mid Z_i = 1) \approx 0.92$ and $\Pr(Y_{i1} = 1 \mid Z_i = 0) \approx 0.02$, but this relationship was not deterministic. The demographic characteristics among referred and non-referred participants were similar. Both groups were approximately 82% male, 64% white, 31% african-american, and 6% were classified as “other” ethinicity. Age distributions were also similar with a median value of 5 years.

The primary goal of this analysis is to estimate the time trend of ADHD prevalence for boys ($G_i = 0$) and girls ($G_i = 1$) separately, and to examine whether this trend differs between them. The impact of race/ethnicity and age at baseline were of secondary interest. Similar to section 4, we examine the impact of assumptions about $\pi(1, g)/\pi(0, g)$ and those in auxiliary model (8). We considered six reasonable analysis approaches plus the naive analysis that ignored the design altogether. We assume that approximately five percent of girls in the population would qualify for referral and among boys this rate is likely to be higher. We considered the values, five, ten, and fifteen percent for boys. In our sample, 25 out of 46 girls were cases, and with five percent prevalence, $\Pr(Z_i = 1 \mid G_i = 1) = 0.05$, we have $\pi(1, 1)/\pi(0, 1) = (25 \cdot 0.95)/(21 \cdot 0.05) = 22.6$. Among boys, there were 113 cases and 96 controls. With $\Pr(Z_i = 1 \mid G_i = 0)$ equal to 0.05, 0.10, and 0.15, $\pi(1, 0)/\pi(0, 0)$ equals 22.4, 10.6, and 6.7, respectively. Next, we considered two linear predictors in auxiliary model (8). In the simpler model, the linear predictor included: Y_{ij}, t_{ij}, age at baseline, gender, African American ethnicity, “other” ethnicity, and all pairwise interactions with Y_{ij}. The more flexible model (8), was identical to the simpler one except the main effect of t_{ij} and its interaction with Y_{ij} were replaced with time-specific indicator variables.

Results from these analyses are displayed in table 4. The naive analysis yielded vastly different conclusions than analyses that acknowledge the biased study design. While t_{ij} was significantly and positively associated with ADHD prevalence in boys ($G_i = 0$) among all analyses that acknowledged the study design, it was significantly and negatively associated with ADHD in the naive analysis. The prevalence time trend for girls in the naive analysis was positive (although not significant), but was flat when study design was taken into account. Similarly, gender appeared to be independent of ADHD prevalence at baseline ($t_{ij} = 0$) in the naive analysis while four of the six other approaches showed substantial evidence of females being at lower risk for ADHD than males. Among the other six analyses, the assumed values of $\pi(1, g)/\pi(0, g)$ had a far larger impact on conclusions than did choice of the linear predictor in model (8). Lower assumed prevalence of referral status ($Z_i = 1$) among boys or equivalently, higher $\pi(1, 0)/\pi(0, 0)$ values, resulted in smaller effect sizes (in magnitude) for gender at baseline. The estimated baseline log odds ratios for girls versus boys ranged from approximately −1.05 to −0.4. The magnitude of the ADHD prevalence time trend for boys was also highly impacted by $\pi(1, 0)/\pi(0, 0)$, with higher values being associated with larger time trend estimates. It is interesting to note that the time trend among females agreed very closely across the six approaches; in all cases, the values of the coefficients for t_{ij} and $t_{ij} \cdot G_i$ were in the opposite direction, and were comparable in magnitude.

6. Discussion

In this manuscript we discussed design and analysis considerations for stratified and unstratified case-control sampling followed by longitudinal followup on a stochastically related binary response vector. We developed a GEE-based estimation strategy as well as robust standard error calculations that incorporate information and uncertainty associated with the study design into the analysis via an ancillary model for case-control status. We found for time-invariant covariate coefficients that the biased design does not have a major impact on inferential validity, as most estimation approaches performed reasonably well. This result may be expected given the well-known performace of logistic regression under case-control sampling. There was one exception which involved the naive analyses with a stratified design. In this case, inferences related to the stratification variable should not be trusted as estimates are likely to exhibit large biases. Misspecification of the ancillary case-control model can lead to invalid inferences on time-varying covariate coefficients; however, as long as this model is reasonably well specified and includes flexible functions of time in the linear predictor, analyses should perform reasonable well. Specification of the sampling ratio was also shown to have a major impact on the validity of analyses related to time-varying covariates, and, since this value is often unknown, appropriate specification of it is a major challenge. Content-specific experts should be involved in determining its value, and sensitivity analyses over a reasonable range should be conducted to examine the impact on inferences. While we found that stratification and strong dependence between the sampling covariate and response vector can improve estimation efficiency, covariance weighting only improved efficiency slightly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Rathouz's contribution to this research was funded in part by National Institute of Mental Health grants R01 MH62437 and R01 MH53554.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders : (DSM-IV). 4th edition. Washington, DC: American Psychiatric Association; 1994.

Azzalini A. Logistic regression for autocorrelated data with application to repeated measures (Corr: 97V84 p989). Biometrika 1994;81:767–775.

Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. Biometrics 1995;51:309–317. [PubMed: 7766784]

Hartung C, Willcutt E, Lahey B, Pelham W, Loney J, Stein M, Keenan K. Sex differences in young children who meet criteria for attention deficit hyperactivity disorder. J Clin Child Adolesc Psychol 2002;31:453–464. [PubMed: 12402565]

Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. Biometrics 1999;55:688–698. [PubMed: 11314994]

Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics 2002;58:342–351. [PubMed: 12071407]

Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika 2001;88:973–985.

Lahey B, Pelham W, Stein M, Loney J, Trapani C, Nugent K, Kipp H, Schmidt E, Lee S, Cale M, Gold E, Hartung C, Willcutt E, Baumann B. Validity of DSM-IV attention-deficit/hyperactivity disorder for younger children. J Am Acad Child Adolesc Psychiatry 1998;37:695–702. [PubMed: 9666624]

Lahey BB, Gordon RA, Loeber R, Stouthamer-Loeber M, Farrington DP. Boys who join gangs: A prospective study of predictors of first gang entry. Journal of Abnormal Child Psychology 1999;27:261–276. [PubMed: 10503645]

Lee A, McMurchy L, Scott A. Re-using data from case-control studies. Stat Med 1997;16:1377–1389. [PubMed: 9232759]

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.

Liang K-Y, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data (Disc: P24-40). Journal of the Royal Statistical Society, Series B: Methodological 1992;54:3–24.

Mancl LA, Leroux BG. Efficiency of regression estimates for clustered data. Biometrics 1996;52:500–511. [PubMed: 10766502]

Neuhaus J, Scott AJ, Wild CJ. The analysis of retrospective family studies. Biometrika 2002;89:23–37.

Neuhaus JM, Jewell NP. The effect of retrospective sampling on binary regression models for clustered data. Biometrics 1990;46:977–990. [PubMed: 2085642]

Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. International Statistical Review 1991;59:25–35.

Neuhaus JM, Scott AJ, Wild CJ. Family-specific approaches to the analysis of case-control family data. Biometrics 2006;62:488–494. [PubMed: 16918913]

Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics: Simulation and Computation 1994;23:939–951.

Qaqish BF, Zhou H, Cai J. On case-control sampling of clustered data. Biometrika 1997;84:983–986.

Schildcrout J, Heagerty P. Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. Biostatistics 2005;6:633–652. [PubMed: 15917376]

Schildcrout JS, Heagerty PJ. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. Biostatistics 2008;9:735–749. [PubMed: 18372397]

Ten Have TR, Kunselman AR, Tran L. A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. Stat Med 1999;18:947–960. [PubMed: 10363333]

Whittemore AS. Logistic regression of family data from case-control studies (Corr: 97V84 p989-990). Biometrika 1995;82:57–67.

Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: A generalized estimating equation approach (Corr: V45 p347). Biometrics 1988;44:1049–1060. [PubMed: 3233245]

Zhao LP, Hsu L, Holte S, Chen Y, Quiaoit F, Prentice RL. Combined association and aggregation analysis of data from case-control family studies. Biometrika 1998;85:299–315.

Table 1

Estimation Strategies: In strategy 1, the sampling ratio, $\pi(1)/\pi(0)$ or $\pi(1, x_1)/\pi(0, x_1)$, is correctly specified, and the model (8) is very flexible. In strategies 2, 3, and 4, we misspecify the sampling ratio by assuming it is one half, one fifth and twice its true value, respectively. In strategies 5, 6, and 7, we assume three less flexible models for (8), given in the text. Approach 8 ignores the design altogether with analyses conducted as if the sample was representative of the target population.

Estimation Strategy	**Specification of $\pi(1, x_1)/\pi(0, x_1)$**	**Specification of Model (8)**
1	$\pi(1, x_1)/\pi(0, x_1)$	(lp-4)
2	$0.5 \cdot \pi(1, x_1)/\pi(0, x_1)$	(lp-4)
3	$0.2 \cdot \pi(1, x_1)/\pi(0, x_1)$	(lp-4)
4	$2.0 \cdot \pi(1, x_1)/\pi(0, x_1)$	(lp-4)
5	$\pi(1, x_1)/\pi(0, x_1)$	(lp-1)
6	$\pi(1, x_1)/\pi(0, x_1)$	(lp-2)
7	$\pi(1, x_1)/\pi(0, x_1)$	(lp-3)
8	Ignored	Ignored

Table 2

Percent bias in parameter estimates and coverage percentages in eight estimation strategies described in table 1 and across 1500 replicates. The **γ** values (−5, 10) and (−3, 3) correspond to model parameters in $\Pr(Z_i \mid Y_{i1})$, and represent strong ($\mathrm{Str}_{z\sim y}$) and weak ($\mathrm{Wk}_{z\sim y}$) dependence of Z_i on Y_{i1}, respectively. GEE_E was used for estimation, and sampling was based Z_i in the unstratified designs (S(z)) and on (Z_i, x_{1i} in the stratified designs (S(z, x_1)). Percent bias in parameter estimates is calculated with $100 \cdot (\hat{\beta}_k - \beta_k)/\beta_k$ for $k \in (0, t, x_{11}, x_{2i})$. Coverage percentages were calcualted as the percent of nominal 95% Wald confidence intervals using robust standard errors spanning the true parameter value.

Estimation Approach	β_0	β_t	β_{x_1}	β_{x_2}	$cp(\beta_0)$	$cp(\beta_t)$	$cp(\hat{\beta}_{x_1})$	$cp(\hat{\beta}_{x_2})$
				$\mathrm{Str}_{z\sim y}$ with $S(z)$				
1	−1	−3	1	3	95	94	95	94
2	−23	−29	2	3	1	19	95	95
3	−52	−57	2	4	0	0	95	96
4	20	26	3	5	11	53	94	93
5	−24	−65	−5	−3	0	0	94	95
6	−5	−13	−1	1	86	79	94	95
7	−4	−12	1	2	87	81	94	94
8	−63	−65	2	4	0	0	96	95
				$\mathrm{Wk}_{z\sim y}$ with $S(z)$				
1	0	−1	0	0	94	94	95	94
2	−11	−15	2	1	56	74	95	95
3	−29	−35	5	4	0	4	95	93
4	8	11	−2	−1	84	90	95	94
5	−17	−46	−3	−3	24	0	95	94
6	−2	−7	−1	−2	94	92	95	94
7	−1	−6	0	0	94	92	95	94
8	−40	−46	7	5	0	0	94	94
				$\mathrm{Str}_{z\sim y}$ with $S(z, x_1)$				
1	−1	−2	−4	2	93	94	92	95
2	−23	−28	−3	3	3	19	93	96
3	−52	−55	−3	4	0	0	93	95
4	20	26	−4	3	19	51	92	94
5	−22	−59	−9	−4	4	0	93	96

Estimation Approach	β_0	β_t	β_{x1}	β_{x2}	$cp(\hat{\beta}_0)$	$cp(\hat{\beta}_t)$	$cp(\hat{\beta}_{x1})$	$cp(\hat{\beta}_{x2})$
6	−4	−11	−6	0	87	80	93	95
7	−4	−11	−5	1	88	82	93	95
8	−65	−59	−62	3	0	0	22	95
				Wk$_{z\sim y}$ with $S(z, x_1)$				
1	0	−1	−1	1	94	94	95	95
2	−11	−15	0	3	63	73	95	96
3	−29	−34	2	5	1	3	96	95
4	9	11	−1	0	85	90	95	95
5	−15	−43	−3	−2	43	0	95	96
6	−2	−6	−2	−1	94	91	95	96
7	−1	−5	−1	1	93	92	95	96
8	−40	−43	−16	6	0	0	90	95

Table 3

Study design and estimation efficiency across 1500 replicates. Relative efficiency of A vs B is defined by the empirical variance of B divided by the empirical variance of A across 1500 replications. We only consider parameter by estimation strategies that were observed to be approximately valid in table 2. The upper portion of the table shows efficiency gains due to study design. We compare the efficiency of $\text{Str}_{z\sim y}$ versus $\text{Wk}_{z\sim y}$ with S(z) and GEE_E, and S(z, x_1) versus S(z) with $\text{Str}_{z\sim y}$ and GEE_E. In the bottom portion of the table, we show efficiency gains due to estimation strategy. First, we compare GEE_E versus GEE_I with $\text{Str}_{z\sim y}$ and S(z), and second we show the impact of using estimation strategy 1 versus other estimation approaches with $\text{Str}_{z\sim y}$, S(z), and GEE_E.

	Study design							
	$\text{Str}_{z\sim y}$ versus $\text{Wk}_{z\sim y}$				**$S(z, x_1)$ versus $S(z)$**			
Estimation strategy	**β_0**	**β_t**	**β_{x_1}**	**β_{x_2}**	**β_0**	**β_t**	**β_{x_1}**	**β_{x_2}**
1	1.84	1.26	1.42	1.42	0.75	1.02	1.32	1.2
2	-	-	1.31	1.31	-	-	1.28	1.16
3	-	-	1.09	1.11	-	-	1.23	1.09
4	-	-	1.37	1.37	-	-	1.38	1.21
5	-	-	1.38	1.39	-	-	1.25	1.23
6	-	-	1.41	1.42	-	-	1.25	1.2
7	-	-	1.39	1.37	-	-	1.29	1.23
8	-	-	1.05	1.07	-	-	-	1.07
	Estimation procedure							
	GEE_E versus GEE_I				Strategy 1 versus others			
	β_0	β_t	β_{x1i}	β_{x2i}	β_0	β_t	β_{x1i}	β_{x2i}
1	1.04	1.08	1.05	1.03	1	1	1	1
2	-	-	1.09	1.07	-	-	0.84	0.83
3	-	-	1.11	1.08	-	-	0.76	0.74
4	-	-	0.99	0.98	-	-	1.31	1.3
5	-	-	1.06	1.03	-	-	0.97	0.99
6	-	-	1.04	1.01	-	-	0.97	0.98
7	-	-	1.04	1.02	-	-	1.04	1.05
8	-	-	1.09	1.06	-	-	0.76	0.73

Table 4

ADHD Study results: A gender stratified design was used to examine the timecourse of ADHD symptom exhibition in males and females separately and the difference in the trajectories. Linear t_{ij} columns correspond to estimates where the functional form of t_{ij} in the intermediate model (8) was assumed to be linear. With flexible t_{ij}, time-specific indicator variables were substituted for linear t_{ij}. We display parameter estimates on the log odds scale and the 95% confidence intervals are in parentheses.

	$\pi(1, 0)/\pi(0, 0) = 22.4$		**$\pi(1, 0)/\pi(0, 0) = 10.6$**		**$\pi(1, 0)/\pi(0, 0) = 6.7$**		**Naive**
	Flexible t_{ij}	**Linear t_{ij}**	**Flexible t_{ij}**	**Linear t_{ij}**	**Flexible t_{ij}**	**Linear t_{ij}**	
Time (years)	0.13 (0.06, 0.19)	0.10 (0.05, 0.16)	0.09 (0.04, 0.14)	0.08 (0.03, 0.12)	0.06 (0.02, 0.11)	0.06 (0.01, 0.10)	−0.04 (−0.07, −0.01)
Age (years) -5	−0.25 (−0.59, 0.09)	−0.18 (−0.51, 0.14)	−0.20 (−0.50, 0.11)	−0.16 (−0.46, 0.13)	−0.17 (−0.45, 0.12)	−0.14 (−0.42, 0.13)	−0.09 (−0.34, 0.15)
Female	−0.41 (−1.12, 0.30)	−0.50 (−1.18, 0.17)	−0.77 (−1.45, −0.08)	−0.80 (−1.46, −0.14)	−1.05 (−1.72, −0.37)	−1.05 (−1.71, −0.39)	0.00 (−0.66, 0.66)
Female · Time	−0.12 (−0.23, −0.01)	−0.10 (−0.20, −0.01)	−0.08 (−0.19, 0.02)	−0.08 (−0.17, 0.02)	−0.06 (−0.16, 0.05)	−0.06 (−0.16, 0.04)	−0.09 (−0.19, 0.01)
Afr Am Ethnicity	1.29 (0.73, 1.85)	0.96 (0.43, 1.48)	1.06 (0.57, 1.56)	0.88 (0.40, 1.36)	0.94 (0.48, 1.41)	0.82 (0.37, 1.27)	0.54 (0.13, 0.95)
Other Ethnicity	0.11 (−0.97, 1.18)	0.00 (−1.05, 1.06)	0.17 (−0.82, 1.16)	0.11 (−0.87, 1.08)	0.22 (−0.74, 1.17)	0.17 (−0.77, 1.12)	0.38 (−0.51, 1.27)
Intercept	−2.21 (−2.66, −1.77)	−2.05 (−2.48, −1.63)	−1.84 (−2.21, −1.46)	−1.75 (−2.12, −1.38)	−1.55 (−1.89, −1.20)	−1.49 (−1.83, −1.16)	−0.05 (−0.36, 0.25)