Statistics and Its Interface Volume 7 (2014) 75–85

Approaches to retrospective sampling for longitudinal transition regression models

Sally Hunsberger*,†, Paul S. Albert‡, and Marie Thoma‡

For binary diseases that relapse and remit, it is often of interest to estimate the effect of covariates on the transition process between disease states over time. The transition process can be characterized by modeling the probability of the binary event given the individual's history. Designing studies that examine the impact of time varying covariates over time can lead to collection of extensive amounts of data. Sometimes it may be possible to collect and store tissue, blood or images and retrospectively analyze this covariate information. In this paper we consider efficient sampling designs that do not require biomarker measurements on all subjects. We describe appropriate estimation methods for transition probabilities and functions of these probabilities, and evaluate efficiency of the estimates from the proposed sampling designs. These new methods are illustrated with data from a longitudinal study of bacterial vaginosis, a common relapsing-remitting vaginal infection of women of child bearing age.

Keywords and phrases: Weighted maximum likelihood, Survey sampling, Markov model.

1. INTRODUCTION

For chronic diseases that flair up and then go away (relapse and remit) or for diseases that can occur repeatedly over time, it is often of interest to estimate the effect of covariates on the transition process between disease states over time. For a binary disease, the transition process can be characterized by modeling the probability of the binary event, given the individual's history of prior events and important time-dependent or subject-specific covariates. These so-called transition models have been used to describe the effect of sleep on episodes of depression or mania in patients with bipolar disorder [9], the relationship between age and asthma events [15], or the relationship of sucking patterns in premature infants to time dependent covariates [17].

With time varying covariates, information on the covariates must be collected at the same time points as the response. This can lead to extensive amounts of data being collected. Sometimes it may be possible to collect and store tissue, blood or images for later evaluation. A retrospective analysis of covariate information may be desirable since it allows any new technology that is developed during the course of a study to be used on samples from all individuals at all time points. Also, since longitudinal studies may be performed over many years, it allows investigators to identify and answer important new scientific questions years after follow-up is complete.

Evaluation of biomarker covariate data (which are often based on measurements obtained from assays on tissue, blood or images) on all subjects and all time points can be expensive. Therefore, it is of interest to consider efficient sampling designs that do not require evaluation of biomarker measurements on all patients. We evaluate the efficiency of sampling subjects from a longitudinal study and describe appropriate estimation techniques. Note, this implies that for a subject who is sampled all covariates at all times will be obtained. This approach allows for efficient estimation of transition probabilities and functions of the transition probabilities.

Others [13, 10] have considered this sampling problem for analyzing binary longitudinal data, but have focused on estimating the marginal probability of response. Schildcrout & Heagerty proposed obtaining covariate data on subjects where the longitudinal responses are not all the same. That is, subjects with all positive responses or all negative responses would not be evaluated, while subjects with both positive and negative response over time would have their longitudinal covariate data evaluated. In their estimation approach they accounted for the sampling design using maximum likelihood with conditional logistic regression models. Further, they showed that, in some situations, this sampling design provides estimates of the parameters associated with covariates that are nearly fully efficient. Neuhaus & Jewell considered an approach based on sampling longitudinal binary response data when the sampling depends on the response patterns. They discussed the use of weighted logistic regression to analyze the data and obtained unbiased estimates of the parameters along with correct estimation of the variance of the parameter estimates when estimating the marginal probability of response. Schildcrout & Heagerty extend their previous paper by stratifying on a coarser categorization of the total number of responses.

In this paper, we focus on transition models for analyzing longitudinal binary data and the efficiency of estimating

*Corresponding author.
†National Cancer Institute.
‡Eunice Kennedy Shriver National Institute of Child Health and Human Development, Intramural Program.

parameters from the transition models under various sampling and estimation methods. The efficiency of the following sampling and corresponding parameter estimation methods for a transition model are compared: 1) A subset of subjects from the full data set are randomly sampled. For each sampled subject all measurements for each time point are obtained and maximum likelihood is used to estimate the parameters of the transition model. 2) A specified number of subjects are randomly sampled from each pattern of response (stratum) and weighted maximum likelihood is used to estimate the parameters of the transition model. Again, all observations per sampled subject are obtained. 3) Schildcrout and Heagerty's (2008) method of only sampling from subjects with longitudinal measurements that are a mix of positive and negative outcomes (if a subject is sampled all measurements are obtained). In this approach the data are analyzed using maximum likelihood with a conditional logistic regression type model. We consider sampling methods when an individual is sampled and all longitudinal measurements are obtained.

The interest in this problem arose from The Longitudinal Study of Vaginal Flora (LSVF), which enrolled 3,620 participants presenting for routine health care at Birmingham, Alabama health clinics from August 1999 to February 2002. Details of this study have been described elsewhere [6]. At baseline and quarterly for up to 1 year, vaginal symptoms were recorded and a vaginal swab was collected to diagnose bacterial vaginosis (BV). Additionally, a vaginal wash sample was obtained at each visit and stored for future evaluation.

Microbiologically, BV is characterized by an imbalance in vaginal flora [12]. In research settings, the diagnosis of BV is often made using Nugent Gram stain criteria [8]. This method assigns a score from 0 to 10 based on the prevalence of anaerobic bacteria and inverse prevalence of lactobacilli [11]. Clinically, a diagnosis of BV is made if three of the four criteria are present: 1) homogenous, white adherent vaginal discharge, 2) vaginal pH > 4.5, 3) detection of include cells on saline wet smear, and 4) presence of amine odor with additional of KOH (whiff test) [1]. This method is preferred in clinical settings because it is relatively simple to perform and allows point of care treatment. BV is a relapsing-remitting condition which is characterized by periods of BV followed by periods without BV.

Recent technological advances in DNA sequencing have opened up a new field of molecular amplification-based research on the human microbiota, including detection of previously uncultivatable organisms. A question of interest within the LSVF was whether these new methods could be applied to stored vaginal wash samples to evaluate human vaginal microflora. A further question is whether these organisms initiate an active period of BV or initiate a remission. Other interest is on examining the relationship between these new biomarkers and the natural history of BV as measured by either Gram staining (dichotomized) or a clinical diagnosis. However, in order to preserve samples and minimize cost, it was of interest to use a sampling design that reduced the number of samples required for analysis. Although this particular study has not been completed, data from the LSVF was used to demonstrate methods described in this paper.

The outline of the paper is as follows. Section 2 discusses the model and estimation. Section 3 describes a simulation study and Section 4 presents the results of the simulation study. Section 5 demonstrates the approach on the BV data set. A discussion follows in Section 6.

2. METHODS

The following first-order Markov regression model [3] can be used to estimate transition probabilities:

$$\text{logit}\ (\text{Prob}(Y_{ij} = 1|Y_{ij-1}, X_{ij})) = \beta_0 + \beta_1 X_{ij} + \beta_2 Y_{ij-1}, \tag{1}$$

where X_{ij} is the biomarker value and Y_{ij} is a binary indicator of outcome status for the j^{th} measurement on the i^{th} subject. We assume the Y_{ij} are equally spaced and observed at the same time intervals for all individuals, which is a reasonable approximation for the LSVF data set.

A slightly more flexible model [16] allows for an interaction between the biomarker and the previous lagged outcome:

$$\begin{aligned}\text{logit}(\text{Prob}(Y_{ij} = 1|Y_{ij-1}, X_{ij})) = \beta_0 + \beta_1 X_{ij} + \\ \beta_2 Y_{ij-1} + \beta_3 X_{ij} Y_{ij-1}.\end{aligned} \tag{2}$$

An underlying assumption in Markov models is that the transition probabilities depend on the past history through a specified number of past or lagged outcomes. Higher order Markov dependence can be incorporated by including additional lagged outcome terms (e.g. a term corresponding to Y_{ij-2}).

For a first-order model, the transition probability matrix for a specified covariate X corresponding to 2 is

$$P(X) = \begin{bmatrix} 1 - p_{01}(X) & p_{01}(X) \\ 1 - p_{11}(X) & p_{11}(X) \end{bmatrix}$$

where $p_{01}(X) = \text{Prob}(Y_{ij} = 1|Y_{ij-1} = 0, X_{ij}) = \text{logit}^{-1}(\beta_0 + \beta_1 X_{ij})$ and $p_{11}(X) = \text{Prob}(Y_{ij} = 1|Y_{ij-1} = 1, X_{ij}) = \text{logit}^{-1}(\beta_0 + \beta_1 X_{ij} + \beta_2 + \beta_3 X_{ij})$. The subscripts on p are used to indicate the states, so for example p_{01} is used to denote transitioning from state 0 to 1. The diagonal elements characterize the probability of staying in the same state and the off-diagonal elements characterize the probability of transitions.

In order to estimate the transition matrices, estimates of the parameters in equation (2) must be obtained. In a longitudinal data set with covariates collected at all time points on all subjects or on a random sample of subjects, maximum likelihood can be used to obtain parameter and variance estimates that are asymptotically consistent.

As discussed in Neuhous & Jewell when outcome dependent sampling is performed and the X_{ij} are sampled with probability only depending on Y_{ij} (not on other Y_{ik} values for subject i or on $\boldsymbol{Y}_j$, where $\boldsymbol{Y}_j$ is the vector of outcome responses for subject j), the maximum likelihood estimates (MLE) of β_1 and β_3 from (2) will be asymptotically unbiased, but estimates of β_0 and β_2 will be biased. Since our interest is in estimating the transition probabilities, it is important to obtain unbiased estimates of all parameters. Thus, we explore using weighted maximum likelihood (WMLE) with outcome dependent sampling to estimate the parameters from the transition model. Due to the outcome dependent sampling, the standard variance estimates obtained from the weighted likelihood using maximum likelihood estimation are not correct.

Obtaining correct weighted maximum likelihood variance estimates has been discussed in the survey literature. We use these methods to calculate correct variance estimates of the WMLE parameter estimates. In survey data it is common for a data set to include subjects with unequal probability of selection. With outcome dependent sampling, unless all strata occur with equal probability, subjects will be sampled with unequal probability. Weights which are the inverse of the probability of selection (or the inverse of the sampling fraction) are used with survey data to account for the unequal probability of selection. In this setting, the probability of selection is the number of subjects sampled from a particular pattern of response (stratum) divided by the total number of subjects in the full data set with that particular pattern of response. The weights are the same for all observations within a subject and for all subjects in a stratum.

The correct variance estimates are derived as follows. Since the WMLE parameter estimates are differentiable functions of the data [5, 14, 4], the correct variance estimates can be obtained by using the Taylor linearization method [2]. We first form the l vector (where l is the number of parameters in the model) of Taylor deviates $\boldsymbol{z}_{hij}$ which is obtained by differentiating the sample weighted estimators (in this case the WMLE's) with respect to their weights [5, 14, 4]. Then

$$\boldsymbol{z}_{hij} = w_{hij}\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial w_{hij}} \qquad \text{and}$$

$$\frac{\partial\hat{\boldsymbol{\beta}}}{\partial w_{hij}} = \left[\sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{m} w_{hij}\boldsymbol{x}_{hij}\boldsymbol{x}'_{hij}(1-\hat{p}_{hij})\hat{p}_{hij}\right]^{-1}\boldsymbol{x}_{hij}(y_{hij}-\hat{p}_{hij})$$

for stratum h, individual i, and measurement j. (The appendix provides details on the derivation of $\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial w_{hij}}$). Now the sample-weighted sum $\boldsymbol{T} = \sum_h^H\sum_i^{I_h}\sum_j^m w_{hij}\boldsymbol{z}_{hij}$ (where H is the number of strata, I_h is the number of individuals sampled in the stratum and m is the number of observations per individual) is an asymptotically consistent linear approximation to the WMLE parameter estimates. Therefore the variance estimate of T is a consistent estimator for the variance estimate of the WMLE parameter estimates [5]. We can compute $\widehat{Var(\hat{\boldsymbol{\beta}})}$ by computing the variance of $\hat{\boldsymbol{T}}$ based on the sample design. In this case, it is that of a multistage stratified cluster sample [7]. Let $\boldsymbol{z}_{hi\cdot} = \sum_{j=1}^m \boldsymbol{z}_{hij}$ and $\overline{\boldsymbol{z}}_h = \frac{1}{I_h}\sum_{i=1}^{I_h}\boldsymbol{z}_{hi\cdot}$ and $\widehat{var}(\boldsymbol{z}_{hi\cdot}) = \sum_{i=1}^{t_h}(\boldsymbol{z}_{hi\cdot}-\overline{\boldsymbol{z}}_h)(\boldsymbol{z}_{hi\cdot}-\overline{\boldsymbol{z}}_h)'/(t_h-1)$ then

$$\widehat{var}\left(\sum_h^H\sum_i^{I_h}\sum_j^m w_{hij}\boldsymbol{z}_{hij}\right) =$$

$$\widehat{var}\left(\sum_h^H\sum_i^{I_h}\boldsymbol{z}_{hi\cdot}\right) = \sum_h^H\sum_i^{I_h}\widehat{var}(\boldsymbol{z}_{hi\cdot}) =$$

$$\sum_{h=1}^H\frac{I_h}{I_h-1}\sum_{i=1}^{I_h}(\boldsymbol{z}_{hi\cdot}-\overline{\boldsymbol{z}}_h)(\boldsymbol{z}_{hi\cdot}-\overline{\boldsymbol{z}}_h)'.$$

In the simulations we also perform the outcome dependent sampling method of Shildcrout & Heagerty where subjects with discrepant outcomes (i.e. a subject has both positive and negative outcomes across time) are randomly sampled (all observations within a subject are obtained) and parameters are estimated by maximum likelihood. We briefly review the estimation procedure and implementation in this paper. Let $S_i = \sum_{j=1}^m Y_{ij}$, then the likelihood is

$$L = \prod_{i=1}^N P(\boldsymbol{Y}_i|0<S_i<m)I_{(0<S_i<m)}$$

where N is the number of subjects in the sample and $I_{(\cdot)}$ is the indicator function so that when the condition is true the function is 1 and 0 otherwise. Now let

$$P(\boldsymbol{Y}_i|0<S_i<m) = \frac{P(Y_{i1}=y_{i1})\prod_{j=2}^m P_{y_{ij-1},y_{ij}}(x_{ij})}{1-P(Y_{i1}=0)\prod_{j=2}^m P_{00}(x_{ij})-P(Y_{i1}=1)\prod_{j=2}^m P_{11}(x_{ij})}$$

where

$$P_{y_{ij-1},1}(x_{ij}) = \frac{\exp(\beta_0+\beta_1x_{ij}+\beta_2y_{ij-1}+\beta_3x_{ij}y_{ij-1})}{1+\exp(\beta_0+\beta_1x_{ij}+\beta_2y_{ij-1}+\beta_3x_{ij}y_{ij-1})}.$$

Note, $P(Y_{i1})$ is from the full data set and not just the outcome dependent sample.

3. SIMULATIONS

We performed simulations to compare bias and efficiency of different sampling and estimation methods. In the rest of this paper we refer to the sampling/estimation methods as follows.

Table 1. Medians(true), median absolute differences (MAD) of parameter estimates and median of estimated asymptotic standard errors (SE). Outcome data is generated according to $logitP(Y_{ij}=1|Y_{ij-1},X_{ij})=\beta_0+\beta_1X_{ij}+\beta_2Y_{ij-1}+\beta_3X_{ij}Y_{ij-1}$. *The covariates are independent and are generated according to Simulation 1 with binary* $x_{ij}\sim binomial(1,.5)$ *or with continuous* $x_{ij}\sim .5+N(0,1)$. *For the top half of the table the marginal probability of a positive response,* $P(Y^*)$, *is .24 and in the bottom half* $P(Y^*)$ *is .024. The mean number of observations sampled for each simulation is 320 for* $\beta_0=-1.5$ *with binary and continuous* x *and 120 and 132 for* $\beta_0=-4.0$ *with binary and continuous* x

	Binary				Continuous			
	NOD MLE	strat WMLE	strat MLE	DOD CMLE	NOD MLE	strat WMLE	strat MLE	DOD CMLE
med $\beta_0(-1.5)$	−1.503	−1.501	−0.268	−1.501	−1.503	−1.501	−0.314	−1.500
MAD	0.118	0.128	0.089	0.125	0.099	0.083	0.057	0.113
med asy SE	0.118	0.125	0.117	0.128	0.099	0.081	0.098	0.114
med $\beta_1(.5)$	0.502	0.505	0.503	0.499	0.502	0.504	0.502	0.501
MAD	0.157	0.237	0.167	0.144	0.080	0.124	0.085	0.077
med asy SE	0.156	0.232	0.160	0.143	0.082	0.122	0.086	0.077
med $\beta_2(.2)$	0.193	0.201	−0.110	0.194	0.196	0.201	−0.126	0.194
MAD	0.231	0.172	0.125	0.219	0.196	0.125	0.086	0.184
med asy SE	0.230	0.176	0.167	0.214	0.193	0.120	0.141	0.186
med$\beta_3(.2)$	0.202	0.200	0.194	0.203	0.198	0.201	0.200	0.202
MAD	0.306	0.315	0.230	0.267	0.164	0.180	0.128	0.145
med asy SE	0.304	0.325	0.227	0.266	0.163	0.176	0.126	0.145
med $\beta_0(-4.0)$	−4.043	−4.001	−1.044	−3.999	−4.031	−4.004	−1.077	−3.995
MAD	0.496	0.247	0.157	0.251	0.388	0.187	0.117	0.226
med asy SE	0.168	0.168	0.169	0.168	0.394	0.184	0.149	0.231
med$\beta_1(.5)$	0.507	0.508	0.503	0.501	0.507	0.513	0.501	0.502
MAD	0.671	0.430	0.238	0.233	0.291	0.247	0.118	0.115
med asy SE	0.168	0.169	0.168	0.168	0.287	0.237	0.118	0.115
med $\beta_2(.2)$	−12.937	0.184	−0.661	0.071	−13.737	0.177	−0.706	0.070
MAD	4.423	0.427	0.362	1.195	18.392	0.367	0.284	0.856
med asy SE	0.168	0.169	0.169	0.168	1640.742	0.312	0.322	0.817
med$\beta_3(.2)$	3.191	0.212	0.210	0.310	0.178	0.216	0.209	0.218
MAD	13.695	0.655	0.497	1.828	3.586	0.387	0.249	0.575
med asy SE	0.168	0.169	0.169	0.169	1499.745	0.361	0.242	0.561

NOD/MLE Non-outcome dependent sampling. Randomly sample any subject from the full data set, use maximum likelihood for estimation.

strat/MLE Outcome dependent stratified sampling. Ten subjects are randomly sampled within each stratum, use maximum likelihood for estimation.

strat/WMLE Outcome dependent stratified sampling. Ten subjects are randomly sampled within each stratum, use weighted maximum likelihood for estimation.

DOD/CMLE Discrepant outcome dependent sampling where only subjects with discrepant outcomes are randomly sampled. Use maximum likelihood estimation from a conditional logistic regression type model.

In all simulations for each replication, longitudinal data were generated for 10,000 subjects at five time points. With five time points there are $2^5=32$ possible strata. Then, a sample of 320 subjects (with all data at each time point) were selected from the data set of 10,000 using the different sampling methods. Weights for each pattern of response (stratum) were estimated using the 10,000 subjects (note, the same stratum weight was used for each subject and observation in a stratum).

In all simulations we explore two values for the intercept, β_0, so that the prevalence of an event is not uncommon ($\beta_0=-1.5$) and also where the prevalence is very small ($\beta_0=-4.0$). The initial values for each simulation were generated with $logit(P(Y_{i0}=1))=\beta_0$ (i.e. no dependence on previous responses or covariates). We generated data to explore the bias and efficiency of the sampling/estimation methods with different covariate structures and when the outcome model is misspecified. The following simulations were performed.

Simulation 1 (Results in Table 1). The covariate x_{ij} is independent across time and subject and is unrelated to the lag term. Two scenarios are considered for the covariate x. One is with the x's generated with $x_{ij}\sim binomial(1,.5)$ and one is with the x's generated with $x_{ij}\sim .5+N(0,1)$. Outcome data were generated according to equation (2) with $\beta_1=.5,\beta_2=.2$ and $\beta_3=.2$.

Table 2. Medians(true), median absolute differences (MAD) of parameter estimates and median of estimated asymptotic standard errors(SE). Outcome data is generated according to $logitP(Y_{ij}=1|Y_{ij-1},X_{ij})=\beta_0+\beta_1X_{ij}+\beta_2Y_{ij-1}+\beta_3X_{ij}Y_{ij-1}$. *The covariates are correlated and are generated according to Simulation 2 and 3. With results in the first 4 columns from the model with* $x_{ij}=.5+b_i+\epsilon_{ij}$, $b_i\sim N(0,4)$ *and* $\epsilon_{ij}\sim N(0,1)$. *Results in the last four columns from the model with* $x_{ij}=.5+b_i+\epsilon_{ij}+y_{ij-1}$ *where* $b_i\sim N(0,4)$ *and* $\epsilon_{ij}\sim N(0,1)$. *For the top half of the table the marginal probability of a positive response,* $P(Y^*)$ *is .30 and .40 for the correlation structures of Simulations 2 and 3 respectively. For the bottom half* $P(Y^*)$ *is .043 and .058 for correlation structures of Simulations 2 and 3 respectively. The mean number of observations sampled for each simulation is 320 for* $\beta_0=-1.5$ *for both correlation structures and 241 and 278 for* $\beta_0=-4.0$ *with correlation structures of Simulations 2 and 3 respectively*

	Correlation within x's				Correlation with x's and with lag			
	NOD MLE	strat WMLE	strat MLE	DOD CMLE	NOD MLE	strat WMLE	strat MLE	DOD CMLE
med $\beta_0(-1.5)$	−1.501	−1.497	−0.430	−1.262	−1.501	−1.497	−0.332	−1.022
MAD	0.098	0.078	0.059	0.103	0.105	0.068	0.052	0.094
med asy SE	0.098	0.078	0.100	0.110	0.106	0.069	0.098	0.110
med $\beta_1(.5)$	0.502	0.510	0.393	0.391	0.502	0.513	0.334	0.263
MAD	0.051	0.074	0.043	0.050	0.060	0.083	0.047	0.067
med asy SE	0.050	0.064	0.049	0.051	0.059	0.070	0.054	0.066
med $\beta_2(.2)$	0.187	0.187	−0.628	−0.028	0.198	0.172	−0.950	−0.043
MAD	0.214	0.142	0.115	0.183	0.256	0.223	0.146	0.208
med asy SE	0.214	0.139	0.173	0.193	0.254	0.203	0.207	0.227
med $\beta_3(.2)$	0.203	0.197	0.160	0.238	0.200	0.195	0.138	0.255
MAD	0.095	0.089	0.065	0.082	0.096	0.100	0.067	0.093
med asy SE	0.095	0.084	0.075	0.086	0.096	0.091	0.080	0.095
med $\beta_0(-4.0)$	−4.017	−4.037	−1.200	−3.665	−4.011	−4.064	−1.062	−2.672
MAD	0.281	0.240	0.117	0.175	0.271	0.282	0.119	0.195
med asy SE	0.283	0.218	0.168	0.192	0.270	0.236	0.175	0.211
med$\beta_1(.5)$	0.504	0.557	0.302	0.296	0.503	0.569	0.283	−0.104
MAD	0.089	0.172	0.037	0.066	0.091	0.192	0.038	0.062
med asy SE	0.091	0.121	0.045	0.060	0.088	0.134	0.048	0.063
med $\beta_2(.2)$	0.039	0.202	−1.222	−0.104	0.126	0.236	−1.446	−0.957
MAD	1.373	0.419	0.277	0.510	1.101	0.459	0.289	0.480
med asy SE	1.313	0.382	0.342	0.527	1.092	0.428	0.374	0.504
med $\beta_3(.2)$	0.234	0.157	0.259	0.387	0.220	0.140	0.226	0.752
MAD	0.344	0.195	0.073	0.134	0.242	0.206	0.064	0.114
med asy SE	0.327	0.152	0.081	0.134	0.234	0.156	0.079	0.116

Simulation 2 (Results in Table 2). The covariate x_{ij} is correlated across time within a subject so $x_{ij}=.5+b_i+\epsilon_{ij}$ where $b_i\sim N(0,4)$ and $\epsilon_{ij}\sim N(0,1)$. With these parameters, ρ, the correlations between the x's is $\rho=\frac{\sigma_b^2}{(\sigma_b^2+\sigma_\epsilon^2)}=.8$ where $\sigma_b^2=4$ and $\sigma_\epsilon^2=1$. Outcome data were generated according to equation (2) with $\beta_1=.5,\beta_2=.2$ and $\beta_3=.2$.

Simulation 3 (Results in Table 2). The covariate x_{ij} depends on the lag term and there is correlation in $\boldsymbol{x}_i$. The data is generated with $x_{ij}=.5+b_i+\epsilon_{ij}+y_{ij-1}$ where $b_i\sim N(0,4)$ and $\epsilon_{ij}\sim N(0,1)$. Outcome data were generated according to equation (2) with $\beta_1=.5,\beta_2=.2$ and $\beta_3=.2$.

Simulation 4 (Results in Table 3). The analysis model is misspecified. Data were generated according to a model with 2 lag terms and binary x_{ij}. The model is logit$(Prob(Y_{ij}=1|Y_{ij-1},Y_{ij-2},X_{ij})=\beta_0+\beta_1x_{ij}+\beta_2y_{ij-1}+\beta_3x_{ij}y_{ij-1}+\beta_4y_{ij-2}$ with $x_{ij}\sim binomial(1,.5)$ with $\beta_1=.5,\beta_2=.2,\beta_3=.2and\beta_4=.2$. Model (2) is then fit to the data.

In some replications of the simulations not all strata had at least ten subjects. To accommodate this, the stratified sampling methods use all subjects in those strata and the total number of subjects in the sample is reduced. Therefore, in order to maintain the same sample size across methods the sample size from the stratified methods is used for all methods. Because some replications in the simulations have strata with less than 10 subjects the total sample size varies from replication to replication.

When presenting the simulation results, we provide the median estimate for each parameter, the median absolute deviations (MAD) of the simulated parameter estimates and the median of the asymptotic estimates of the standard errors. The medians and median absolute deviations are provided since sometimes there are large outliers that heavily

Table 3. Medians(true), median absolute differences (MAD) of parameter estimates and median of estimated asymptotic standard errors (SE). Outcome data is generated according to $logit P(Y_{ij} = 1|Y_{ij-1}, X_{ij}) = \beta_0 + \beta_1 X_{ij} + \beta_2 Y_{ij-1} + \beta_3 X_{ij} Y_{ij-1} + \beta_4 Y_{ij-2}$ *and* $x_{ij} \sim binomial(1, .5)$. *The estimation model is misspecified and is given in equation 2. For the top half of the table the marginal probability of a positive response,* $P(Y^*)$ *is .247 and the average sample size is 320. For the bottom half* $P(Y^*)$ *is .024 and the mean number of observations sampled was 123*

	NOD MLE	strat WMLE	strat MLE	DOD CMLE
med $\beta_0(-1.5)$	−1.501	−1.497	−0.430	−1.262
MAD	0.098	0.078	0.059	0.103
med asy SE	0.098	0.078	0.100	0.110
med $\beta_1(.5)$	0.502	0.510	0.393	0.391
MAD	0.051	0.074	0.043	0.050
med asy SE	0.050	0.064	0.049	0.051
med $\beta_2(.2)$	0.187	0.187	−0.628	−0.028
MAD	0.214	0.142	0.115	0.183
med asy SE	0.214	0.139	0.173	0.193
med $\beta_3(.2)$	0.203	0.197	0.160	0.238
MAD	0.095	0.089	0.065	0.082
med asy SE	0.095	0.084	0.075	0.086
med $\beta_0(-4.0)$	−4.017	−4.037	−1.200	−3.665
MAD	0.281	0.240	0.117	0.175
med asy SE	0.283	0.218	0.168	0.192
med $\beta_1(.5)$	0.504	0.557	0.302	0.296
MAD	0.089	0.172	0.037	0.066
med asy SE	0.091	0.121	0.045	0.060
med $\beta_2(.2)$	0.039	0.202	−1.222	−0.104
MAD	1.373	0.419	0.277	0.510
med asy SE	1.313	0.382	0.342	0.527
med $\beta_3(.2)$	0.234	0.157	0.259	0.387
MAD	0.344	0.195	0.073	0.134
med asy SE	0.327	0.152	0.081	0.134

influence the mean and standard errors. All model fits converge in all simulations.

We were also interested in investigating whether there were general statements that could be made about optimal designs. To this end we considered "asymptotic" efficiency of parameter estimates under the various sampling and estimation procedures with a data set approximating asymptotic conditions. To implement this, we generated a data set with 1,000,000 subjects and 5 time points per subject. We generated data according to equation 2 with $x_{ij} \sim binomial(1, .5)$. The parameter value for β_0 was varied from −4 to 0 at intervals of .1 and the parameter value for β_1 was varied from 0 to .5 at intervals of .02. The parameter value for β_2 and β_3 were held constant at .2. For each set of parameter values the x_{ij} values were the same and the same random variates were used to generate the outcome variables. Once a data set was generated, weights were calculated from the large data set, 1,000 patients per stratum were sampled for the stratified methods and 32,000 observations were sampled in the NOD and DOD methods. We then created contour plots of the ratio of the variance of the estimates for all pairs of methods for β_0 and β_1. Note, since the data sets are very large all estimates of β_1 are unbiased. Estimates of β_0 are unbiased for all methods except strat/MLE.

4. SIMULATION RESULTS

For Simulation 1 where the covariate x_{ij} are independent across time and unrelated to the lag term, Table 1 shows that the parameter estimates associated with x (β_1 and β_3) are unbiased for all sampling and estimation methods when the marginal probability of a positive response (defined as $P(Y^*)$) is largest. But, as expected β_0 and β_2 are biased for the strat/MLE method with the other methods providing unbiased estimates of these parameters. When $P(Y^*)$ is small the estimates of β_2 and β_3 from the NOD/MLE and DOD/CMLE methods are biased. The estimates from the strat/WMLE are all unbiased while the estimates from the strat/MLE are biased for β_0 and β_2 and unbiased for β_1 and β_3. The results are the same for continuous or binary x_{ij}. We only discuss efficiency of estimates that are unbiased. For β_1 and β_3 the strat/MLE either provides similar MAD or smaller MAD than the other three methods. For β_0 and β_2, the strat/WMLE method either provides similar or smaller standard errors than the other methods.

For all methods, the MAD estimates of the variability in the simulated parameter estimates is either smaller or very close to the median of the asymptotic estimates of the standard errors when the prevalence (probability of a positive response) is largest. When the prevalence is low and the x_{ij} covariate is binary the asymptotic estimates can underestimate the variability of the parameter estimates while for continuous x_{ij} the asymptotic standard error estimates and the Monte-carlo estimates are close. This result suggests that asymptotic standard errors perform well even when the prevalence is low for continuous covariates.

In Simulations 2 and 3, we study the effects of correlation on the bias and efficiency of the sampling/estimation methods. In Table 2, the first four columns of the numerical section of the table show results with correlation among the x_{ij} but independent of the lag term (Simulation 2) while in the last four columns the x_{ij} are correlated within a person and also correlated with the lag term (Simulation 3). The strat/MLE and DOD/CMLE methods estimates are all biased for each of the four scenarios in Table 2 so they will not be discussed further when discussing this table. When the marginal probability of a positive response ($P(Y^*)$) is largest .30 and .40 for Simulations 2 and 3 respectively the parameter estimates for the NOD/MLE and strat/WMLE both give unbiased estimates. The variability in the estimates is similar for both methods. When the marginal probability is small (.043 and

.058) the estimates for β_0 and β_1 are unbiased with the variability larger for β_1 with the strat/WLME. β_2 is biased for the NOD/MLE method but unbiased for the strat/WMLE method and β_3 is biased for both methods. This suggests that the strat/MLE and DOD/CMLE are very sensitive to the correlation structure in the covariates, while the strat/WMLE and NOD/MLE are not sensitive unless the prevalence is very low, with the strat/WMLE less sensitive than the NOD/MLE for low prevalence. The MAD Monte-Carlo standard error is very close to the median of the asymptotic estimates of the standard errors for all parameters and scenarios.

We also examine the effect on the parameters when the estimation model is misspecified according to the generation model (Simulation 4). Table 3 shows that the strat/MLE method and the DOD/CMLE method are biased so they will not be discussed further when discussing this table. When the marginal probability of a positive response ($P(Y^*)$) is largest (.247) the parameter estimates for the NOD/MLE and strat/WMLE both give unbiased estimates. The variability in the estimates is similar for both methods with the exception of β_2 where the variability is smaller for NOD/MLE. When the marginal probability is small (.024) the estimates for β_0 and β_1 are unbiased with the variability larger for β_1 with the strat/WLME. β_2 is biased for the NOD/MLE method but unbiased for the strat/WMLE method and β_3 is biased for both methods. The MAD monte-Carlo standard error is very close to the median of the asymptotic estimates of the standard errors for all parameters and scenarios.

This suggests that asymptotic standard errors perform well even when the model is misspecified. We compute asymptotic relative efficiencies over a large range of values of β_0 and β_1 to make general design statements. The results are based on the correct model specification.

Figure 1 gives the contour plots of the ratio of the approximate asymptotic variances for the different sampling/estimation methods (obtained with the generation of a very large data set as described in section 3). The first column shows the ratio of the variance estimates for β_0 for the three unbiased methods. The last two columns show the ratio of the variance estimates for β_1 for all the methods. The figure shows that the strat/WMLE method is more efficient than the NOD/MLE method and the DOD/CMLE method for all combinations of β_0 and β_1 studied. The efficiency comparisons depend on the prevalence (β_0) and little on β_1. For β_1 the strat/WMLE method is the most efficient.

5. EXAMPLE

We analyzed data from the Longitudinal Study of Vaginal Flora. The original intent was to compare stored vaginal wash samples of vaginal microflora with the Nugent Gram stain method of diagnosing BV. The vaginal wash molecular amplification data has not been collected. Therefore, in this paper we compare the clinical diagnosis of BV to the Gram stain method. In particular, we examine the following questions: Does a positive Gram stain increase the probability of a positive clinical diagnosis of BV? Does a positive clinical diagnosis of BV increase the odds of a subsequent positive clinical diagnosis of BV?

In total 1,710 women had both measurements at all 5 time points. Although for simplicity we only include observations with measurements at all time points, the methods could be applied with missing data if the data were missing at random. For an analysis with missing covariate information the data would be stratified on the outcome measurements. Missing covariate data would be imputed and the strat/WMLE method could be used.

This example is helpful to understand the proposed methods since clinical diagnosis of BV and a Gram stain were obtained on all subjects. The full data set can be used to obtain the best estimates of the parameters and then the sampling/estimation method results can be compared to the full data set estimates. For the sampling methods, we assume the clinical BV measurements were obtained on all women and we will sample (either randomly or based on outcome) to determine which women's Gram stain measurements will be included. Rather than performing the sampling a single time, bootstrap parameter estimates will be provided (details of the bootstrap method are provided below).

The full data set is analyzed to choose the terms in the Markov model that will be used in the sampling analysis. Only Markov models with two or fewer lag terms will be considered since there are only five longitudinal time points. The estimation routine was highly dependent on starting values and converged to local maxima when two-way interactions were considered, so only the first-degree interactions between the Gram stain variable and the lag terms were considered. The most flexible model we considered was

$$\begin{aligned}\text{logit}(P(Y_{ij}=1|Y_{ij-2},Y_{ij-1},X_{ij})) &= \beta_0+\beta_1X_{ij}+\\ &\beta_2Y_{i,j-1}+\beta_3Y_{i,j-2}+\beta_4X_{i,j-1}Y_{i,j-1}+\\ &\beta_5X_{i,j-2}Y_{i,j-2}+\beta_6Y_{i,j-1}Y_{i,j-2}\end{aligned}$$

where the Y_{ij} are the clinical BV response data at the j^{th} time point and the X_{ij} are the Gram stain BV response data at the j^{th} time point for women i. In this model each of the interaction terms were not individually significant and a simultaneous test of all interaction terms was also not significant (all testing was performed using a likelihood ratio test). A reduced model with the first four terms $(\beta_0,\beta_1,\beta_2,\beta_3)$ was then examined. In this model each of the lag terms was significant so they were included in the final model for the bootstrap analysis.

In order to evaluate the performance of the sampling methods on this BV data set, the following bootstrap proce-

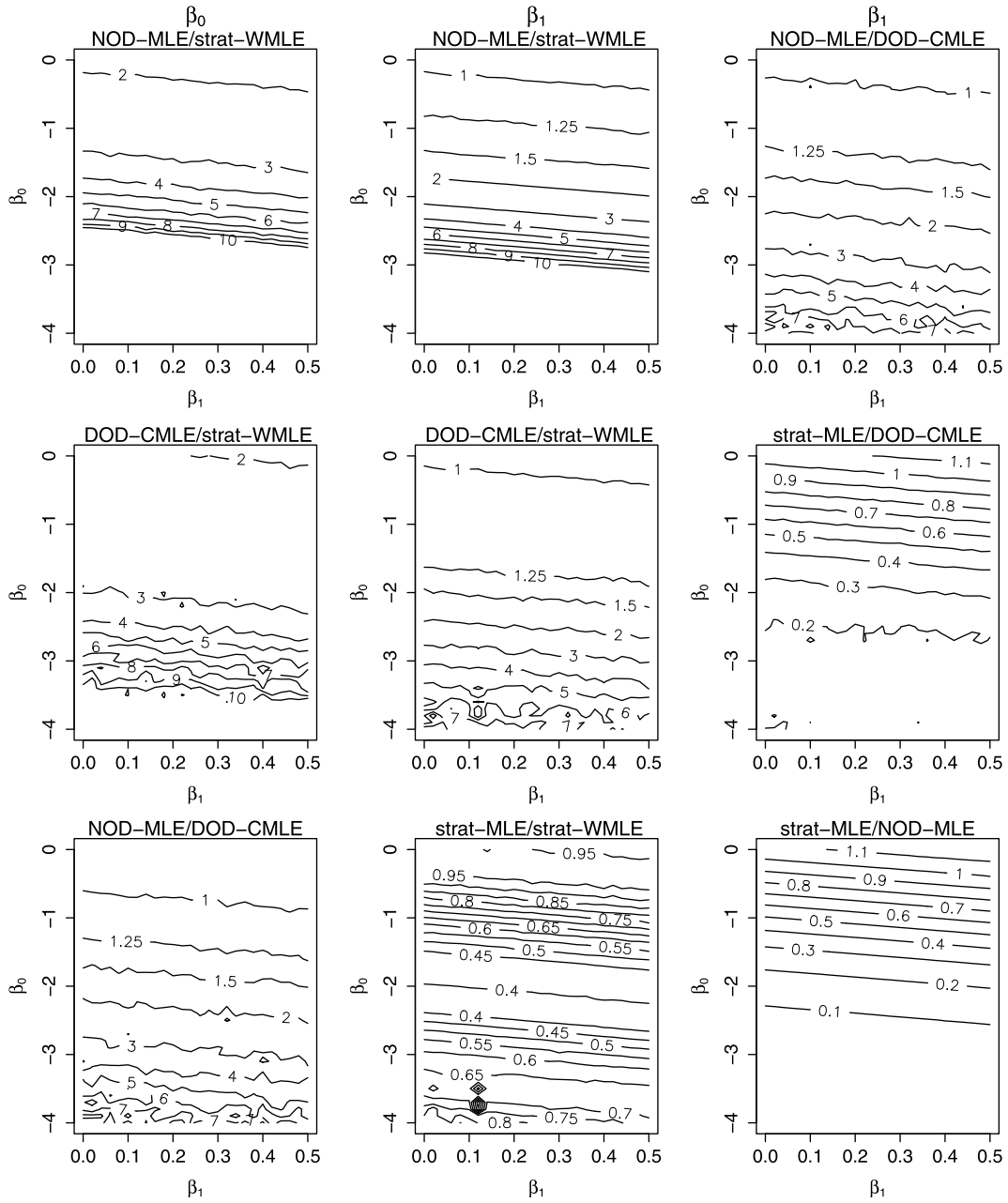

Figure 1. Contour plots of the ratio of the asymptotic variance estimates for the different sampling/estimation methods. The first column shows the ratio of the variance estimates for β_0 for the three unbiased methods. The last two columns show the ratio of the variance estimates for β_1 for the all methods.

dure was used. Weights were calculated based on the original sample. Repeated random sampling was performed using the three sampling methods. The samples were then analyzed using the previously discussed estimation methods with the model being a Markov regression model with two lag terms and no interactions.

If less than 10 observations were available in a stratum, all the observations were sampled and the total sample size was reduced (as in the simulation). Therefore, 129 subjects were sampled for all methods.

Table 4 gives the median of the parameter estimates and the median of the estimates of the asymptotic standard errors of the bootstrap parameter estimates along with the parameter estimates with all of the data in the full data set used. For the bootstrap estimates for each replication we perform a Wald test for testing $\beta_i = 0$. We then report the

Table 4. Parameter estimates from example when the full data set is analyzed and each of the 4 different sampling analysis methods are used. The full data set is resampled 10,000 times. The mean number of observations sampled was 129. Medians, median of estimated asymptotic standard errors and the proportion of times each bootstrap Wald test would reject the hypothesis of $\beta_i = 0$ testing at the 2-sided .05 level. The analysis model is $logitP(Y_{ij} = 1|Y_{ij-1}, X_{ij}) = \beta_0 + \beta_1 X_{ij} + \beta_2 Y_{ij-1} + \beta_3 Y_{ij-2}$

	Full MLE	NOD MLE	strat WMLE	strat MLE	DOD CMLE
β_0	−3.94	−3.995	−3.945	−0.700	−3.916
median asy SE	.129[a]	0.480	0.351	0.189	0.220
proportion rejections	–[b]	1.00	1.00	1.00	1.00
β_1	-.059	−0.077	−0.054	−0.204	0.005
median asy SE	.183[a]	0.705	0.402	0.238	0.284
proportion rejections	–[b]	0.015	0.0373	0.020	0.007
β_2	1.54	1.438	1.537	−0.493	1.616
median asy SE	.251[a]	0.958	0.312	0.256	0.384
proportion rejections	–[b]	0.372	1.000	0.185	0.989
β_3	1.92	1.878	1.921	−0.007	1.929
median asy SE	.205[a]	0.792	0.357	0.234	0.315
proportion rejections	–[b]	0.623	1.000	0.000	1.000

[a]Estimated asymptotic standard error using all data.
[b]The p-values for the test of $\beta_0 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$ for the full data set are all $< .001$. The p-value for $\beta_1 = 0$ is .374.

proportion of times the p-value is less than 0.025 (2-sided 0.05 significance level).

It can be seen from the table that the strat/WMLE provides less biased estimates than the NOD/MLE and DOD/CMLE method and gives smaller standard errors. It is not expected that the strat/MLE method would give an unbiased estimate of β_0, but the estimate of β_1 should be unbiased. In this data set, the strat/MLE estimates of β_1 are more biased than the other methods. The reason the NOD/MLE method performs so poorly is that $P(Y^*)$ for the data set is .044, which is very low, and results in the NOD/MLE having a very high proportion of subjects who have no positive observations. The strat/WMLE performs well since the overall sample size is quite large (1,710 subjects), giving good estimates of the weights that are used in the estimation method for the stratified sample.

The full data set analysis shows that a positive Gram stain diagnosis of BV does not increase the probability of observing a positive clinical diagnosis of BV(p-value for test of $\beta_1 = 0$ is 0.374). The other terms are all highly significantly different from zero (all p-values <.001) indicating that a positive BV clinical diagnosis in either of the two previous visits increases the probability of subsequent positive clinical diagnosis. The proportion of rejections for the strat/WMLE and DOD/CMLE methods agree with the full data set results in that when the p-value for the test of a parameter being equal to zero is very small, the proportion of rejections is close to one. For the test of β_1 the proportion of rejections at the two-sided .05 level is small and this is consistent with the full data set results.

6. DISCUSSION

We explored bias and efficiency of approaches to sampling for a longitudinal transition regression model. In practice, randomly sampling observations and using maximum likelihood estimation works well unless the overall probability of a positive response is very small. When this happens, the probability of sampling all zero measurements is high; thus, the parameter estimates have large variability and can be very biased.

Stratifying observations based on the patterns of response and sampling a specified number of observations can give unbiased estimates if a weighted MLE approach is used. In this paper we examine longitudinal data with five times points so the number of possible strata is not too large. In cases where there are many time points it would be difficult to sample over all possible strata. In this case, other types of stratification and sampling would need to be implemented. For example, stratification by types of transitions or number of responses could be performed. The weights associated with each strata could be calculated and the strat/WMLE's could be calculated. Further simulations to examine properties of the sampling schemes could be performed.

When the marginal probability of a response is not too small, the standard errors of the strat/WLME parameter estimates are larger than that of the NOD/MLE estimates. However, for small marginal response probabilities, the strat/WMLE estimates have smaller standard errors than NOD/MLE. This is illustrated in the example data, where the strat/WMLE method provided the most efficient

estimation of the relationship between gram staining and clinical assessments of BV.

The DOD/CMLE method performs similarly to the NOD/MLE method when the overall probability of a positive response is not too small. For the small prevalence scenarios studied here the DOD/CMLE method gave biased results.

In addition to conducting simulations and data analysis we compared the asymptotic variances across methods and found that the strat/WMLE had higher efficiency than other methods over a wide range of prevalences and effect sizes (i.e., values of β_0 and β_1). These results show the advantages of the proposed weighted approach when the study is large enough to estimate the weights precisely.

Even though the weights may have large variance in practical situations, we showed that the strat/WMLE approach does close to or better than the other approaches in many situations, and most dramatically when the prevalence is low. There are other advantages of this approach as well. In the situations we studied it was more robust to misspecification of the order of dependence in the transition model, specifically when the covariates were correlated across time or with lag terms.

In the simulations where the prevalence was small, even though all of the model fits converged, some estimates were very biased. This was due to the fact that the particular sample had very few events and so the log-likelihood function was very flat thus not allowing for good estimation of the parameters. This demonstrates the issue with random sampling when the prevalence is low. Once a random sample is taken, if no events are sampled, the data set will be uninformative.

Functions of transition probabilities such as first passage time use all estimates of parameters in the estimate. Therefore, it is important to estimate all parameters well. This is in contrast to studies where it is only of interest to test one covariate (say a treatment effect). As we have shown here when transition probabilities and functions of transition probabilities are of interest the strat/WMLE approach should be used.

APPENDIX A

A.1 Variance estimate for weighted maximum likelihood estimators

Let the weighted pseudo-likelihood estimating equations for the px1 vector of parameters, $\boldsymbol{\beta}$ be

$$U(\hat{\boldsymbol{\beta}}) = \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{m} w_{hij}\boldsymbol{x}_{hij}(y_{hij} - p_{\hat{hij}})$$

where $\boldsymbol{x}_{hij}$ is a px1 vector. Then

$$\frac{\partial U(\hat{\boldsymbol{\beta}})}{\partial w_{hij}} = \boldsymbol{x}_{hij}y_{hij} - \boldsymbol{x}_{hij}\hat{p}_{hij} - \sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{m} w_{hij}\boldsymbol{x}_{hij}\frac{\partial \hat{p}_{hij}}{\partial \boldsymbol{\beta}}' \frac{\partial \beta}{\partial w_{hij}} \tag{3}$$

$$\frac{\partial \hat{p}_{hij}}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_{hij}(1-\hat{p}_{hij})\hat{p}_{hij} \tag{4}$$

note the chain rule is used since $\hat{\boldsymbol{\beta}}$ is a function of w_{hij}. Now since $\frac{\partial U(\hat{\boldsymbol{\beta}})}{\partial w_{hij}} = 0$, and substituting 4 into 3 gives

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial w_{hij}} = \left[\sum_{h=1}^{H}\sum_{i=1}^{I_h}\sum_{j=1}^{m} w_{hij}\boldsymbol{x}_{hij}\boldsymbol{x}'_{hij}(1-\hat{p}_{hij})\hat{p}_{hij}\right]^{-1} \boldsymbol{x}_{hij}(y_{hij}-\hat{p}_{hij})$$

ACKNOWLEDGEMENTS

The authors would like to thank Mark Klebanoff and Barry Graubard for helpful discussions. We thank the Center for Information Technology, NIH, for providing access to the high performance computational capabilities of the Biowulf cluster computer system.

Received 6 February 2013

REFERENCES

[1] Amsel, R., Totten, P. A., Spiegel, C. A., Chen, K. C. S., Eschenbach, D. and Holmes, K. K. (1983). Nonspecific vaginitis. Diagnostic criteria and microbial and epidemiologic associations. *American Journal of Medicine* **74** 14–22.

[2] Binder, D. (1996). Taylor linearization for single phase and two phase samples: A cookbook approach. *Survey Methodology* 17–26.

[3] Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*. Chapman & Hall CRC. MR1014891

[4] Demnati, A. and Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology* **30** 4–13.

[5] Graubard, B. I., Rao, R. S. and Gastwirth, J. L. (2005). Using the Peters-Belsen method to measure health care disparities from complex survey data. *Statistics in Medicine* **24** 2659–2668. MR2196206

[6] Klebanoff, M., Schwebke, J. R., Zhang, J., Nansel, T. R., Yu, K.-F. and Andrews, W. W. (2004). Vulvovaginal symptoms in women with bacterial vaginosis. *Obstetrics and Gynecology* **104** 267–272.

[7] Korn, E. L., Graubard, B. I., Groves, R. M., Kalton, G., Rao, J. N. K., Schwartz, N. and Skinner, C. (1999). *Analysis of Health Surveys*. John Wiley & Sons.

[8] Koumans, E. H., Markowitz, L. E. and Hogan, V. (2002). Indications for therapy and treatment recommendations for bacterial vaginosis in nonpregnant and pregnant women: A synthesis of data. *Clinical Infectious Diseases* **35** S152–S172.

[9] Leibenluft, E., Albert, P. S., Rosenthal, N. E. and Wehr, T. A. (1996). Relationship between sleep and mood in patients with rapid-cycling bipolar disorder. *Psychiatry Research* **63** 161–168.

[10] Neuhaus, J. M. and Jewell, N. P. (1990). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics* **46** 977–990.

[11] Nugent, R. P., Krohn, M. A. and Hillier, S. L. (1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *Journal of Clinical Microbiology* **29** 297–301.

[12] Pirotta, M., Fethers, K. A. and Bradshaw, C. S. (2009). Bacterial vaginosis – More questions than answers. *Australian Family Physician* **38** 394–397.

[13] Schildcrout, J. S. and Heagerty, P. J. (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics* **9** 735–749.

[14] Shah, B. V. (2004). Comment on "Linerization variance estimators for survey data" by A. Demnati and J.N.K. Rao. *Survey Methodology* **30** 16.

[15] Ware, J. H. and Lipsitz, S. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine* **7** 95–107.

[16] Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44** 1019–1031. MR0980997

[17] Zhang, P. and Medoff-Cooper, B. (1996). A Markov regression model for nutritive sucking data. *Biometrics* **52** 112–124.

Sally Hunsberger
Biostatistics Research Branch
6130 Executive Blvd, rm 8120
Rockville, MD 20852
USA
E-mail address: sallyh@mail.nih.gov

Paul S. Albert
Biostatistics and Bioinformatics Branch
6130 Executive Blvd, rm 8120
Rockville, MD 20852
USA
E-mail address: albertp@mail.nih.gov

Marie Thoma
Epidemiology Branch
6100 Executive Blvd
Rockville, MD 20852
USA
E-mail address: thomame@mail.nih.gov