OPEN

Extending the Case–Control Design to Longitudinal Data: Stratified Sampling Based on Repeated Binary Outcomes

Jonathan S. Schildcrout,[a] *Enrique F. Schisterman,*[b] *Nathaniel D. Mercaldo,*[a] *Paul J. Rathouz,*[c] *and Patrick J. Heagerty*[d]

Abstract: We detail study design options that generalize case–control sampling when longitudinal outcome data are already collected as part of a primary cohort study, but new exposure data must be retrospectively processed for a secondary analysis. Furthermore, we assume that cost will limit the size of the subsample that can be evaluated. We describe a novel class of stratified outcome-dependent sampling designs for longitudinal binary response data where distinct strata are created for subjects who never, sometimes, and always experienced the event of interest during longitudinal follow-up. Individual designs within this class are differentiated by the stratum-specific sampling probabilities. We show for parameters associated with time-varying exposures, subjects who experience the event/outcome at some but not at all of the follow-up times (i.e., those who exhibit response variation) are highly informative. If the time-varying exposure varies exclusively within individuals (i.e., intraclass correlation coefficient is 0), then sampling all subjects with response variability can yield highly precise parameter estimates even when compared with an analysis of the original cohort. The flexibility of the designs and analysis procedures also permits estimation of parameters that correspond to time-fixed covariates, and we show that with an imputation-based estimation procedure, baseline covariate associations can be estimated with very high precision irrespective of the design. We demonstrate features of the designs and analysis procedures via a plasmode simulation using data from the Lung Health Study.

(*Epidemiology* 2018;29: 67–75)

We describe a novel class of study designs and associated analysis procedures for longitudinal binary response data.[1–3] The designs can be used in settings where the response and basic covariate data are available, but an expensive exposure variable must be ascertained retrospectively in order to conduct a new study. To generalize case–control sampling, we propose a stratified sampling based on three strata (0, 1, and 2) empirically defined by summarizing each subject's response over time into a simple total count. With Y_{ij} denoting the binary response value for subject i at observation number j, and n_i the number of times subject i was observed, we define stratum 0, 1, and 2 with $\sum_j Y_{ij} = 0$, $0 < \sum_j Y_{ij} < n_i$, and $\sum_j Y_{ij} = n_i$, respectively. Within this class of stratified sampling designs, we assign members of a stratum (r) a common sampling probability, $\pi(r)$, $r = 0, 1, or\ 2$, and alternative designs are distinguished by the choices for the stratum-specific sampling probabilities: $\pi(0)$, $\pi(1)$, and $\pi(2)$.

An important feature of longitudinal data arises from the fact that both response and exposure variability may result from within-subject changes over time, or from subject-to-subject variability. In this article, we will show that *who* is informative (i.e., which strata should be sampled with high probability) depends upon the response-exposure associations that are of interest. We emphasize the importance of considering the empirical sources of exposure and response variability (between-subject and within-subject) when conducting the designs.

To demonstrate the utility of the designs, we use data collected at annual clinic visits from the Lung Health Study.[4–6] The Lung Health Study protocols were approved by the institutional review boards at each participating clinical center, and participants were enrolled after written informed consent was obtained. Based on genetic associations observed with chronic

Editor's Note: A Commentary on this article appears on p.76.

Submitted July 20, 2016; accepted September 27, 2017.

From the [a]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN; [b]Epidemiology Branch, Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD; [c]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI; and [d]Department of Biostatistics, University of Washington, Seattle, WA.

Supported, in part, by the NIH grants R01 HL094786 from the National Heart Lung and Blood Institute, the NIH grant K07 CA172294 from the National Cancer Institute, the Long-Range Research Initiative of the American Chemistry Council, and the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health.

Disclosure: The authors have no conflicts of interest. P.J.R. is a Charter Member of a Data Safety Monitoring Board for Sunovian Pharmaceuticals, Inc., in Fort Lee, NJ. Sunovian is a pharmaceutical and drug development company.

Data for the analyses conducted here can be downloaded from the database for genotypes and phenotypes (dbGaP). The code for conducting analyses is available from the online electronic appendix; http://links.lww.com/EDE/B286 and http://links.lww.com/EDE/B287.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: jonathan.schildcrout@vanderbilt.edu Department of Biostatistics, Vanderbilt University Medical Center, 2525 West End Ave, Suite 11000, Nashville, Tennessee 37203.

ISSN: 1044-3983/18/2901-0067
DOI: 10.1097/EDE.0000000000000764

bronchitis and chronic obstructive pulmonary disease in smokers,[7] the exemplar analysis will examine the short term, causal effect of current smoking on chronic bronchitis risk in those with and without at least one T allele on single nucleotide polymorphism rs7671167, and the difference between the two effects. Importantly, it will also study the extent to which other, baseline characteristics are risk factors for chronic bronchitis.

The Lung Health Study collected genetic, covariate, and outcome data for all participants and made them available on the National Center for Biotechnology Information database of genotypes and phenotypes (dbGaP[8]; access number phs000335.v2.p2). Links to instruction pages for downloading dbGaP data are provided in the eAppendix; http://links.lww.com/EDE/B284. We use a subset of Lung Health Study data to conduct a plasmode simulation study[9] that examines the utility of the proposed retrospective study designs under scenarios where the longitudinal outcome and covariate data are available on all subjects but genetic data must be ascertained retrospectively at a sample size limiting cost.

ESTIMANDS AND ESTIMATORS FOR THE LUNG HEALTH STUDY ANALYSIS

The population represented by Lung Health Study[5] participants is middle-aged smokers with mild-to-moderate COPD. Although the Lung Health Study was a randomized clinical trial that examined smoking cessation interventions, treatment allocation was not available to us. For this analysis, we seek to emulate a hypothetical sequentially randomized study where participants are randomized to be a smoker or nonsmoker at different longitudinal follow-up visits. We, therefore, conduct a multiple logistic regression analysis that adjusts or controls for potential confounders of the smoking–chronic bronchitis relationship and other baseline characteristics that are also of scientific interest for their association with chronic bronchitis. Our primary goal with these data is to illustrate possible novel sampling designs that could be used to efficiently evaluate a new candidate biomarker that may predict chronic bronchitis or may modify the effect of smoking.

Let $\mu_{ij} = Pr[CB_{ij} = 1 | \mathbf{X}_{ij}]$ be the prevalence of chronic bronchitis in a population defined by subject i's covariate values $\mathbf{X}_{ij}$ at the j^{th} visit, and let smk_{ij} be a 0/1 indicator of current smoking. In the analysis dataset, smk_{ij} varies both within and between individuals (intracluster correlation coefficient = 0.82), and for this reason, we adopt the "between–within" model described in Neuhaus and Kalbfleisch[10] to partition exposure variability. The between–within model decomposes smk_{ij} into two components, one that varies exclusively within and the other that varies exclusively between individuals. Specifically, let $smk_{b,i} = \overline{smk}_i = \sum_j smk_{ij} / n_i$, then we can write $smk_{ij} = (smk_{ij} - smk_{b,i}) + smk_{b,i} = smk_{w,ij} + smk_{b,i}$. Thus, $ICC(smk_{w,ij}) = 0$ and $ICC(smk_{b,i}) = 1$, and the two components can be modeled separately. For the Lung Health Study analysis, the analytical model is given by:

$$logit(\mu_{ij}^{lhs}) = \beta_0 + \beta_w smk_{w,ij} + \beta_s snp_i + \beta_{ws} smk_{w,ij} \cdot snp_i + \beta_b smk_{b,i} + \beta_c c_{ij} \quad (1)$$

where $snp_i = 1(0)$ if the T allele is present (absent) at rs7671167, and c_{ij} is a vector of confounders and other risk factors that includes baseline covariates lifetime pack–years smoked, average number of cigarettes smoked per day before the Lung Health Study, gender, age, body mass index, and nine site-specific indicator variables; and the time-varying covariate, study year. Notice that β_{ws} corresponds to a difference in the log odds ratio for smokers (versus nonsmokers) between those with and without the T allele. Because the estimate was very close to 0, we excluded the interaction between snp_i and $smk_{b,i}$ in this between–within model.

Although the between–within model is not a standard data-generating, causal model,[11,12] it is popular in longitudinal analyses and has been shown to be useful for estimating causal effects under specific assumptions. For example, consider a simplified version of (1), and assume the true model is given by $logit(\mu_{ij}) = \beta_0 + \alpha \cdot smk_{ij} + \beta_u u_i \equiv \beta_0 + \alpha \cdot smk_{w,ij} + \alpha \cdot smk_{b,i} + \beta_u u_i$ so that $\beta_w = \beta_b$. Also assume that the subject level confounder u_i is unmeasured and the model fit to the data is $logit(\mu_{ij}^{bw}) = \beta_{bw,0} + \beta_w smk_{w,ij} + \beta_b smk_{b,i}$. The estimate of β_w is consistent for α as long as u_i has a linear relationship with $smk_{b,i}$.[13] Fitting the standard model $logit(\mu_{ij}^*) = \beta_0^* + \alpha^* \cdot smk_{ij}$ results in an inconsistent estimate of α. This estimate of β_w has exhibited robustness to the "linear with the average" criterion[12] although it has shown sensitivity to high levels of heteroscedasticity[14] and can potentially induce confounding due to conditioning on a collider.[15,16] For the purpose of providing insights into the operating characteristics of the study designs, we assume the between–within model assumptions are not violated in a severe way. A key benefit of the model is the explicit decomposition of smk_{ij} into $smk_{w,ij}$ and $smk_{b,i}$, which vary exclusively within-subjects ($ICC = 0$) and between-subjects ($ICC = 1$), respectively. Design considerations for between- and within-subject covariate effects will be shown to be quite different.

Lung Health Study Cohort

Table 1 shows demographic and other characteristics of 3,932 participants. Seventy-six percentage of participants possessed at least one T allele at rs7671167. The vast majority were observed at four follow-up visits, and at each visit, between 71 and 76 percentage were smokers. Chronic bronchitis occurred in 28–29% of visits. In all, 2,064 participants never experienced chronic bronchitis, 1,404 experienced it at some but not all visits, and 464 experienced chronic bronchitis at all visits.

TABLE 1. Demographics and Features of the Lung Health Study Cohort at Baseline and Over the Course of Four Annual Visits to Study Clinics: Continuous Variable Are Summarized with the [5: 25: 50: 75: 95]th Percentiles, and Categorical Variables Are Summarized with Proportions or Raw Numbers Observed

Variables	Patient Level	Year 0	Year 1	Year 2	Year 3
No. patients	3,932				
Patient-level summaries					
Age (y)	[37: 43: 49: 54: 58]				
BMI (kg/m^2)	[20: 23: 25: 28: 33]				
Pack years	[17: 28: 37: 50: 76]				
Cigarettes/d	[10: 20: 30: 40: 55]				
Proportion female	0.37				
Proportion with ≥ 1 T allele	0.76				
Number smoking					
At no visits	750				
At some visits	623				
At all visits	2,559				
Number with chronic bronchitis					
At no visits (R = 0)	2,064				
At some visits (R = 1)	1,404				
At all visits (R = 2)	464				
Longitudinal summaries					
Number observed		3,932	3,859	3,845	3,930
Proportion smoking		0.76	0.74	0.72	0.71
Proportion with chronic bronchitis		0.29	0.28	0.28	0.28

Cigarettes per day captures the number of cigarettes per day before enrolling in the Lung Health Study.

DESIGN

To formalize the design options, assume a representative cohort of N participants with longitudinal binary outcomes ($\{\mathbf{Y}_i\}_{i=1}^N$, $\mathbf{Y}_i = \{Y_{i1},..,Y_{in_i}\}$) and *o*bserved covariate ($\{\mathbf{X}_{oi}\}_{i=1}^N$) data. In the Lung Health Study, Y_{ij} is the binary indicator for presence/absence of chronic bronchitis for subject i at visit j, and n_i is the number of times that subject i was observed. Assume additional measurement is needed to obtain a new *e*xposure variable $\{X_{ei}\}_{i=1}^N$ (e.g., snp_i) and ascertainment costs limit sample size. Thus, for subject i, $\mathbf{X}_i = (X_{ei}, \mathbf{X}_{oi})$ represents the complete design matrix, but X_{ei} is not available initially. Finally, $\mathbf{X}_{ij}$ is the covariate vector for subject i at visit j.

We proposed[1,2] stratified outcome-dependent sampling designs to identify subjects for retrospective X_{ei} ascertainment. Strata are defined by the subject-specific response sum, $\sum_j y_{ij}$. Stratum 0, 1, and 2 is comprised of subjects who had events at none, some, and all follow-up visits, respectively. Within stratum $R = r$, there are N_r subjects and each is sampled with probability $\pi(r)$, which is a design parameter under the control of the investigators and controls the magnitude of possible oversampling within select strata. The sampling probabilities result in an expected sample size from stratum r equal to $N_r^s = N_r \cdot \pi(r)$. We denote unique candidate designs with $D[N_0^s, N_1^s, N_2^s]$, indicating the size of the subsample taken from each stratum.

For the estimation of time-varying covariate regression parameters, including interactions with time-varying covariates, we have previously shown that subjects with any outcome variability (i.e., those belonging to stratum $R = 1$) are more informative than those who have constant responses over time.[1,2] The information difference between subjects with and without variation in their longitudinal outcomes is particularly large if the time-varying covariates vary exclusively within subjects (i.e., intraclass correlation coefficient equals 0) where subjects in strata 0 and 2 provide little to no information. In our example data, this is the case for the within-subject smoking variable where the intraclass correlation coefficient was, by construction, equal to 0. To the extent that there exists between subject variability in a covariate that is of interest, subjects in strata 0 and 2 may provide information regarding the exposure–outcome association.

MODEL

Although the general design ideas apply to a broad class of models, we focus on "marginalized models"[17–19] for binary response data. Marginalized models (like generalized linear mixed models[20,21]) can be estimated with likelihood-based approaches. However, in contrast to conditional generalized linear mixed models, marginalized models estimate marginal

model parameters similar to generalized estimating equations.[22,23] Given $\mathbf{Y}_i$ and $\mathbf{X}_i$, the marginal mean model is

$$\text{logit}(\mu_{ij}^m) \equiv \text{logit}\{E(Y_{ij} \mid \mathbf{X}_{ij})\} = \mathbf{X}_{ij}\beta^m, \quad (2)$$

where β^m is a $p-$dimensional parameter vector. While this model specifies the (marginal) mean of the distribution $[\mathbf{Y}_i \mid \mathbf{X}_i]$, to fully identify $[\mathbf{Y}_i \mid \mathbf{X}_i]$, we construct a second regression model to characterize within-subject response association. We refer to the second regression component as the conditional mean or response dependence model. In this article, we use the marginalized transition and latent variable model[24] where the response dependence model is

$$\text{logit}(\mu_{ij}^c) = \Delta_{ij} + \gamma Y_{ij-1} + b_i, \text{ and } b_i \sim N(0, \sigma^2). \quad (3)$$

The value Δ_{ij} can be thought of as an intercept in a logistic regression model that includes a transition or Markov term Y_{ij-1} and a random intercept b_i. This response dependence model is appealing because with longitudinal data, it is common to observe both serial response dependence that decays with time separations, and long-range dependence that is captured with the random intercept. A more thorough description of marginalized models and specifically Δ_{ij} can be found in the studies by Heagerty,[18] Schildcrout and Heagerty,[24] and Heagerty and Zeger.[25]

ESTIMATION

Within the proposed class of outcome-dependent sampling designs, we deliberately oversample informative subjects, resulting in a nonrepresentative study sample. For example, in many cases the sample will be overrepresented with members of stratum 1 in order to maximize efficiency. Therefore, to validly estimate regression parameters based on the targeted sample, we consider three estimation strategies: ascertainment-corrected maximum likelihood, weighted likelihood, and multiple imputation, which we briefly detail.

Ascertainment-Corrected Maximum Likelihood

Let $\theta = \{\beta^m, \alpha\}$, where $\alpha \equiv (\gamma, \sigma)$, be the parameter vector from a marginalized model and let $L(\theta \mid \mathbf{Y}_i, \mathbf{X}_i) = L_i = Pr(\mathbf{Y}_i \mid \mathbf{X}_i; \theta)$ be the likelihood contribution by subject i under random sampling. An ascertainment-corrected likelihood under the proposed design can be calculated via Bayes' theorem with

$$L_i^c(\theta \mid \mathbf{Y}_i, \mathbf{X}_i) = pr(\mathbf{Y}_i \mid \mathbf{X}_i, S_i = 1; \theta) = \frac{\pi(r_i) \cdot L_i}{\pi(0) \cdot L_{i0} + \pi(2) \cdot L_{i2} + \pi(1) \cdot [1 - L_{i2} - L_{i0}]} \quad (4)$$

where L_{i2} and L_{i0} are the likelihood contributions by subject i under random sampling had $\sum_j y_{ij} = n_i$ and $\sum_j y_{ij} = 0$, respectively, $\pi(r)$ is the probability of being sampled for all members of stratum R = r, $S_i = 1$ if subject i is sampled, and $S_i = 0$ if not. Ascertainment-corrected maximum likelihood maximizes (4), and we are able to estimate parameters in the original model, that is, θ in $Pr(\mathbf{Y}_i \mid \mathbf{X}_i; \theta)$. Notice that this ascertainment-corrected likelihood is similar to the conditional logistic regression likelihood with the exception that we condition on being sampled ($S_i = 1$) rather than on $\sum_j y_{ij}$. While conditional logistic regression is useful in many settings, it cannot estimate baseline covariate effects that are of secondary interest in the Lung Health Study analysis. Further, conditional logistic regression implicitly assumes a random intercept structure (approximately exchangeable correlation), which may be overly simplistic for longitudinal data that often possess serial dependence.

Weighted Likelihood

One of the most common approaches to addressing selection bias induced by design-based oversampling is inverse probability weighting.[26,27] Inverse probability weighting can be used to correct the standard likelihood score equations (i.e., the derivative of the log likelihood) by weighting by the inverse of the subject-specific probability of having been sampled, $\pi(r_i)$, which must equal one of $\pi(0)$, $\pi(1)$, or $\pi(2)$. Parameters are estimated by solving the weighted log-likelihood score equation, $\sum_i [\pi(r_i)^{-1} \cdot \partial \log(L_i) / \partial \theta] = \mathbf{0}$, and robust standard errors[22,26,27] are used to estimate uncertainty.

Multiple Imputation

A weakness of ascertainment-corrected maximum likelihood and weighted likelihood is that even though $(\mathbf{Y}, \mathbf{X}_o)$ are available on all subjects, the only information used in parameter estimation and inference is from subjects selected into the subsample for whom X_{ei} is also available (i.e., those with $S_i = 1$). One of the primary strengths of multiple imputation[28,29] is its efficient use of all available data. This suggests that we may impute X_{ei} for subjects not selected for detailed exposure evaluation (i.e., those for whom $S_i = 0$). There are different approaches for building the required imputation model, and in the eAppendix, http://links.lww.com/EDE/B284, we detail the specific imputation strategy that we use for the present analyses.

RESULTS FROM ANALYSES OF THE LUNG HEALTH STUDY

We illustrate the potential advantages of an outcome-dependent sampling design over random sampling using the Lung Health Study cohort. We are interested in risk factors for chronic bronchitis, and in particular, the impact of current smoking for those with and without at least one copy of the T allele at rs7671167. Equation 1 shows the regression model. While the primary estimation targets are β_w, β_{ws}, and $\beta_w + \beta_{ws}$, we are also interested in other covariate associations with chronic bronchitis risk (i.e., those described by the parameter β_c).

Chronic bronchitis was captured on the Lung Health Study annual questionnaire, and $CB_{ij} = 1$ if subject i experienced chronic bronchitis (at least three months of a phlegm-producing cough) in the year preceding visit j. The smoking variable, smk_{ij}, was set to 1 if the participant tested positive on a cotinine test at the prior $(j-1)$ or present (j) visit. Otherwise, $smk_{ij} = 0$. We used this definition primarily to improve the likelihood that those with $smk_{ij} = 0$ were, in fact, nonsmokers. One disadvantage of this approach is that we removed the first annual follow-up visit from the analysis because all subjects smoked before the study.

Designs and Analysis Procedures

Even though the Lung Health Study genotyped all subjects, we are interested in circumstances where we may only genotype a subset. For illustration we assume that approximately 1,600 of the 3,932 subjects can be genotyped, and we investigate several targeted designs. With random sampling, we randomly selected 1,600 subjects to be included for genotyping. The other designs we considered were outcome-dependent sampling designs: $D[100, 1400, 100]$, $D[275, 1050, 275]$ and $D[450, 700, 450]$. We sampled individuals independently, and in light of the stratum sizes (**Table 1**), sampling probabilities for the three outcome-dependent sampling designs were [0.048, 0.997, 0.216], [0.133, 0.748, 0.593], and [0.218, 0.499, 0.970]. Importantly, $D[100, 1400, 100]$ sampled effectively all participants in stratum 1, and so it includes all subjects who exhibited response variation over the course of the study. Results from earlier work suggest that this design can yield high precision for time-varying covariate effect estimates (e.g., for β_w and β_{ws}). After conducting each of the three outcome-dependent sampling study designs, we analyzed the resulting data with ascertainment-corrected maximum likelihood, weighted likelihood, and multiple imputation. When conducting the random sampling design, maximum likelihood using subsampled individuals only and an multiple imputation procedure that imputed X_{ei} in unsampled subjects were used for analyses. We replicated the process of sampling participants from the cohort and analyzing the observed data 500 times. We report averages across the replicates for parameter and uncertainty estimates.

Full Cohort Analysis

In **Table 2**, we show the results from the full cohort logistic regression analysis conducted with standard maximum likelihood and from the random sampling, D[100, 1400, 100] and D[275, 1050, 275] study designs. Results for the D[450, 700, 450] study design can be found in eTable 1, http://links.lww.com/EDE/B285. Under the full cohort analysis, current smoking was highly associated with chronic bronchitis risk, and presence of the T allele modified the effect of smoking. The odds ratio (95% confidence interval) for chronic bronchitis comparing smoking with nonsmoking visits was $exp(1.25) = 3.49(2.37, 5.13)$ in those without the T allele, and $exp(1.25 - 0.57) = 1.97(1.60, 2.44)$ in those with at least one T allele copy. The ratio of odds ratios for those with and without the T allele was estimated to be $exp(-0.57) = 0.57(0.37, 0.88)$ consistent with the T allele ameliorating the impact of smoking on chronic bronchitis risk.

Subsample Analyses

Across 500 replicates, on average, all designs and analysis procedure combinations approximately reproduced the point estimates from the FC analysis (**Table 2**), so we describe study design by analysis procedure combination performance by comparing the estimated precision. Results are displayed as standard errors for all parameters in **Table 2**. However, in the **Figure**, we focus on four key parameters in order to gain insights into the utility of the proposed designs. In the **Figure**, we present the ratio of the average estimated variance using a random sampling design with maximum likelihood analysis to each other design/analysis procedure combination. Thus, for each design/analysis procedure combination, we describe an estimate of "relative efficiency" compared with a standard random sampling design and analysis. It is worth noting that the maximum possible efficiency gain over random sampling is $3932/1600 \approx 2.46$, which is the comparison of the maximum likelihood analysis using the full cohort to maximum likelihood analysis using the random sampling design.

Figure panels A and B examine time-varying covariate parameters β_w and β_{ws}, and we observe that with the $D[100, 1400, 100]$ design, both ascertainment-corrected maximum likelihood and multiple imputation estimation procedures are very precise. The impressive precision using a targeted sample with less than half the total cohort is due to two important features of the design and model: (1) nearly all the 1,404 subjects with response variation (i.e., those in stratum 1) are sampled under the design and (2) all variability in $smk_{w,ij}$ and $smk_{w,ij} \cdot snp_i$ occured within subjects. These results are consistent with what we have observed in the past.[1,2] Estimation precision drops as fewer subjects from stratum 1 are included in the outcome-dependent sampling, which is the case with the $D[275, 1050, 275]$ and $D[450, 700, 450]$ designs.

One interesting observation is that for estimates of β_w and β_{ws} the relative variance pattern was not monotonic for weighted likelihood analysis under the three outcome-dependent sampling designs. Relative variance is highest for the $D[275, 1050, 275]$ design and is lower for $D[100, 1400, 100]$ and $D[450, 700, 450]$. There appears to be a tradeoff between the study design and weighted likelihood weighting scheme. An extreme design, for example, $D[100, 1400, 100]$, samples the most informative subjects; however, with weighted likelihood, the design induces an extreme weighting scheme that leads to imprecise estimates. Further research is required to find a general solution for this study design-weighting scheme tradeoff.

Figure panels C and D show that the outcome-dependent sampling designs proposed here are not particularly

TABLE 2. Results from the Full Cohort Analysis and 500 Replicates of Random Sampling and the $D[100,1400,100]$ and $D[275,1050,275]$ Designs

	Full Cohort	Random Sampling		D[100, 1400, 100]			D[275, 1050, 275]		
Variables	**ML**	**ML**	**MI**	**ACML**	**WL**	**MI**	**ACML**	**WL**	**MI**
Primary estimation targets									
SNP (β_s)	-0.08 (0.07)	-0.09 (0.11)	-0.09 (0.11)	-0.15 (0.15)	-0.10 (0.20)	-0.14 (0.15)	-0.08 (0.12)	-0.08 (0.13)	-0.07 (0.12)
Smoking (within; β_w)	1.25 (0.20)	1.25 (0.31)	1.24 (0.28)	1.18 (0.21)	1.33 (0.30)	1.27 (0.21)	1.28 (0.25)	1.28 (0.25)	1.28 (0.23)
SNP by smoking (within; β_{ws})	-0.57 (0.22)	-0.57 (0.36)	-0.55 (0.34)	-0.54 (0.23)	-0.63 (0.33)	-0.59 (0.23)	-0.60 (0.28)	-0.59 (0.28)	-0.60 (0.27)
Time-varying covariates									
Study year (per 3 years)	0.00 (0.01)	0.00 (0.02)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.02)	0.00 (0.01)
Time-fixed/baseline covariates									
Smoking (between; β_b)	2.05 (0.10)	2.06 (0.15)	2.05 (0.10)	1.93 (0.21)	2.09 (0.25)	2.05 (0.10)	2.01 (0.15)	2.07 (0.17)	2.05 (0.10)
Female	-0.15 (0.06)	-0.14 (0.10)	-0.15 (0.06)	-0.21 (0.15)	-0.14 (0.19)	-0.15 (0.06)	-0.17 (0.11)	-0.14 (0.12)	-0.15 (0.06)
Cigarettes/d (per 10 cigarettes/d)	0.16 (0.02)	0.16 (0.04)	0.16 (0.02)	0.08 (0.05)	0.16 (0.07)	0.16 (0.02)	0.12 (0.04)	0.16 (0.04)	0.16 (0.02)
Pack years (per 10 pack years)	0.11 (0.04)	0.10 (0.06)	0.11 (0.04)	0.14 (0.08)	0.12 (0.12)	0.11 (0.04)	0.13 (0.06)	0.11 (0.08)	0.11 (0.04)
Age (per 10 years)	0.08 (0.05)	0.08 (0.08)	0.08 (0.05)	0.05 (0.11)	0.06 (0.15)	0.08 (0.05)	0.04 (0.08)	0.08 (0.09)	0.08 (0.05)
BMI (per 5 kg/m^2)	0.00 (0.04)	0.00 (0.06)	0.00 (0.04)	-0.05 (0.09)	-0.01 (0.12)	0.00 (0.04)	-0.03 (0.07)	0.00 (0.08)	0.00 (0.04)
Response dependence model									
γ	0.53 (0.10)	0.53 (0.15)	0.53 (0.10)	0.54 (0.10)	0.52 (0.10)	0.53 (0.10)	0.53 (0.11)	0.52 (0.11)	0.53 (0.10)
$log(\sigma)$	0.82 (0.04)	0.82 (0.06)	0.82 (0.04)	0.83 (0.07)	0.80 (0.07)	0.82 (0.04)	0.82 (0.05)	0.82 (0.05)	0.82 (0.04)

We report average estimates and, in parentheses, average estimated standard errors on the log odds ratio (logistic) scale.
ACML, ascertainment-corrected maximum likelihood; MI, multiple imputation; ML, maximum likelihood; WL, weighted likelihood.

useful if the target of inference is a subject-level or time-fixed covariate parameter. For example, $D[100,1400,100]$ yields lower precision estimates of β_s than random sampling with maximum likelihood analyses, although precision improves as proportionately more subjects are sampled from strata 0 and 2 with $D[275,1050,275]$ and $D[450,700,450]$.

Even though multiple imputation was not clearly more precise than ascertainment-corrected maximum likelihood for β_w, β_{ws}, and β_s, the utility of multiple imputation is clearly displayed in baseline covariate effects such as β_{gender}. Irrespective of the study design multiple imputation yielded β_{gender} estimates that were as precise as the full cohort analysis since the covariate, gender, is available on all subjects in the dataset, including those who were not included in the outcome-dependent sample. By imputing X_{ei}, we simply allow ourselves to exploit all the existing data in order to summarize the gender–chronic bronchitis association. Without imputing X_{ei}, we effectively throw away available covariate information.

DISCUSSION

We discussed a class of epidemiologic study designs for longitudinal binary response data that can be viewed as an extension to the case–control design. Rather than using a single scalar response to create two sampling strata (cases and controls), we summarize the response vector to create three sampling strata (no events, some events, all events). By oversampling subjects with response variation, we can obtain substantial precision gains for time-varying covariate effects and for their interactions with baseline variables. Furthermore, supplementing the design with multiple imputation analyses yields near full efficiency for the effects of baseline covariates.

Our work provides two key results: (1) subjects with response variability are highly informative for estimating time-varying covariate effects, or equivalently subjects without response variability are relatively uninformative and (2) near full efficiency can be obtained for parameters associated with time-varying covariates that vary exclusively within individuals when all subjects with response variability are sampled. Insight into such results is provided by what is known about conditional logistic regression precision for longitudinal binary data. In contrast to random intercepts models, conditional logistic regression only uses within-cluster response and exposure variation to estimate parameters. Generalized linear mixed models "borrow strength" by exploiting between– and within-subject variability. As discussed in

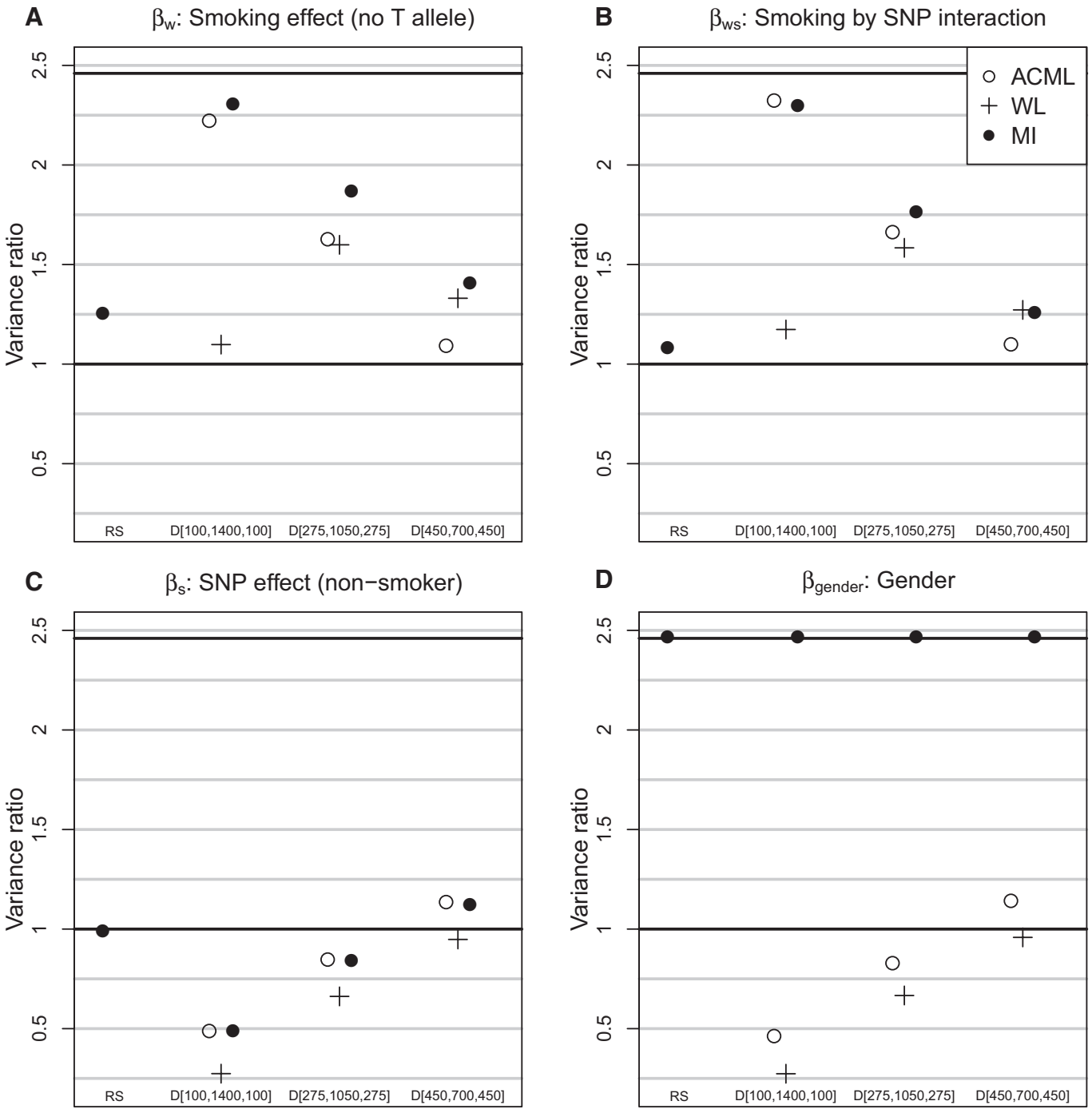

FIGURE. Relative variance: average estimated variance based on a random sampling design with maximum likelihood analysis divided by average estimated variance based on each design by analysis procedure combination (500 replicates). SNP, single nucleotide polymorphism; ACML, ascertainment-corrected maximum likelihood; MI, multiple imputation; and WL, weighted likelihood.

the studies by Neuhaus and Kalbfleisch[10] and Neuhaus and Lesperance[30], generalized linear mixed models are more efficient than conditional logistic regression for all parameters except those associated with time-varying covariates that vary exclusively within subjects (i.e., the intracluster correlation in the covariate is 0), where conditional logistic regression and generalized linear mixed models are equally efficient. That is, toward estimating such parameters, only subjects with response variability contribute information. In the Lung Health Study, the within-subject smoking, $smk_{w,ij}$ variable, had intracluster

correlation equal to 0. Thus, the $D[100, 1400, 100]$, which sampled nearly all subjects with response variability, yielded very precise estimates of β_w and β_{ws}.

Even though we were interested in marginal model parameters, one might also be interested in conditional model estimation (e.g., from mixed effects models). If interest is exclusively in β_w and β_{ws} from a conditional model, Sjölander et al[31] showed that a highly informative study design and estimation procedure combination is one that samples only those with within-subject response and exposure variability (i.e., those who are doubly discordant) and that conducts analyses with conditional logistic regression. That is, we would sample only those who exhibited CB_{ij} and smk_{ij} variability over time. Such a design represents a highly resource-efficient approach and would result in a sample size that is smaller than what we have proposed. Further, it estimates parameters that are analogous to β_w and β_{ws}. What is lost by conducting such a design and analysis is the ability to estimate other parameters efficiently (or at all), to do prediction, to "marginalize" over the exposure distribution to obtain marginal risk contrasts and also the possibility of invalid inferences when the response dependence model differs from a simple exchangeable structure.

The weighted likelihood analysis procedure is not as efficient as ascertainment-corrected maximum likelihood and multiple imputation procedures in settings we studied. However, it is well known that there is a bias-variance tradeoff between maximum likelihood procedures and weighted likelihood in the presence of model misspecification.[32–34] Maximum likelihood is more efficient but less robust than weighted likelihood approaches. Such misspecification might arise, for example, if the mean model for $[\mathbf{Y}_i \mid \mathbf{X}_i]$ ignores an important interaction.

In this article, we have assumed that within-subject smoking is not an endogeneous exposure that is influenced by past outcomes. If the primary exposure of interest is driven by past outcomes, then tailored statistical methods would be needed that can properly address time-dependent confounding.[35,36] Additional research is warranted that considers the potential for efficient outcome-dependent sampling designs in such a complex time-dependent context.

ACKNOWLEDGMENTS

The authors thank the supported effort of the faculty and staff members of the Johns Hopkins University Bayview Genetics Research Facility, NHLBI grant HL066583 (Garcia/Barnes, PI) and NHGRI grant HG004738 (Barnes/Hansel, PI). The Lung Health Study was supported by U.S. Government contract No. N01-HR-46002 from the Division of Lung Diseases of the National Heart, Lung and Blood Institute. Data were downloaded from the NCBI database of genotypes and phenotypes (accession number phs000335.v2.p2). The authors would like to thank the reviewers for their thoughtful comments that led to substantial improvements to this paper.

REFERENCES

1. Schildcrout JS, Heagerty PJ. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*. 2008;9:735–749.
2. Schildcrout JS, Heagerty PJ. Outcome-dependent sampling from existing cohorts with longitudinal binary response data: study planning and analysis. *Biometrics*. 2011;67:1583–1593.
3. Schildcrout JS, Rathouz PJ, Zelnick LR, et al. Biased sampling designs to improve research efficiency: factors influencing pulmonary function over time in children with asthma. *Ann Appl Stat*. 2015;9:731–753.
4. Connett JE, Kusek JW, Bailey WC, et al. Design of the lung health study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Control Clin Trials*. 1993;14(2 Suppl):3S–19S.
5. Anthonisen NR, Connett JE, Kiley JP, et al. Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of FEV1. The Lung Health Study. *JAMA*. 1994;272:1497–1505.
6. Kanner RE, Connett JE, Williams DE, et al. Effects of randomized assignment to a smoking cessation intervention and changes in smoking habits on respiratory symptoms in smokers with early chronic obstructive pulmonary disease: the Lung Health Study. *Am J Med*. 1999;106:410–416.
7. Lee JH, Cho MH, Hersh CP, et al.; COPDGene and ECLIPSE Investigators. Genetic susceptibility for chronic bronchitis in chronic obstructive pulmonary disease. *Respir Res*. 2014;15:113.
8. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbgap database of genotypes and phenotypes. *Nature Genetics*. 2007;39:1181–1186.
9. Franklin JM, Schneeweiss S, Polinski JM, et al. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
10. Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*. 1998;54:638–645.
11. Goetgeluk S, Vansteelandt S. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*. 2008;64:772–780.
12. Brumback BA, Dailey AB, Brumback LC, et al. Adjusting for confounding by cluster using generalized linear mixed models. *Stat Probabil Lett*. 2010;80:1650–1654.
13. Neuhaus JM, McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J R Stat Soc Series B Stat Methodol*. 2006;68:859–872.
14. Brumback BA, Li L, Cai Z. On the use of between-within models to adjust for confounding due to unmeasured cluster-level covariates. *Commun Stat Simul Comput*. 2017;46:3841–3854.
15. Frisell T, Öberg S, Kuja-Halkola R, et al. Sibling comparison designs: bias from non-shared confounders and measurement error. *Epidemiology*. 2012;23:713–720.
16. Sjölander A, Frisell T, Öberg S. Causal interpretations of between-within models for twin research. *Epidemiologic Methods*. 2012;1:217.
17. Azzalini A. Logistic regression for autocorrelated data with application to repeated measures (Corr: 97V84 p989). *Biometrika*. 1994;81:767–775.
18. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*. 1999;55:688–698.
19. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*. 2002;58:342–351.
20. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. *Biometrics*. 1984;40:961–971.
21. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88:9–25.
22. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
23. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42:121–130.
24. Schildcrout JS, Heagerty PJ. Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics*. 2007;63:322–331.
25. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Stat Sci*. 2000;15:1–26.
26. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47:663–685.
27. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846–866.
28. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–592.

29. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* Hoboken, NJ: John Wiley & Sons; 2014.
30. Neuhaus JM, Lesperance ML. Estimation efficiency in a binary mixed-effects model setting. *Biometrika*. 1996;83:441–446.
31. Sjölander A, Johansson ALV, Lundholm C, et al. Analysis of 1: 1 matched cohort studies and twin studies, with binary exposures and binary outcomes. *Stat Sci*. 2012;27:395–411.
32. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *J R Stat Soc Series C Appl Stat*. 1999;48:457–468.
33. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. *J R Stat Soc Series B Methodol*. 1986;48: 170–182.
34. Xie Y, Manski CF. The logit model and response-based samples. *Sociol Methods Res*. 1989;17:283–302.
35. Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11:561–570.
36. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11:550–560.