Published in final edited form as:

Biometrics. 2011 December ; 67(4): 1583–1593. doi:10.1111/j.1541-0420.2011.01582.x.

Outcome dependent sampling from existing cohorts with longitudinal binary response data: study planning and analysis

Jonathan S. Schildcrout and

Departments of Biostatistics and Anesthesiology, Vanderbilt University School of Medicine, 1161 21st Avenue South, S-2323 Medical Center North, Nashville, TN 37232, USA

Patrick J. Heagerty

Department of Biostatistics, University of Washington, F-600 Health Sciences Building, Campus Mail Stop 357232, Seattle, WA, 98105-7232

Jonathan S. Schildcrout: jonathan.schildcrout@vanderbilt.edu; Patrick J. Heagerty: heagerty@u.washington.edu

Summary

When novel scientific questions arise after longitudinal binary data have been collected, the subsequent selection of subjects from the cohort for whom further detailed assessment will be undertaken is often necessary in order to efficiently collect new information. Key examples of additional data collection include retrospective questionnaire data, novel data linkage, or evaluation of stored biological specimens. In such cases, all data required for the new analyses are available except for the new target predictor or exposure. We propose a class of longitudinal outcome dependent sampling schemes and detail a design corrected conditional maximum likelihood analysis for highly efficient estimation of time-varying and time-invariant covariate coefficients when resource limitations prohibit exposure ascertainment on all participants. Additionally, we detail an important study planning phase that exploits available cohort data to proactively examine the feasibility of any proposed substudy as well as to inform decisions regarding the most desirable study design. The proposed designs and associated analyses are discussed in the context of a study which seeks to examine the modifying effect of an interleukin-10 cytokine single nucleotide polymorphism on asthma symptom regression in adolescents participating Childhood Asthma Management Program Continuation Study. Using this example we assume that all data necessary to conduct the study are available except subject-specific genotype data. We also assume that these data would be ascertained by analyzing stored blood samples, the cost of which limits the sample size.

Keywords

Outcome dependent sampling; epidemiological study design; binary data; longitudinal data analysis; marginal models; marginalized models; time-dependent covariates

1. Introduction

The Childhood Asthma Management Program (CAMP) is a large scale randomized clinical trial of 1041 children with mild to moderate asthma. The primary scientific aim is to examine the long term effect of daily inhaled anti-inflammatory medications (budesonide, albuterol, and nedocromil) on lung growth (CAMP Research Group, 1999 and 2000). The original CAMP study, which enrolled five to twelve years olds, followed children over the course of five years. The CAMP Continuation Study (CAMPCS) was a four and a half year observational follow-up study intended to complete assessment of the long term effects of treatment on key asthma outcomes. During CAMPCS, participants were primarily adolescents and teenagers who were returned to their doctors and were prescribed

medication based on standard criteria. While the severity of asthma tends to decrease through the course of adolescence, it is hypothesized that genetic factors may modify this regression. One example of such a candidate factor is a polymorphism at a locus of the Interleukin-10 (IL10) cytokine. IL10 is an anti-inflammatory cytokine that lowers production of macrophages and mononuclear cells. By examining CAMP participants and their parents, Lyon et al (2004) found associations between single nucleotide polymorphisms (SNPs) on the IL10 gene and several asthma phenotypes. The analysis we wish to consider examines the modifying effect of the IL10 SNP on the change, or regression, of asthma symptoms during adolescence.

CAMPCS sought to follow participants quarterly with two in-person clinic visits and two telephone contacts (also referred to as 'visits') per year. At each visit participants completed an interim history report that captured symptoms and clinical events occurring between the present and the most recently past visit. One questionnaire item asked participants to report the number of times since the most recent visit that they had contacted a physician concerning asthma symptoms. We will consider the dichotomized variable defined as 1 if a physician had been contacted at least once, and 0 otherwise. We will examine the course of symptom prevalence in the population for those with and without the IL10 SNP using a longitudinal, logistic regression analysis of the "physician was contacted" phenotype and focus on a SNP indicator, time (since enrollment), and the interaction between genotype and time as the predictors of interest.

Due to the increasing availability of existing cohort study data, administrative information, and electronic medical records data, researchers commonly investigate novel study hypotheses while only requiring ascertainment of select key exposure or confounder variables. For example, the CAMPCS study had collected all covariate and outcome data on the entire cohort, yet genotype data were collected at a later time. If ascertainment costs associated with the desired but missing target covariate are high in terms of monetary cost and/or patient burden (e.g., requiring a new blood sample), and if the outcome is relatively rare, then researchers must carefully consider how they sample subjects from the original cohort in order to estimate parameters efficiently. In the CAMPCS study, only ten percent of all interim history reports (across all children and over time) noted that a physician had been contacted, and 60 percent of all participants had never contacted a physician over the course of follow-up. As is the case with univariate case-control, nested case-control, and other epidemiologic designs, targeted sampling of rare endpoints can be crucial for the construction of adequately powered retrospective longitudinal studies.

To address high exposure ascertainment costs, we discuss planning and analysis of retrospective outcome dependent sampling (ODS) designs for longitudinal binary response data. Specifically, we consider a class of ODS designs that sample based on a summary of an individual's subject-specific response vector, followed by a maximum likelihood analysis that corrects for the ascertainment scheme. In addition, we advocate the use of a proactive comparative feasibility study prior to choosing a final sampling design that aims to exploit existing data to make informed choices regarding the appropriate study design.

Previous research has explored select retrospective longitudinal sampling designs and analyses. Cai et al. (2001) proposed inverse probability of sampling weighted generalized estimating equations for retrospective sampling of longitudinal binary response data. Pfeiffer et al (2005) proposed a longitudinal data analogue to the the case-cohort (Prentice, 1986) and the bidirectional, case-crossover designs (MacClure, 1991; Navidi, 1998), and Park and Kim (2004) describe a design similar to the case-crossover design, that samples all subjects at pre-specified but random times and additionally observes all subjects when they have events. Neuhaus, Scott and Wild (2006) discussed an outcome dependent sampling

framework for longitudinal data with profile likelihood based analyses. Finally, Schildcrout and Rathouz (2010) addressed longitudinal follow-up on individuals following case-control and stratified case-control sampling.

The basic class of designs discussed here is a generalization of one proposed by Schildcrout and Heagerty (2008) in which subjects exhibiting response variation (i.e., ‘responders’ who have at least one positive binary outcome, but not all positive outcomes) are eligible for the ODS study while non-responders (e.g., those in whom none or all of the binary responses are positive) are not. We demonstrated that estimators for the regression coefficient of time-varying covariates, X_{ij} where i denotes subject and j denotes time, are often highly efficient even relative to analysis of the original full cohort while offering the practical and monetary advantage of sub-sampling a fraction of subjects (e.g., fifty percent). However, the ODS design of Schildcrout and Heagerty (2008) was shown to be relatively inefficient for covariates that do not vary within an individual, $X_{ij} \equiv X_i$, or covariates that had large between-person variation, $\text{var}(\bar{X}_i)$, relative to the within-person variation, $\text{var}(X_{ij} - \bar{X}_i)$. To potentially increase estimation efficiency for a broader array of covariates, we propose an extended class of ODS designs that specifies selection of both “responders” and “non-responders” but possibly with different probabilities. The generalization allows us to consider a range of sampling probabilities for subjects categorized into distinct strata based on their longitudinal data: 1) those who never exhibit symptoms; 2) those who have symptoms on some but not all visits; and 3) those who have symptoms on all visits. Evaluation of a set of select alternative designs specified in terms of outcome-strata sampling probabilities allows us to choose a design which is most efficient within the class for key regression coefficients. Furthermore, simulation-based evaluation allows us to explore the power of the proposed study design under various alternative hypotheses.

This manuscript is organized as follows. In Section 2, we propose an extended class of ODS study designs, an Ascertainment Corrected Maximum Likelihood (ACML) estimation procedure, and an important study planning feasibility analysis to proactively anticipate the resulting precision of ODS designs within the candidate class. We evaluate design and estimation efficiency using simulation studies in section 3. In section 4 we revisit the CAMPCS study and discuss implementation of the proposed design class, the proactive study feasibility analysis, and ACML estimation. Finally, we conclude with a discussion in section 5.

2. Design, Estimation, and Planning

Towards efficient estimation of coefficients for a broad class of target covariate distributions, we propose a stratified random sampling approach where sampling probabilities are a function of a coarse categorization of the sum of subject-specific responses, $\sum_j y_{ij}$, where y_{ij} denotes the outcome for subject i at the j th measurement time. Specifically, if $\mathbf{y}_i \equiv (y_{i1}, \dots, y_{in_i})$ is subject i 's, $i \in \{1, 2, \dots, N\}$, binary response vector from a cohort with N total subjects, and S_i is the indicator that subject i is included in the outcome dependent sample, then the subject-specific probability of being sampled is given by,

$$pr(S_i=1|\mathbf{y}_i) \equiv \pi\left(\sum_j y_{ij}\right) = \begin{cases} \pi(0), & \sum_j y_{ij}=0 \\ \pi\{(0, n_i)\}, & 0 < \sum_j y_{ij} < n_i \\ \pi(n_i), & \sum_j y_{ij}=n_i. \end{cases} \quad (1)$$

We will assume the original cohort from which the ODS sample is drawn is representative of the target population and has a total of N subjects where $(N_0, N_{(0,n_i)}, N_{n_i})$ denotes the

number of subjects in each response stratum $[\sum_j y_{ij} = 0, 0 < \sum_j y_{ij} < n_i, \sum_j y_{ij} = n_i]$ with $N_0 + N_{(0,n_i)} + N_{n_i} = N$. Within this class, unique designs are defined by $[\pi(0), \pi\{(0, n_i)\}, \pi(n_i)]$, and for each design, on average, $(N_0^s, N_{(0,n_i)}^s, N_{n_i}^s)$ subjects are selected from the cohort with $N^s = N_0^s + N_{(0,n_i)}^s + N_{n_i}^s$. We denote designs by $D[N_0^s, N_{(0,n_i)}^s, N_{n_i}^s]$.

Let $\mathbf{X}_i$ be the $n_i \times p$ design matrix for subject i , and assume the multivariate $[\mathbf{Y}_i | \mathbf{X}_i]$ distribution is known up to a parameter θ [i.e., $pr(\mathbf{y}_i | \mathbf{x}_i) = pr(\mathbf{y}_i | \mathbf{x}_i; \theta)$]. Under random sampling the contribution of subject i to the likelihood is given by $L_i(\theta | \mathbf{y}_i, \mathbf{x}_i) = pr(\mathbf{y}_i | \mathbf{x}_i; \theta)$. For a generalized linear mixed model (Stiratelli et al, 1984; Breslow and Clayton, 1993) with mean model given by $\mu_{g,ij} = \text{logit}^{-1}(\mathbf{X}_{ij}\boldsymbol{\beta}^{glimm} + \mathbf{b}_i)$, the induced likelihood contribution

for subject i is given by $L_i(\theta^{glimm}) = pr(\mathbf{y}_i | \mathbf{x}_i; \theta^{glimm}) = \int \mathbf{b}_i \left[\prod_{j=1}^{n_i} \mu_{g,ij}^{y_{ij}} (1 - \mu_{g,ij})^{1-y_{ij}} \right] dF(\mathbf{b}_i)$. In the marginalized model class (e.g., Heagerty, 1999; Schildcrout and Heagerty, 2007), the marginal mean $\mu_{m,ij} = \text{logit}^{-1}(\mathbf{X}_{ij}\boldsymbol{\beta}^{mmm})$ captures the relationship between responses $\mathbf{Y}_i$ and covariates $\mathbf{X}_i$ and conditional mean or response dependence model $\mu_{c,ij} = \text{logit}^{-1}(\Delta_{ij} + \mathbf{Z}_{ij}\boldsymbol{\alpha})$ describes the relationship among responses. The dependence model design matrix $\mathbf{Z}_i$ may include transition terms or random effects. The parameter $\boldsymbol{\alpha}$ captures the strength of response dependence, and the value Δ_{ij} , described in manuscripts such as Heagerty and Zeger (2000) and Lee and Daniels (2008), links the marginal and condition means. The contribution of subject i to the likelihood for the marginalized model under random

sampling is given by $L_i(\theta) = \int \mathbf{z}_i \left[\prod_{j=1}^{n_i} \mu_{c,ij}^{y_{ij}} (1 - \mu_{c,ij})^{1-y_{ij}} \right] dF(\mathbf{z}_i)$.

Regardless of the specific form of longitudinal model, when using an ODS design the likelihood contribution for subject i is conditional on their being selected for ascertainment. Therefore, the key to constructing the likelihood function is the inclusion of an additional term representing the ascertainment probability. Algorithmically, what is needed are the score and information corresponding to the basic longitudinal model, $pr(\mathbf{y}_i | \mathbf{x}_i; \theta)$, which are now available from public and/or commercial software, and the first and second derivatives of the ascertainment term, $pr(S_i = 1 | \mathbf{x}_i)$. Using Bayes Theorem, we can show that the ascertainment corrected likelihood contribution for subject i is given by

$$\begin{aligned} L_i^c(\theta | \mathbf{y}_i, \mathbf{x}_i) &= pr(\mathbf{y}_i | \mathbf{x}_i, S_i = 1) \\ &= pr(S_i = 1 | \mathbf{y}_i, \mathbf{x}_i) \cdot pr(\mathbf{y}_i | \mathbf{x}_i; \theta) \cdot \{pr(S_i = 1 | \mathbf{x}_i; \theta)\}^{-1} \\ &= \pi\left(\sum_j y_{ij}\right) \cdot L_i \cdot [\pi(0) - \pi\{(0, n_i)\}]L_{i0} + [\pi(n_i) - \pi\{(0, n_i)\}]L_{in_i} + \pi\{(0, n_i)\}]^{-1} \end{aligned} \quad (2)$$

where $L_{i0} = pr(\sum_j y_{ij} = 0 | \mathbf{x}_i; \theta)$ and $L_{in_i} = pr(\sum_j y_{ij} = n_i | \mathbf{x}_i; \theta)$. The ascertainment corrected log-likelihood $[l_i^c \equiv \log(L_i^c)]$ score function for parameter θ is given by,

$$\begin{aligned} \frac{\partial l_i^c}{\partial \theta} &= \sum_{i=1}^{N^s} \frac{\partial L_i}{\partial \theta} \frac{1}{L_i} \\ &\quad - \sum_{i=1}^{N^s} \left\{ (\pi(0) - \pi\{(0, n_i)\}) \frac{\partial L_{i0}}{\partial \theta} + (\pi(n_i) - \pi\{(0, n_i)\}) \frac{\partial L_{in_i}}{\partial \theta} \right\} \cdot \{\pi\{(0, n_i)\} + (\pi(0) - \pi\{(0, n_i)\})L_{i0} + (\pi(n_i) - \pi\{(0, n_i)\})\}^{-1} \end{aligned} \quad (3)$$

By maximizing the ascertainment corrected likelihood we are able to estimate all parameters corresponding to the full cohort model, $pr(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta})$. Note that the key terms in the ascertainment correction to the likelihood (and score equations) are simply obtained from standard longitudinal model likelihood and score equation calculations for two outcome scenarios: L_{ini} where all $y_{ij} \equiv 1$; and L_{i0} where all $y_{ij} \equiv 0$. Therefore, practical implementation of ODS can be based on standard longitudinal likelihood calculations where for each subject their direct likelihood contribution is combined with two “pseudo-outcome” contributions that form the basis of the ascertainment correction.

While the proposed design is agnostic to model choice, we will focus on marginalized models in the presentation that follows. Finally, the use of ascertainment corrected maximum likelihood is commonly considered in family studies where clusters of data (families) are sampled conditional on at least one affected subject (see Glidden and Liang, 2002 or Clayton 2003 for discussion of details).

2.1 Comparative Design/Feasibility Analysis

We now propose a pre-implementation evaluation that permits exploration and comparison of the precision of select alternative designs prior to choosing and executing a specific ascertainment plan, $[\pi(0), \pi\{(0, n_i)\}, \pi(n_i)]$. As opposed to standard power, sample size, and precision calculations for prospective studies, in this retrospective design setting, we have all covariate and outcome information on subjects except for a single key covariate. Therefore, we are far more informed and have to make fewer assumptions than we would normally make when planning a prospective study. If we could simulate exposure, $\mathbf{x}_{e,i}$, then alternative sampling plans could be compared. However, to respect the ODS design we actually need to simulate $\mathbf{x}_{e,i}$ given the available data $\mathbf{y}_i$ and $\mathbf{x}_{o,i}$. Our proposed procedure treats the desired, and not yet ascertained, exposure measure as missing data, and an expectation-maximization (EM; Dempster, Laird, and Rubin, 1986) algorithm allows us to estimate the ‘missing’ exposure distribution conditional on the observed data. We can then simulate data under various potential designs in order to evaluate the resulting power and/or precision.

The key using existing cohort data to simulate alternative sub-sampling designs is the specification of two models: the regression model of interest; and the marginal exposure probabilities. In order to describe the details we first establish notation. Let $(\mathbf{x}_{o,i}, \mathbf{y}_i)$ be subject i ’s observed data prior to ascertaining the key exposure $\mathbf{x}_{e,i}$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{(o)}, \boldsymbol{\beta}^{(e)}, \boldsymbol{\alpha})$ be the vector of parameters for the target (marginalized) model $pr(\mathbf{y}_i | \mathbf{x}_{o,i}, \mathbf{x}_{e,i}; \boldsymbol{\theta})$, and let $\boldsymbol{\omega}$ be the parameter vector for an exposure prevalence model conditional on measured covariates: $pr(\mathbf{x}_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega})$. We want to generate the ‘missing’ exposure $\mathbf{x}_{e,i}$ consistent with assumed value for $\boldsymbol{\beta}^{(e)}$ and conditional on the outcomes that have been observed.

From Bayes’ Theorem,

$$pr(\mathbf{x}_{e,i} | \mathbf{x}_{o,i}, \mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\omega}) \propto pr(\mathbf{y}_i | \mathbf{x}_{o,i}, \mathbf{x}_{e,i}; \boldsymbol{\theta}) \cdot pr(\mathbf{x}_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega}), \quad (4)$$

and if $(\boldsymbol{\theta}, \boldsymbol{\omega})$ is known, we may repeatedly generate $\mathbf{x}_{e,i}$ from this posterior distribution, and for each replicate, conduct an ODS design and associated analysis. Summary statistics across replicates (e.g., average estimated variance) can then be used to guide the choice of a final sampling plan.

Specification of the parameters $(\boldsymbol{\theta}, \boldsymbol{\omega})$ needed to evaluate (4) would generally be challenging and involves making assumptions regarding associations that can be informed by the

observed cohort data $(\mathbf{x}_{o,i}, \mathbf{y}_i)$. Therefore, we propose an adaptive specification of θ that uses $\hat{\theta}(\beta^{(e)}, \omega)$ such that estimates of both $\beta^{(o)}$ and dependence parameters α can be obtained from the observed data combined with the exposure assumptions given by $\beta^{(e)}$ and ω .

An EM algorithm for estimating $\hat{\theta}(\beta^{(e)}, \omega)$ is implemented in the standard two steps. An initial estimate of the conditional exposure probability given by (4) can be obtained using the marginal prevalence model and the assumed parameter ω . The iterative algorithm proceeds with:

Expectation step—Calculate

- a. Likelihood contributions: $pr(\mathbf{y}_i | \mathbf{x}_{o,i}, \mathbf{x}_{e,i}; \widehat{\beta^{(o)}}, \widehat{\alpha}, \beta^{(e)})$ for each of $\mathbf{x}_{e,i} \in \{0, 1\}$
- b. Conditional probabilities:

$$\widehat{P}_i \{ \mathbf{x}_{e,i} = x | \mathbf{y}_i, \mathbf{x}_{o,i}; \widehat{\theta}(\beta^{(e)}, \omega) \} = pr \{ \mathbf{x}_{e,i} = x | \mathbf{y}_i, \mathbf{x}_{o,i}; \widehat{\beta^{(o)}}, \widehat{\alpha}, \beta^{(e)}, \omega \} \text{ for } x \in \{0, 1\}$$

Maximization step—Holding $\beta^{(e)}$ and ω fixed, calculate the maximum profile likelihood

estimate $(\widehat{\beta^{(o)}}, \widehat{\alpha})$ for the j^{th} iteration by maximizing the expected log-likelihood:

$$\widehat{\theta}^{(j+1)}(\beta^{(e)}, \omega) : \arg\max_{\theta} Q \{ \theta(\beta^{(e)}, \omega) | \widehat{\theta}^{(j)}(\beta^{(e)}, \omega) \}$$

$$Q[\theta | \widehat{\theta}^{(j)}] = \sum_{i=1}^N \sum_{x=0}^1 \log \{ pr(\mathbf{y}_i | \mathbf{x}_{o,i}, \mathbf{x}_{e,i} = x; \beta^{(o)}, \alpha, \beta^{(e)}) \} \times \widehat{P}_i \{ \mathbf{x}_{e,i} = x | \mathbf{y}_i, \mathbf{x}_{o,i}; \widehat{\theta}^{(j)}(\beta^{(e)}, \omega) \}$$

The EM steps are iterated until a specified convergence criterion is achieved. Given that the ultimate goal is to simulate data using final estimates of the conditional exposure probabilities, $\widehat{P}_i \{ \mathbf{x}_{e,i} = x | \mathbf{y}_i, \mathbf{x}_{o,i}; \widehat{\theta}(\beta^{(e)}, \omega) \}$, it is reasonable to monitor these target estimates for convergence. Once $\mathbf{x}_{e,i}$ can be appropriately generated, then $m = 1, 2, \dots, M$ simulated

data sets can be created with $(\mathbf{x}_{o,i}, \mathbf{x}_{e,i}^{(m)}, \mathbf{y}_i)$ available on all $i \in \{1, 2, \dots, N\}$ subjects. Finally, we may evaluate any proposed ODS design by subsampling from the three outcome strata using probabilities $[\pi(0), \pi\{(0, n_i)\}, \pi(n_i)]$ and estimating the target parameter, $\beta^{(e)}$, and standard error, using ACML.

3. Evaluation of Proposed Designs and Estimation Procedure

Using simulations we explore the performance of the proposed class of designs under combinations of characteristics that may be present in typical cohort studies. Motivated by the CAMP example, we focus on regression parameters that capture group-by-time interactions which would be used to test for various longitudinal trends across subgroups of subjects. In different simulation scenarios we contrast key characteristics of longitudinal binary data: baseline response prevalence (low and moderate); and cluster size (small and moderate to large clusters; fixed and variable cluster sizes).

To conduct the simulation study, we generated data from a first order marginalized transition population model (Heagerty, 2002) model,

$$\text{logit}(\mu_{m,ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 g_i + \beta_3 t_{ij} g_i; \text{logit}(\mu_{c,ij}) = \Delta_{ij} + \alpha_1 y_{ij-1}. \quad (5)$$

The covariate value g_i was binary and subject-level (e.g., presence of two copies of the minor allele, T on the sixth locus of IL10 cytokine), and it was generated under the assumption that $pr(g_i = 1) = 0.25$. The time vector for subject i , t_i , was equal to $(1, 2, \dots, n_i)$, and under the scenarios studied the number of observations per subject, n_i , was: 1) 4, 2) uniformly distributed between three and seven, $U(3, 7)$, 3) 10, and 4) $U(5, 15)$. The primary scientific goal of the ODS designs focuses on valid and efficient estimation of $(\beta_1, \beta_2, \beta_3)$. When n_i was equal to 4 and when it was $U(3, 7)$, $(\beta_0, \beta_1, \beta_2, \beta_3, \alpha_1)$ was set equal to $(-1.5, -0.15, 3, 0.15, 2)$. The combination of parameter values and cluster sizes led to, on average, stratum-specific sample sizes, $(N_0, N_{(0,n_i)}, N_{n_i})$, approximately equal to $(475, 375, 150)$ when $n_i = 4$ and $(440, 425, 135)$ when $n_i \sim U(3, 7)$. In the moderate to large cluster settings where n_i was 10 or was $U(5, 15)$, $(\beta_0, \beta_1, \beta_2, \beta_3, \alpha_1)$ was set equal to $(-2.85 \text{ or } -1.85, -.15, 1, .15, 1.5)$. When $\beta_0 = -1.85$, on average, $(N_0, N_{(0,n_i)}, N_{n_i})$ was approximately $(400, 600, 0)$, and when $\beta_0 = -2.85$, symptom prevalence dropped substantially, and $(N_0, N_{(0,n_i)}, N_{n_i})$ was, on average, approximately $(650, 350, 0)$.

To illustrate the ODS designs, we will assume that, at the planning stage, g_i is unavailable on all subjects, but that it can be ascertained retrospectively. We also assume resource limitations permit inclusion of 250 participants from the larger cohort of 1000. Note that to examine the large sample properties, we also conducted simulations based on sampling 2500 participants from a larger cohort of 10000. These results are not shown. Estimators were unbiased and the relative efficiency of designs and estimation procedures were very much the same as those observed in the scenarios discussed here.

For each of 1000 replications of the simulation, we generated the original cohort, and then sampled from it according to four designs. The first design was random sampling of 250 participants. The three ODS designs were defined by stratum-specific sampling fractions, $[\pi(0), \pi\{(0, n_i)\}, \pi(n_i)]$. In the small cluster scenarios, where n_i was equal to 4 or was $U(3, 7)$, we sampled with probabilities so that, on average, $(N_0^s, N_{(0,n_i)}^s, N_{n_i}^s) \in \{(0, 250, 0), (15, 220, 15), (30, 190, 30)\}$. In the large cluster scenarios, which approximate the CAMPCS study, it was very rare to observe $\sum_j y_{ij} = n_i$, and the three ODS designs sampled with probability so that $(N_0^s, N_{(0,n_i)}^s, N_{n_i}^s) \in \{(0, 250, 0), (15, 235, 0), (45, 205, 0)\}$.

With random sampling, we use maximum likelihood (ML) for estimation. With ODS, we used ACML as described earlier, and for comparison, we also consider the weighted estimating equations (WEE) approach of Cai et al, (2001) where weights were equal to the inverse of the probability of being sampled or observed (Robins et al, 1994). The estimating equation we solved for WEE is a reweighted version of score equations for the marginalized

model, and is given by, $\sum_{i=1}^{N_s} \left\{ \pi \left(\sum_j y_{ij} \right)^{-1} \cdot \partial l_i / \partial \theta \right\} = \mathbf{0}$, where l_i is the log-likelihood for the marginalized transition model. Since the correlation structure is properly specified in the estimating equation, the WEE estimator is efficient within the class of estimators. Empirical variances of the solution, $\hat{\theta}_{WEE}$, are used for comparisons with ACML. Simulation results are shown in Table 1.

3.1 Results for Time-varying Covariate Parameters (β_1, β_3)

We now consider estimates for time-varying covariate coefficients, β_1 and β_3 . In the scenarios studied parameter estimates for all designs and estimators were effectively unbiased with none exceeding 4% in magnitude. These are not shown. Similarly, estimated standard errors were largely unbiased with the largest percent bias observed when $n_i = 4$, design $D[30, 190, 30]$ was implemented, and with WEE estimation. In this case the

(rounded) average estimated standard error was 6.7×10^{-2} while the empirically derived standard error in the estimators was slightly smaller, 6.3×10^{-2} .

Compared to random sampling with ML estimation, the ODS designs and estimation procedures were superior with regards to estimation efficiency and standard error reductions ranged from 10% to 30%. The most efficient design tended to be $D[0, 250, 0]$. This was most pronounced in the large sample scenario with relatively high symptom prevalence (i.e., when $\beta_0 = -1.85$). For example, when $n_i \sim U(5, 15)$, the empirical standard errors for β_1 and β_3 under $D[0, 250, 0]$ were 2.1×10^{-2} and 2.9×10^{-2} , respectively. These were at least 17% smaller than those from the other designs. In the small cluster scenarios (e.g., $n_i = 4$) and in the large cluster scenarios with low prevalence (e.g., when $\beta_0 = -2.85$), sampling several of the non-responders led to approximately equally efficient estimators as compared to $D[0, 250, 0]$. In fact, the $D[15, 235, 0]$ design may be the most efficient for rarer outcomes. For example, empirical standard errors for β_1 and β_3 were 3.1×10^{-2} and 4.2×10^{-2} , respectively, for $D[0, 250, 0]$ with low prevalence and when $n_i \sim U(5, 15)$. For $D[15, 235, 0]$ they were 2.9×10^{-2} and 3.9×10^{-2} , respectively.

When the cluster sizes were fixed, properly specified WEE estimators were as efficient as ACML estimators. However, when cluster sizes varied, efficiency gains of ACML over WEE were observed. With $D[15, 220, 15]$, and $n_i \sim U(3, 7)$ empirical standard errors for β_1 and β_3 were 4.5×10^{-2} and 8.2×10^{-2} , respectively. With WEE, they were 5.2×10^{-2} and 9.3×10^{-2} , respectively. Similar efficiency gains were observed under $D[15, 235, 0]$ with $n_i \sim U(5, 15)$.

3.2 Results for time-invariant covariate parameter β_2

Efficiency gains of ODS with ACML estimation were also observed with β_2 though were less dramatic than with β_1 and β_3 . In all scenarios considered, one of the ODS designs with ACML estimation was most efficient among the designs and estimation procedure combinations, but this varied widely by scenario. For example, in the large cluster, high prevalence scenarios $D[0, 250, 0]$ was most efficient among the designs; however it was least efficient with small clusters. In fact when in the small sample scenarios, it was even less efficient than random sampling (e.g., empirical standard errors were 5.25×10^{-1} for $D[0, 250, 0]$ and were 3.80×10^{-1} under random sampling). The inefficiency of the design $D[0, 250, 0]$ with small clusters is understandable since it is well known that with classical conditional logistic regression for estimation we are unable to estimate any time invariant covariate coefficients. With $n_i = 3$ our ascertainment conditioning is not on the exact sum (like classical conditional logistic regression) but rather the only slightly coarse sum of either 1 or 2, and therefore we would expect the efficiency of the design to be relatively poor. With small cluster sizes and $D[15, 220, 15]$ and with large clusters, $\beta_0 = -2.85$, and $D[15, 235, 0]$ efficiency gains with ACML over WEE were substantial (e.g. 2.64×10^{-1} versus 4.15×10^{-1}). This is an expected result since for WEE estimators sampled non-responders were assigned very large inverse probability weights relative to the responders, and circumstances in which few subjects receive very large weights relative to others are known to lead to inefficient estimators.

4. Example: Childhood Asthma Management Program

We wish to examine the time course of symptoms during the first two years of CAMPCS for children with and without a polymorphism at a specific locus in the gene for the IL10 cytokine. Of the 1041 children who participated in the CAMP study, 941 had completed CAMPCS interim history reports containing the longitudinal outcome, “called physician regarding symptoms,” and 538 also had genotypic data related to a specific locus of IL10 cytokine. For the sake of comparing sampling schemes, we limited the age range to 12 to 16

years, and we removed all children who had less than four interim history reports completed. The 476 children remaining compose the original cohort in our hypothetical study. While we have genotypic data on all children, we act as if we do not and want to conduct a substudy that, due to resource limitations, can only collect genotypic data on a fraction of the full cohort.

Table 2 shows characteristics of the CAMPCS cohort. Of particular interest, 60 percent of participants never contacted their physician for symptoms, and thirty percent of children had two copies of the minor allele T on the sixth locus of the IL10 cytokine. To conduct the hypothetical substudy, we assume a recessive genetic model using the indicator variable, $g_i = I_i(TT)$, which must be ascertained retrospectively using stored blood samples. We further assume that resources permit us to include 190 children (40 percent) in the substudy. We chose to subsample 190 children for simplicity since this was also the number of children who exhibited response variation during follow-up.

To separate cohort and longitudinal effects, we decomposed $age_i(t_{ij})$ into a baseline age, $age_i(0)$, and time since baseline t_{ij} (which is a target covariate). The longitudinal binary response variable, y_i , was 1 if the participant or parent had called a physician due to symptoms between successive clinic visits and 0 otherwise. Our analysis focuses on parameters β_g and β_{gt} in the marginalized transition and latent variable model (Schildcrout and Heagerty, 2007),

$$\begin{aligned} \text{logit}(\mu_{m,ij}) &= \beta_0 + \beta_t t_{ij} + \beta_g g_i + \beta_{gt} g_i t_{ij} + \beta_{age} age_i(0) + \beta_f I_i(female) + \dots \\ \text{logit}(\mu_{c,ij}) &= \Delta_{ij} + \alpha_1 y_{ij-1} + b_i \end{aligned} \quad (6)$$

where $I_i(female)$ is 1 if subject i is female and 0 if male, and $b_i \sim N(0, \alpha_2^2)$. Other adjustment variables included race, city, and season (to control for seasonal confounding). Each of these were included in the regression model as indicator variables, e.g., three for race, seven for city, and three for season.

4.1 Comparative Feasibility Analysis for CAMPCS

We now consider the planning phase of a substudy from CAMPCS. At this point we assume all information on the subjects is available except for g_i . We compare sampling strategies defined by $[\pi(0), \pi\{(0, n_i)\}, \pi(n_i)]$ to identify the most efficient design for estimating target coefficients. Among the subjects in the original cohort study, 285 never called their physician, 190 called their physician at least once, but not at every study visit, and only one called his/her physician at every clinic visit. To conduct the proactive feasibility analysis we followed the protocol detailed in section 2.1.

We considered three ODS sampling schemes: $D[0, 190, 0]$, $D[10, 180, 0]$, and $D[25, 165, 0]$. The corresponding values of $[\pi(0), \pi\{(0, n_i)\}, \pi(n_i)]$ for these strategies are $(0, 1, 0)$, $(10/285, 180/190, 0)$, and $(25/285, 165/190, 0)$, respectively. Because the values of ω and $\beta^{(e)}$ are not known during this phase of the study, we evaluated six comparative feasibility (CF) study scenarios based on three values of ω in $pr(x_{e,i} | \mathbf{x}_{o,i}; \omega)$ and two values of (β_g, β_{gt}) . These scenarios are described explicitly in Table 3. In comparative feasibility (CF) studies CF-1, CF-2, and CF-3, (β_g, β_{gt}) was set to the null value $(0, 0)$, and in CF-4, CF-5, and CF-6, it was set to $(0.35, -0.37)$. In CF-1 and CF-4, ω was set so that $pr(x_{e,i} | \mathbf{x}_{o,i}; \omega) = 0.2$, and in CF-3 and CF-6 it was set so that $pr(x_{e,i} | \mathbf{x}_{o,i}; \omega) = 0.3$. In CF-2 and CF-5, $pr(x_{e,i} | \mathbf{x}_{o,i}; \omega)$ was allowed to vary with race and city. For these data, $pr(x_{e,i} | \mathbf{x}_{o,i}; \omega) = pr(g_i = 1 | \text{race}_i, \text{city}_i; \omega)$. In general one would need to rely on prior data to inform the marginal exposure model assumptions in order to conduct study planning evaluations. However, for illustration

we also used assumptions for $pr(x_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega})$ based on fitted values from the logistic regression model,

$$\text{logit}\{pr(g_i=1|\text{race}_i, \text{city}_i; \boldsymbol{\omega})\} = \omega_0 + I_i(\text{race})\omega_r + I_i(\text{city})\omega_c, \quad (7)$$

although in practice g_i would not be available at the planning stage. This decision simply allows us to explore more complex exposure models that are consistent with the structure present in the example data. Among all CF scenarios CF-5 most accurately depicts the relationships contained in the CAMPCS study. It is very important to note that in a real study, this model is not available. We use it here as a gold standard. Scientific topic experts are extremely important for positing reasonable models for the unknown exposure distribution. Table 4 displays coefficient estimates from the exposure distribution logistic regression model. Race was related to genotype although city was not. The log odds ratio associated with being black, hispanic, and other races, relative to being white, were 0.63 (standard error: 0.36), 1.07 (0.40), and 0.99 (0.36), respectively. Though CF-3 and CF-6 capture the marginal $pr(g_i = 1) = 0.30$ correctly, they ignore the relationship between g_i and race. This misspecification could potentially lead to overestimates or underestimates of design precision.

Table 5 shows standard error estimates (e.g., square root of the average variance estimates) and null hypothesis rejection probability estimates for β_g and β_{gt} under CF-1 to CF-6 across 500 replicates of data generated from model (4). The efficiency gains with the ODS designs relative to random sampling are moderate for β_g and are large for β_{gt} . In this analysis all ODS designs appear to be approximately equally efficient. In CF-5 standard error and power estimates for β_g with D[10,180,0] were approximately 0.34 and 0.18, respectively. With random sampling the standard error estimate was approximately 18% larger (0.39), and the power was approximately 17% lower (0.15). For β_{gt} the standard error was 50% larger with random sampling as compared to the D[10,180,0] design (0.36 versus 0.24). Similarly, the power was reduced by approximately 50% (0.19 versus 0.40).

Standard errors were affected by assumptions regarding $pr(x_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega})$. For example with D[0,190,0] standard errors for β_{gt} were approximately 0.28 when $pr(x_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega}) = 0.2$, and 0.24 when $pr(x_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega}) = f(\text{race}, \text{city})$. However, in these data, capturing the relationship between g_i and race correctly did not impact standard error estimates, since assuming $pr(x_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega}) = 0.3$, also yielded standard errors equal to 0.24. Had the covariate relationships been stronger, we might expect larger differences in standard error estimates between, say CF-5 and CF-6. Finally, we observed little to no association between effect size and standard error estimates. For example, comparing CF-1 to CF4 under design D[25,165,0] for β_g led to standard error estimates of 0.39 and 0.38, respectively. Larger differences in standard error estimates are anticipated if assumed covariate effects are far from the ‘true’ values.

In summary, the comparative feasibility analysis revealed that all ODS designs were far more efficient than random sampling; however, at best, the power to detect target covariate effects was approximately 0.4. Specification of $pr(x_{e,i} | \mathbf{x}_{o,i}; \boldsymbol{\omega})$ had a moderate effect on standard error estimates; however, assumptions regarding (β_g, β_{gt}) had no effect on precision. Our results only apply to these data and to the assumptions we have made. Results would change with a different dataset and with different assumptions, which points to the importance of conducting this comparative feasibility analysis exercise.

4.2 CAMPCS Results

Table 6 shows the results from analyses of CAMPCS. There was little evidence to suggest a major effect of the IL10 locus 6 polymorphism on asthma symptom prevalence among adolescents ages 12 to 16. For example, with the D[10,180,0] design the log odds ratio associated with 1 year time increments was 0.10 (standard error: 0.14) for those without the polymorphism and $0.10 - 0.41 = -0.31$ for those with it. The parameter estimate associated with the main effect of IL10 was 0.36 (0.34) indicating a positive but non-significant association between g_i and symptom prevalence at baseline. While estimates of target coefficients ($\beta_r, \beta_g, \beta_{gt}$) were similar across study designs, all ODS designs (which were equally efficient) were far more efficient than random sampling. Additionally, though we only sampled approximately 40 percent of children from the full cohort (190/476), standard errors for time-varying covariate coefficients under ODS were only 17 percent larger than those from analyses of the entire cohort.

To verify the accuracy of the comparative feasibility analysis discussed in the previous subsection, we may compare the estimated standard errors obtained in the CAMPCS study to those of the feasibility analysis under scenarios CF-5. Table 6 shows that the estimated standard errors for the main effect of the IL10 polymorphism and the interaction with time were 0.32 and 0.23 in the CAMP analysis using design D[0,190,0]. These values were very well estimated by the CF-5 analysis shown in the fifth line of table 5. Also note that in CF-1 and CF-4 where the prevalence was misspecified and assumed to be 0.2, standard error estimates were overestimated. For example, in CF-4 the anticipated standard error for β_{gt} was 0.28 for D[10, 180, 0], while in the actual CAMPCS analysis, it was 0.23.

5. Discussion

We have discussed planning and analysis of ODS designs for longitudinal, binary response data in settings where all but a key exposure or possibly a confounder variable is available prior to the study. These designs are particularly effective when exposure ascertainment costs limit sample size and when there is insufficient response variability to estimate exposure effects efficiently with standard, random sampling designs. With the increasing availability of cohort study and administrative data, we believe these and other ODS designs for longitudinal data will become increasingly important.

We proposed a class of ODS designs for longitudinal binary response data where sampling is based on three strata defined by summaries of subject-specific response vectors: 1) the never-responders or those who do not exhibit symptoms, 2) the responders or those who exhibit response variation, and 3) the ever-responders or those who exhibit symptoms at all observations. While these designs are often more efficient than random sampling with rare outcomes, it is important to proactively examine the feasibility of the study as a whole and to examine sampling probabilities from the three strata that maximize design efficiency. We posed a comparative feasibility analysis, that can be used for this purpose. By positing the marginal distribution of the ‘missing’ exposure and the dependence of the response on this exposure, we are able to explore the efficiency of several sampling schemes. This planning tool relies on an EM algorithm to generate a number of datasets that can then be used to conduct replicates of the ODS designs. By summarizing the behavior of the estimators across replicates, informed design choices can be made.

The designs proposed here can be extended to circumstances in which sampling is based not only on summaries of subject-specific response vectors, but also on another variable that is correlated with the retrospectively ascertained exposure. Though not considered here, other estimators, including the profile likelihood estimators of Neuhaus et al (2006), may use the available data (all but the key exposure) on non-sampled subjects more efficiently than the

ACML estimator. Extensions to the continuous case scenarios will provide challenges as evaluation of integrals are required to identify the ascertainment corrected likelihood. However, the range of possible designs is far greater with continuous response data. The work of Zhou et al (2002, 2007) and Weaver et al (2005), who have developed ODS designs for univariate continuous data, provide important insights into how this may be accomplished.

Acknowledgments

This research was funded by the National Heart Lung and Blood Institute grant number R01 HL94786. The work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. The Childhood Asthma Management Program trial and CAMP Continuation Study were supported by contracts N01-HR-16044, 16045, 16046, 16047, 16048, 16049, 16050, 16051, and 16052 with the National Heart, Lung, and Blood Institute and General Clinical Research Center grants M01RR00051, M01RR0099718-24, M01RR02719-14, and RR00036 from the National Center for Research Resources. The CAMP Genetics Ancillary Study is supported by grants U01HL075419, U01HL65899, and P01HL083069 from the National Heart Lung and Blood Institute.

References

- Anderson JA. Separate sample logistic discrimination. *Biometrika*. 1972; 59:19–35.
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 1993; 88:9–25.
- Cai J, Qaqish B, Zhou H. Marginal analysis for cluster-based case-control studies. *Sankhyā, Series B*. 2001; 63:326–337.
- CAMP, R. G. Recruitment of participants in the childhood Asthma Management Program (CAMP). I. Description of methods: Childhood Asthma Management Program Research Group. *J Asthma*. 1999; 36:217–237. [PubMed: 10350219]
- CAMP, R. G. Long-term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine*. 2000; 343 (15):1054–1063. [PubMed: 11027739]
- Clayton D. Conditional likelihood inference under complex ascertainment using data augmentation. *Biometrika*. 2003; 90:976–981.
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (C/R: P22-37). *Journal of the Royal Statistical Society, Series B: Methodological*. 1977; 39:1–22.
- Glidden D, Liang KY. Ascertainment adjustment in complex diseases. *Genetic Epidemiology*. 2002; 23:201–208. [PubMed: 12384971]
- Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*. 1999; 55:688–698. [PubMed: 11314994]
- Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*. 2002; 58:342–351. [PubMed: 12071407]
- Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*. 2000; 15:1–26.
- Lee K, Daniels MJ. Marginalized models for longitudinal ordinal data with application to quality of life studies. *Stat Med*. 2008; 27:4359–4380. [PubMed: 18613246]
- Lyon H, Lange C, Lake S, Silverman E, Randolph A, Kwiatkowski D, Raby B, Lazarus R, Weiland K, Laird N, Weiss S. IL10 gene polymorphisms are associated with asthma phenotypes in children. *Genet Epidemiol*. 2004; 26:155–165. [PubMed: 14748015]
- Maclure M. The case-crossover design: A method for studying transient effects on risk of acute events. *American Journal of Epidemiology*. 1991; 133:144–153. [PubMed: 1985444]
- Navidi W. Bidirectional case-crossover designs for exposures with time trends. *Biometrics*. 1998; 54:596–605. [PubMed: 9629646]
- Neuhaus JM, Jewell NP. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*. 1990; 46:977–990. [PubMed: 2085642]
- Neuhaus JM, Scott AJ, Wild CJ. Family-specific approaches to the analysis of case-control family data. *Biometrics*. 2006; 62:488–494. [PubMed: 16918913]

- Park E, Kim Y. Analysis of longitudinal data in case-control studies. *Biometrika*. 2004; 91:321–330.
- Pfeiffer RM, Ryan L, Litonjua A, Pee D. A case-cohort design for assessing covariate effects in longitudinal studies. *Biometrics*. 2005; 61:982–991. [PubMed: 16401271]
- Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–412.
- Qaqish BF, Zhou H, Cai J. On case-control sampling of clustered data. *Biometrika*. 1997; 84:983–986.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Schildcrout J, Heagerty P. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*. 2008; 9:735–749. [PubMed: 18372397]
- Schildcrout JS, Heagerty PJ. Marginalized Models for Moderate to Long Series of Longitudinal Binary Response Data. *Biometrics*. 2007; 63:322–331. [PubMed: 17688485]
- Schildcrout JS, Rathouz PJ. Longitudinal Studies of Binary Response Data Following Case-Control and Stratified Case-Control Sampling: Design and Analysis. *Biometrics*. 2009
- Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binary response. *Biometrics*. 1984; 40:961–971. [PubMed: 6534418]
- Weaver MA, Zhou H. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*. 2005; 100:459–469.
- Zhou H, Chen J, Rissanen TH, Korrick SA, Hu H, Salonen JT, Longnecker MP. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*. 2007; 18:461–468. [PubMed: 17568219]
- Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*. 2002; 58:413–421. [PubMed: 12071415]

Table 1

Average estimated standard error $[-] (\times 10^{-1})$ and the empirical standard error or the standard deviation of the parameter estimates $(\times 10^{-1})$ across 1000 replications $(-)$ of the simulation. All parameter estimates were unbiased (i.e., less than 4 percent bias for all estimators) and so they are not shown. Data for the original cohorts were generated from a marginalized transition model with parameter values shown in the first column of the table. From the original cohort, the random sampling design, and the ODS designs, sampled a total of 250 subjects from the original cohort of 1000. Estimation used maximum likelihood (ML) for the random sampling design and both ACML and WEE for the ODS designs. The value n_i corresponds to the distribution of cluster sizes, $(N_0, N_{(0;n_i)}, N_{n_i})$ is the average number subjects allocated to each of the three response strata $(\sum_j y_{ij} = 0, 0 < \sum_j y_{ij} < n_i, \sum_j y_{ij} = n_i)$, and $D[N_0^s, N_{(0;n_i)}^s, N_{n_i}^s]$ denotes the design with the average number of subjects sampled within the three strata.

$(\beta_0, \beta_1, \beta_2, \beta_3, a_1)$	n_i	$(N_0, N_{(0;n_i)}, N_{n_i})$	$(N_0^s, N_{(0;n_i)}^s, N_{n_i}^s)$	Estimation Procedure	β_0	β_1	β_2	β_3	β_1
$(-1.5, -0.15, 3, 0.15, 2)$	4	(475, 375, 150)	Random Sampling	ML	[1.81] (1.77)	[0.94] (0.93)	[3.65] (3.80)	[1.80] (1.85)	[2.29] (2.35)
			D[0, 250, 0]	ACML	[2.69] (2.76)	[0.58] (0.59)	[5.17] (5.27)	[1.10] (1.06)	[2.93] (2.82)
			D[15, 220, 15]	ACML	[1.93] (1.85)	[0.62] (0.60)	[3.36] (3.31)	[1.17] (1.18)	[2.15] (2.02)
			D[15, 220, 15]	WEE	[2.35] (2.32)	[0.63] (0.62)	[4.00] (4.27)	[1.22] (1.23)	[2.69] (2.74)
	U(3,7)	(440, 425, 135)	D[30, 190, 30]	ACML	[1.75] (1.72)	[0.66] (0.63)	[3.10] (2.95)	[1.26] (1.20)	[2.02] (1.95)
			D[30, 190, 30]	WEE	[1.87] (1.85)	[0.67] (0.63)	[3.40] (3.35)	[1.29] (1.24)	[2.22] (2.22)
			Random Sampling	ML	[1.72] (1.79)	[0.66] (0.65)	[3.40] (3.39)	[1.20] (1.19)	[2.05] (2.00)
			D[0, 250, 0]	ACML	[2.21] (2.20)	[0.44] (0.44)	[4.12] (4.16)	[0.81] (0.81)	[2.26] (2.29)
			D[15, 220, 15]	ACML	[1.79] (1.78)	[0.47] (0.45)	[3.13] (3.04)	[0.85] (0.82)	[1.85] (1.84)
			D[15, 220, 15]	WEE	[2.25] (2.30)	[0.53] (0.52)	[3.74] (4.05)	[0.93] (0.93)	[2.24] (2.32)
			D[30, 190, 30]	ACML	[1.67] (1.77)	[0.50] (0.50)	[2.96] (3.12)	[0.91] (0.93)	[1.79] (1.81)
			D[30, 190, 30]	WEE	[1.81] (1.87)	[0.52] (0.53)	[3.25] (3.48)	[0.94] (0.97)	[1.93] (1.95)
$(-1.85, -0.15, 1, 0.15, 1.5)$	10	(400, 600, 0)	Random Sampling	ML	[1.61] (1.64)	[0.35] (0.35)	[2.62] (2.61)	[0.51] (0.53)	[1.51] (1.51)
			D[0, 250, 0]	ACML	[1.49] (1.49)	[0.23] (0.24)	[2.04] (2.01)	[0.34] (0.34)	[1.18] (1.17)
			D[15, 235, 0]	ACML	[1.59] (1.58)	[0.28] (0.29)	[2.33] (2.30)	[0.41] (0.43)	[1.35] (1.32)
			D[15, 235, 0]	WEE	[1.93] (1.91)	[0.28] (0.29)	[2.82] (2.91)	[0.41] (0.43)	[1.59] (1.59)
	U(5,15)	(400, 600, 0)	D[45, 205, 0]	ACML	[1.52] (1.53)	[0.30] (0.30)	[2.37] (2.36)	[0.44] (0.44)	[1.37] (1.37)
			D[45, 205, 0]	WEE	[1.55] (1.56)	[0.30] (0.30)	[2.44] (2.42)	[0.44] (0.44)	[1.40] (1.40)
			Random Sampling	ML	[1.54] (1.48)	[0.32] (0.31)	[2.48] (2.45)	[0.44] (0.45)	[1.53] (1.49)
			D[0, 250, 0]	ACML	[1.50] (1.47)	[0.22] (0.21)	[2.04] (1.96)	[0.30] (0.29)	[1.19] (1.19)
			D[15, 235, 0]	ACML	[1.58] (1.54)	[0.26] (0.26)	[2.30] (2.32)	[0.36] (0.37)	[1.36] (1.36)
			D[15, 235, 0]	WEE	[1.94] (1.90)	[0.28] (0.28)	[2.90] (3.04)	[0.39] (0.40)	[1.57] (1.64)

$(\beta_0, \beta_1, \beta_2, \beta_3, \alpha_1)$	n_i	$(N_0, N_{(0,n_i)}, N_{n_i})$	$(N_0^s, N_{(0,n_i)}^s, N_{n_i}^s)$	Estimation Procedure	β_0	β_1	β_2	β_3	β_1
$(-2.85, -0.15, 1, 0.15, 1.5)$	10	$(650, 350, 0)$	D[45, 205, 0]	ACML	[1.48] (1.47)	[0.27] (0.28)	[2.31] (2.26)	[0.39] (0.38)	[1.39] (1.39)
			D[45, 205, 0]	WEE	[1.52] (1.51)	[0.28] (0.28)	[2.39] (2.30)	[0.39] (0.38)	[1.41] (1.42)
			Random Sampling	ML	[2.33] (2.36)	[0.53] (0.53)	[3.49] (3.34)	[0.71] (0.69)	[2.46] (2.45)
			D[0, 250, 0]	ACML	[2.79] (2.73)	[0.35] (0.34)	[3.30] (3.27)	[0.47] (0.45)	[2.02] (1.95)
			D[15, 235, 0]	ACML	[2.05] (2.01)	[0.32] (0.31)	[2.68] (2.64)	[0.43] (0.43)	[1.74] (1.71)
	U(5,15)	$(650, 350, 0)$	D[15, 235, 0]	WEE	[2.59] (2.62)	[0.32] (0.31)	[3.88] (4.15)	[0.43] (0.43)	[2.50] (2.68)
			D[45, 205, 0]	ACML	[1.84] (1.82)	[0.34] (0.33)	[2.61] (2.60)	[0.46] (0.44)	[1.76] (1.72)
			D[45, 205, 0]	WEE	[1.92] (1.93)	[0.34] (0.33)	[2.93] (2.95)	[0.46] (0.44)	[1.98] (2.01)
			Random Sampling	ML	[2.23] (2.14)	[0.48] (0.47)	[3.30] (3.16)	[0.62] (0.60)	[2.47] (2.65)
			D[0, 250, 0]	ACML	[2.82] (2.85)	[0.33] (0.31)	[3.32] (3.30)	[0.42] (0.42)	[2.01] (2.00)
U(5,15)	$(650, 350, 0)$	D[15, 235, 0]	ACML	[2.04] (2.06)	[0.30] (0.29)	[2.66] (2.63)	[0.39] (0.39)	[1.73] (1.73)	
		D[15, 235, 0]	WEE	[2.61] (2.63)	[0.34] (0.34)	[4.08] (4.09)	[0.47] (0.49)	[2.43] (2.65)	
		D[45, 205, 0]	ACML	[1.80] (1.78)	[0.32] (0.32)	[2.55] (2.58)	[0.41] (0.42)	[1.76] (1.76)	
		D[45, 205, 0]	WEE	[1.89] (1.88)	[0.33] (0.33)	[2.94] (2.98)	[0.44] (0.46)	[1.97] (2.02)	

Table 2

Demographics characteristics and experiences of children in the CAMPCS with IL10 cytokine SNP information. Continuous variables are summarized with the 10th, 25th, 50th, 75th, and 90th percentiles of their distributions, subjects per city with frequencies, and other categorical variables with proportions

Number of children	476
City	
Albuquerque	44
Baltimore	64
Boston	59
Denver	53
San Diego	42
Seattle	68
St. Louis	76
Toronto	70
Baseline age (years)	12.2, 13, 14.3, 15.6, 16.4
Female	0.36
Race	
White	0.75
Black	0.10
Hispanic	0.07
Other	0.08
Follow-up visits	6, 7, 7, 8, 8
Years of follow-up	1.6, 1.8, 1.9, 1.9, 2.0
Physician seen for symptoms (as a proportion of visits)	0.1
Physician never seen for symptom (as a proportion of children)	0.6
Two copies of minor allele T on locus 6 of the IL10 cytokine	0.3

Table 3

Study planning scenarios considered for the comparative feasibility analysis. We examined six combinations of $\text{pr}(\mathbf{x}_{e,i} | \mathbf{x}_{o,i}; \omega) \equiv \text{pr}(g_i = 1 | \text{race}_i, \text{city}_i; \omega)$ and $\boldsymbol{\beta}^{(e)} \equiv (\beta_g, \beta_{gt})$ in order to assess the impact that these values have on parameter uncertainty estimates. In CF-1, CF-2, CF-5, and CF-6, we assumed a fixed prevalence for $g_i \equiv I_i(\text{TT})$. In CF-1 and CF-2 it was fixed at 0.2, and in CF-5 and CF-6 it was fixed at 0.3. In CF-3 and CF-4, it was allowed to vary as a function of gender and city. We calculated $\text{pr}(g_i = 1 | \text{age at baseline}_i, \text{city}_i; \omega)$ by using the fitted values from the logistic regression model on all 476 participants shown in table 4. This was intended to capture the “true” value of $\text{pr}(g_i = 1 | \mathbf{x}_{o,i}; \omega)$. We consider two values for (β_g, β_{gt}) ; (0, 0) in CF-1, CF-2, and CF-3, and (0.35, -0.37) in CF-4, CF-5, and CF-6. The former value assumes the null hypothesis of no main effect or time interaction for the IL10 SNP. The latter value is approximately equal to the value observed in the original, full cohort analysis.

	(β_g, β_{gt})	$\text{pr}(g_i = 1 \mathbf{x}_{o,i}; \omega)$
CF-1	(0, 0)	0.2
CF-2	(0, 0)	$f(\text{race}, \text{city})$
CF-3	(0, 0)	0.3
CF-4	(0.35, -0.37)	0.2
CF-5	(0.35, -0.37)	$f(\text{race}, \text{city})$
CF-6	(0.35, -0.37)	0.3

Table 4

Logistic regression parameter estimates for $pr(g_i = 1 | \text{race}_i, \text{city}_i; \omega)$ used in CF-2 and CF-5

	Estimate	Standard Error	Z value	P-value
Intercept	-1.159	0.299	-3.878	< 0.001
Race				
Black	0.628	0.363	1.728	0.084
Hispanic	1.073	0.395	2.717	0.007
Other	0.991	0.357	2.774	0.006
City				
Albuquerque	-0.348	0.481	-0.724	0.469
Boston	-0.349	0.426	-0.817	0.414
Denver	0.227	0.422	0.538	0.590
San Diego	0.179	0.455	0.394	0.694
Seattle	0.265	0.398	0.666	0.506
St. Louis	0.162	0.388	0.416	0.677
Toronto	0.154	0.403	0.381	0.703

Table 5

Results from the comparative feasibility study summarizing parameter and standard error estimates across 500 datasets generated from model (4) using EM algorithm described in section 2.1. Random sampling and three ODS designs were considered, $D[0, 190, 0]$, $D[10, 180, 0]$ and $D[25, 165, 0]$. Six feasibility study scenarios were examined with $\text{pr}(g_1 = 1 | \mathbf{x}_{0,i}; \omega)$ equal to 0.2 in CF-1 and CF-4, and 0.3 in CF-3 and CF-6. In CF-2 and CF-5 it was allowed to vary as a function of race and city according to the estimates shown in table 4. For $\beta^{(e)} \equiv (\beta_g, \beta_{gt})$ we considered the values (0, 0) and (0.35, -0.37). Standard error is the square root of the average estimated variance, and $\text{pr}(\text{Reject } H_0)$ is the proportion of datasets in which the null hypothesis (i.e., the parameter is zero) was rejected at the two-sided $\alpha = 0.05$ significance level.

	(β_g, β_{gt})	$pr(x_{e,i} \mathbf{x}_{0,i}; \omega)$	CFS Setting	β_g			β_{gt}				
				Random Sampling	D[0,190,0]	D[10,180,0]	D[25,165,0]	Random Sampling	D[0,190,0]	D[10,180,0]	D[25,165,0]
Standard Error	(0,0)	0.2	CF-1	0.46	0.38	0.39	0.39	0.42	0.28	0.28	0.29
	(0,0)	f(race, site)	CF-2	0.40	0.33	0.34	0.34	0.36	0.24	0.24	0.25
	(0,0)	0.3	CF-3	0.40	0.32	0.34	0.34	0.36	0.24	0.24	0.25
	(0.35,-0.37)	0.2	CF-4	0.44	0.37	0.38	0.38	0.42	0.28	0.28	0.28
	(0.35,-0.37)	f(race, site)	CF-5	0.39	0.33	0.34	0.34	0.36	0.24	0.24	0.25
	(0.35,-0.37)	0.3	CF-6	0.38	0.32	0.33	0.33	0.36	0.24	0.24	0.25
pr(Reject H_0)	(0,0)	0.2	CF-1	0.06	0.06	0.06	0.06	0.04	0.05	0.04	0.04
	(0,0)	f(race, site)	CF-2	0.05	0.07	0.05	0.05	0.03	0.04	0.04	0.05
	(0,0)	0.3	CF-3	0.06	0.03	0.04	0.04	0.04	0.04	0.03	0.03
	(0.35,-0.37)	0.2	CF-4	0.13	0.15	0.12	0.15	0.14	0.23	0.25	0.22
	(0.35,-0.37)	f(race, site)	CF-5	0.15	0.20	0.18	0.19	0.18	0.38	0.40	0.35
	(0.35,-0.37)	0.3	CF-6	0.15	0.17	0.18	0.17	0.16	0.34	0.34	0.30

Table 6

Analysis of the CAMPCS averaged across 500 replicated studies. At each replicate we subsampled according to four designs: random sampling of 190 individuals and then three ODS designs. Design $D[N_0^s, N_{(0,n_i)}^s, N_{n_i}^s]$ sampled N_0^s individuals who did not exhibit symptoms during follow-up, $N_{(0,n_i)}^s$ individuals who exhibited symptoms on at least one but not all interim history reports, and $N_{n_i}^s$ who exhibited symptoms on all of them. We display average parameter estimates and square roots of average variances to approximate standard errors. Estimates from the random sampling design were obtained via maximum likelihood while those for the ODS designs were obtained via ACML.

	Original cohort	Random Sampling	$D[0, 190, 0]$	$D[10, 180, 0]$	$D[25, 165, 0]$
Intercept	-1.77 (0.26)	-1.83 (0.42)	-1.18 (0.37)	-1.38 (0.39)	-1.54 (0.39)
Time (per year)	0.08 (0.12)	0.09 (0.20)	0.09 (0.14)	0.10 (0.14)	0.10 (0.14)
IL-10 Locus 6	0.35 (0.24)	0.36 (0.38)	0.36 (0.32)	0.36 (0.34)	0.36 (0.33)
Time (per year) · IL-10 Locus 6	-0.37 (0.21)	-0.38 (0.35)	-0.43 (0.23)	-0.41 (0.23)	-0.41 (0.24)
Baseline age (years)	-0.17 (0.05)	-0.17 (0.08)	-0.14 (0.08)	-0.17 (0.09)	-0.19 (0.09)
Female (versus male)	0.26 (0.16)	0.25 (0.27)	0.44 (0.25)	0.47 (0.27)	0.45 (0.26)
Race (versus white)					
African American	0.02 (0.27)	0.01 (0.44)	-1.13 (0.44)	-0.94 (0.42)	-0.67 (0.41)
Asian	0.44 (0.31)	0.42 (0.51)	-0.50 (0.50)	-0.30 (0.50)	-0.08 (0.48)
Other	-0.32 (0.33)	-0.37 (0.56)	-0.12 (0.51)	-0.21 (0.56)	-0.25 (0.53)
Season (versus Winter)					
Spring	-0.28 (0.14)	-0.28 (0.23)	-0.29 (0.16)	-0.28 (0.16)	-0.28 (0.16)
Summer	-0.89 (0.16)	-0.91 (0.26)	-0.94 (0.18)	-0.92 (0.18)	-0.92 (0.18)
Fall	-0.33 (0.15)	-0.34 (0.24)	-0.29 (0.16)	-0.29 (0.16)	-0.30 (0.17)
City (versus Baltimore)					
Albuquerque	-0.52 (0.37)	-0.54 (0.61)	-0.66 (0.62)	-0.73 (0.63)	-0.68 (0.59)
Boston	0.35 (0.29)	0.36 (0.48)	0.14 (0.41)	0.20 (0.45)	0.25 (0.45)
Denver	-0.36 (0.33)	-0.35 (0.53)	-0.87 (0.53)	-0.80 (0.54)	-0.64 (0.51)
San Diego	-0.21 (0.36)	-0.23 (0.59)	-0.07 (0.53)	-0.13 (0.57)	-0.14 (0.56)
Seattle	0.08 (0.29)	0.11 (0.47)	-0.17 (0.41)	-0.16 (0.46)	-0.13 (0.46)
St. Louis	-0.15 (0.29)	-0.15 (0.47)	0.06 (0.41)	0.01 (0.46)	-0.02 (0.46)
Toronto	-0.42 (0.31)	-0.42 (0.50)	-0.77 (0.47)	-0.79 (0.50)	-0.72 (0.49)
Dependence model parameters					
γ	0.04 (0.22)	0.04 (0.35)	0.02 (0.22)	0.02 (0.22)	0.02 (0.24)
$\log(\sigma)$	0.25 (0.10)	0.18 (0.18)	-0.25 (0.30)	-0.01 (0.21)	0.08 (0.17)

Note: CAMPCS data were obtained with permission from the CAMP Data Coordinating Center.