Published in final edited form as:

Biometrics. 2013 June ; 69(2): . doi:10.1111/biom.12013.

Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics

Jonathan S. Schildcrout^{1,*}, Shawn P. Garbett², and Patrick J. Heagerty³

¹Departments of Biostatistics and Anesthesiology, Vanderbilt University School of Medicine, Nashville, TN, USA

²Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN, USA

³Department of Biostatistics, University of Washington, Seattle, WA, USA

Summary

The analysis of longitudinal trajectories usually focuses on evaluation of explanatory factors that are either associated with rates of change, or with overall mean levels of a continuous outcome variable. In this manuscript we introduce valid design and analysis methods that permit outcome dependent sampling of longitudinal data for scenarios where all outcome data currently exist, but a targeted sub-study is being planned in order to collect additional key exposure information on a limited number of subjects. We propose a stratified sampling based on specific summaries of individual longitudinal trajectories, and we detail an ascertainment corrected maximum likelihood approach for estimation using the resulting biased sample of subjects. In addition, we demonstrate that the efficiency of an outcome-based sampling design relative to use of a simple random sample depends highly on the choice of outcome summary statistic used to direct sampling, and we show a natural link between the goals of the longitudinal regression model and corresponding desirable designs. Using data from the Childhood Asthma Management Program, where genetic information required retrospective ascertainment, we study a range of designs that examine lung function profiles over four years of follow-up for children classified according to their genotype for the IL 13 cytokine.

Keywords

asthma; biased sampling; linear mixed effects models; longitudinal data analysis; outcome dependent sampling; time-dependent covariates

1. Introduction

1.1 Overview

The expansion of accessible administrative databases and linked electronic medical records is generating new opportunities for conducting clinical studies of factors associated with longitudinal changes in health status. For example, pharmacoepidemiologic studies can link detailed changes in drug treatment to subsequent patient outcomes, while genetic studies can

*jonathan.schildcrout@vanderbilt.edu.

Supplementary Materials

The R code used to conduct the analyses described in section 4 and shown in table 4 are available with this paper at the Biometrics website on Wiley Online Library.

link patient characteristics with rates of disease progression or with individual response to therapy. Although key outcomes and select covariates are becoming readily available, many scientific questions require the additional collection of costly variables that are the primary predictors of interest. Key examples include patient genotype, or clinical status as characterized by a transcription, metabolomic, or other high-dimensional profile. Frequently the collection of additional exposure information is expensive, and cost constraints limit the number of subjects for whom it can be ascertained. It is crucial to use available data to develop a sampling design that targets subjects who are likely to be the most informative.

1.2 Motivating Example

The Childhood Asthma Management Program (CAMP; CAMP Research Group, 1999 and 2000) trial is the most comprehensive study of the long term effects of anti-inflammatory medication on lung growth in children with mild to moderate asthma. In CAMP, 1041 children in eight cities were randomized to one of two anti-inflammatory medications (budesonide and nedocrilil) or to placebo. A primary aim of the study was to compare among treatment groups at the end of the study period lung function as measured by post-bronchodilator forced expiratory volume in 1 second (FEV_1). Although for the primary endpoint there was no observed treatment effect, many substudies have followed CAMP and have observed, for example, that FEV_1 is lower in CAMP participants greater than 10 years than in age-sex matched children from cohorts of children without asthma (Strunk et al. 2006). Further, a number of substudies have sought to examine genetic modifiers of asthma severity and lung function (e.g., Lyon et al. 2004). In order to conduct these genetic substudies, genotype data had to be ascertained retrospectively. Therefore, subsequent substudies incurred an additional cost with the collection of new covariate data. Even though the CAMP trial genotyped the majority of children for single nucleotide polymorphisms (SNPs), in many cases, retrospective ascertainment of a key exposure variable can be resource intensive and therefore can limit sample size. In circumstances where nearly all data to address a scientific hypothesis are available except for a single, expensive exposure or confounder variable, targetted, outcome dependent sampling (ODS) designs can be highly efficient relative to randomly subsampling from the original cohort.

Our specific interest is in examining the longitudinal lung function profile over the course of follow-up for subjects with and without at least one variant allele (G allele) on locus 2 of the IL 13 cytokine. We use the CAMP data to illustrate the impact that targetted sampling can have on efficiency when subsampling from a cohort, and specifically when interest is in a longitudinal continuous response. We consider the situation where key explanatory data such as genotype data are expensive to ascertain, but we have FEV_1 and other covariate data for subjects over the course of the longitudinal follow-up. The goal is to conduct a new substudy where we would estimate the rate of change in FEV_1 for those subjects with and without the candidate IL13 polymorphism. We also assume that genotypic data are not available at the time the substudy is conceived, and they must be ascertained retrospectively at a cost that limits the sample size. In order to make inference regarding rates of change it would be natural to use a linear mixed effects model with random intercepts and slopes that characterize heterogeneity in FEV_1 level (the intercept) and in FEV_1 changes over time. The parameters addressing our study hypotheses correspond to the main effects of the genotype indicator (presence of a variant allele), time since enrollment, and the genotype-by-time interaction. The key scientific issue is how to efficiently select a subset of individuals on whom detailed data will be ascertained, and specifically whether sampling should be based on the observed outcome profiles.

1.3 Background

While case-control designs are ubiquitous in epidemiology, there is a relatively small literature on ODS for continuous longitudinal response data. The seminal work of Anderson (1972) and Prentice and Pyke (1979) showed that one may ignore the biased case-control design and analyze study data with logistic regression as if the data had been collected prospectively. In contrast, most generalized ODS designs require explicit adjustment for the biased design. For example, with continuous univariate response data Lawless, Kalbfleisch, and Wild (1999) describe a series of likelihood and pseudo-likelihood based approaches to correct for the biased design (or for missing data). Zhou et al. (2002) proposed a design that combines a sample drawn from the tails of a continuous response with a random sample and developed an estimated likelihood approach for analysis. Weaver and Zhou (2005) and Zhou et al. (2007) proposed an empirical likelihood extension to the analyses, and Zhou et al. (2011), proposed a two-phase extension. A number of authors have discussed unplanned, outcome dependent follow-up in the longitudinal binary data setting (e.g., Lin and Ying, 2001; Buzkova and Lumley, 2009) and others have discussed it in the continuous response setting (e.g., Lipsitz et al., 2002). Towards planned ODS, Neuhaus, Scott, and Wild (2006) have discussed a general profile likelihood based framework for ODS with longitudinal data. Schildcrout and Heagerty (2008, 2011) described an ascertainment corrected likelihood approach for subject-level sampling with longitudinal binary response data. Schildcrout and Rathouz (2010) and Schildcrout et al. (2012) addressed subject-level and observation-level sampling, respectively, using an auxiliary variable that is related to the binary response variable in combination with semi-parametric estimation.

The present research addresses outcome dependent sampling for longitudinal, continuous response data where we sample subjects based on low-dimensional summary statistics that reflect key features of their response profile. An ascertainment corrected maximum likelihood procedure is proposed for inference. In Section 2 we introduce the target model, the class of outcome dependent sampling designs for continuous response data, the ascertainment corrected likelihood used for estimation and inference and an algorithm for parameter estimation and inference using an *ascertainment corrected maximum likelihood* (ACML) approach. In Section 3 we examine the finite sampling operating characteristics of these estimators and in Section 4 we analyze the CAMP data to examine a range of designs and analytical approaches. Finally, Section 5 provides a summary and a discussion.

2. Model, Design, and Conditional Likelihood

We now introduce the longitudinal model, the novel ODS sampling design, and a simple ascertainment corrected maximum likelihood approach that accounts for the intentionally biased sampling design.

2.1 Model

The ODS design and estimation strategy we propose presumes that the original cohort of N subjects has the longitudinal outcome data and basic covariate data available, but in order to address a new scientific question a novel exposure must be ascertained. If complete data were available, each would admit (Y_i, X_i) , $i \in 1, 2 \dots N$, where Y_i is a n_i -vector of response values, and X_i is a $n_i \times p$ design matrix. In analyses of the CAMP data our interest is in lung function over the course of follow-up, and the differences in change in lung function for those with and without the IL13 SNP. More generally, we would like to evaluate a time-invariant covariate (e.g. group), a time-varying covariate, and their interaction.

We adopt the linear mixed model of Laird and Ware (1982) given as: $Y_i = X_i\beta + Z_i b_i + \epsilon_i$, where β is a p -vector of fixed-effect coefficients, Z_i is a $n_i \times q$ design matrix for the random

effects, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, and $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma)$. For the CAMP example the design matrix, $\mathbf{X}_i$ is given by $(\mathbf{1}, \mathbf{T}_i, \mathbf{X}_{e,i}, \mathbf{X}_{e,i} \cdot \mathbf{T}_i, \mathbf{X}_{o,i})$, where $\mathbf{X}_{e,i}$ is an expensive and not yet ascertained predictor of interest, $\mathbf{T}_i = \{T_{ij}\}_{j \in 1,2,\dots,n_i}$ is the vector of times and $\mathbf{X}_{o,i}$ is the matrix of pre-existing or inexpensive confounder variables. The random effects design matrix is given by $\mathbf{Z}_i = (\mathbf{1}, \mathbf{T}_i)$, $\mathbf{b}_i = (b_{0i}, b_{1i})$, and $\mathbf{D}_i$ is the 2×2 covariance matrix that contains the variance components $(\sigma_0^2, \sigma_1^2, \rho)$, where ρ is the correlation between b_{0i} and b_{1i} . For ease of exposition, we will assume $\Sigma \equiv \sigma_e^2 \cdot \mathbf{I}_{n_i}$ where $\mathbf{I}_{n_i}$ is a diagonal identity matrix. We may write the multivariate density for this model with the induced multivariate normal distribution:

$$f(\mathbf{Y}_i | \mathbf{X}_i; \theta) = (2\pi)^{-n_i/2} |\mathbf{V}_i|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{Y}_i - \boldsymbol{\mu}_i)^t \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right\}, \quad (1)$$

where $\theta = (\boldsymbol{\beta}, \sigma_0, \sigma_1, \rho)$, $\boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta}$, $\mathbf{V}_i = \mathbf{Z}_i \mathbf{D}_i \mathbf{Z}_i^t + \sigma_e^2 \mathbf{I}$. If a sample of N_r subjects was drawn randomly from the target population, inferences for target model parameters could be made

by maximizing the log-likelihood: $l(\theta; \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{N_r} l_i(\theta; \mathbf{Y}_i, \mathbf{X}_i) = \sum_{i=1}^{N_r} \log f(\mathbf{Y}_i | \mathbf{X}_i; \theta)$. Details regarding first and second partial derivatives with respect to all parameters are shown in, for example, Laird and Ware (1982), Lindstrom and Bates (1988), and McCulloch and Searle (2001).

2.2 ODS study design

The innovation that we detail has two key facets: we propose sampling based on a user-defined *low-dimensional summary* of the outcome vector $\mathbf{Y}_i$; and we propose a *coarse sampling* design where sampling is dependent on strata defined by the summary measure, but not on the exact value of the summary measure. Intuitively, a case-control design oversamples those subjects who are unusual in terms of their outcome, and such a design provides for efficient inference. To generalize the concept to longitudinal data one needs to consider characteristics of the response vector that would naturally identify an individual as interesting or relatively unusual. In order to formalize this idea let $Q_i = g(\mathbf{Y}_i, \mathbf{X}_{oi})$ define a function of the response and observed covariates that summarizes important features of the response vectors. For example, if $\mathbf{W}_i = \frac{1}{n_i} \mathbf{1}^t$, then $Q_i = \mathbf{W}_i \mathbf{Y}_i$ is subject i 's mean value.

Alternatively, if $\mathbf{X}_{Ti} = (\mathbf{1}, \mathbf{T}_i)$ where $\mathbf{T}_i$ is a vector of times, and $\mathbf{W}_i = (\mathbf{X}_{Ti}^t \mathbf{X}_{Ti})^{-1} \mathbf{X}_{Ti}^t$, then $\mathbf{W}_i \mathbf{Y}_i$ contains the intercept and slope of subject i 's regression of $\mathbf{Y}_i$ on $\mathbf{T}_i$. With continuous response data such as those discussed here, there are a number of possibly important sampling variables, Q_i , including those that may be multivariate. For the sake of this research, we will focus on Q_i that is a linear function of the response vector $\mathbf{Y}_i$. However, other summaries may also be used (e.g., standardized subject-specific slope that can account for a variable number of observations per subject, n_i).

When developing a sampling plan based on Q_i one needs to also consider the analysis that will account for the design. We will detail a form of conditional likelihood to account for the proposed sampling. However, Q_i may be a sufficient statistic for key parameters such that conditioning on the exact value of Q_i would eliminate information regarding parameters of interest. Therefore, we focus on a coarse sampling design that has the practical advantage of being easy to specify and implement, and has the statistical advantage of not leading to an over-conditioning when conducting analysis of the resulting biased sample.

In order to detail the proposed design first consider a univariate Q_i where we partition support into K regions: $\{(-\infty, k_1], \dots, (k_{K-1}, \infty)\} = \{R^1, R^2, \dots, R^K\}$. Within each R^k we randomly sample subjects for whom $X_{e,i}$ is then ascertained. The probability of being

sampled depends on all data (Y_i, X_i) only through stratum membership (or Q_i). If S_i is an indicator that X_{ei} was ascertained on subject i , then $pr(S_i = 1 | Y_i, X_i) = pr(S_i = 1 | Q_i \in R^k) = \pi(R^k)$. For a three stratum design, a reasonable outcome dependent sampling scheme may be given by: $\pi(R^1)$ if $Q_i < k_1$, $\pi(R^2)$, if $k_1 < Q_i < k_2$, and $\pi(R^3)$ if $Q_i > k_2$.

Figure 1 shows three designs based on subject-specific simple linear regressions of the response vector on a time variable. In all panels, black pluses correspond to sampled subjects and grey dots correspond to unsampled subjects. In a) and b), we show three region designs where sampling is based on the intercept and slope, respectively. In c), we show a design that samples jointly on the intercept and slope. In all three designs, there is a central region, R^2 in a) and b), and R^1 in c), that captures the relatively “typical” individuals. Outside this region, R^1 and R^3 in a) and b), and R^2 in c), are those subjects who are less “typical” in that their Q_i value tends towards the extremes of the distributions. We highlight the over-representation of the less typical subjects in the sample.

Intuitively, efficiency gains from ODS are obtained via improved observed response and exposure variation. Unlike univariate response data, with longitudinal data, there are two distinct sources of variation (i.e., between and within subjects), and the choice of Q_i should target the appropriate source of variation. For example, one would expect that sampling based on a subject-specific response mean or intercept would yield improved observed among-subject response variation thereby leading to efficiency improvements in time-invariant covariate coefficients. However, sampling based on a subject-specific slope (or variance) would likely lead to greater observed within-subject variation leading to improvements in estimates of time-varying covariate coefficients. Further, we may presume that sampling based on the subject-specific intercept and slope jointly could improve efficiency for a larger range of coefficients than sampling based on either one individually.

2.3 An Ascertainment Corrected Likelihood for the ODS Design

Because the ODS designs proposed here yield biased samples, naive likelihood analyses will not provide valid inference. By applying and/or extending standard missing data approaches (e.g., Rubin, 1976; Robins et al., 1994; Chen, 2004; Lawless, Kalbfleisch, and Wild, 1999) and ODS analysis approaches (Zhou et al., 2002, 2007); Weaver and Zhou, 2005; Neuhaus et al., 2002, 2006) there are a number of possible likelihood-based and semi-parametric estimation strategies that could be used in this setting. Here, we will discuss a relatively simple ascertainment corrected likelihood that conditions on having X_{ei} ascertained. Specifically, since $f(Y_i | X_i; \theta)$ is the multivariate density for subject i under random sampling from a population, the density for those who are included in the ODS is given by

$$\begin{aligned} f(Y_i | X_i, S_i=1; \theta) &= \pi(q_i) f(Y_i | X_i; \theta) \{pr(S_i=1 | X_i; \theta)\}^{-1} \\ &= \pi(q_i) f(Y_i | X_i; \theta) \left\{ \sum_{k=1}^K \pi(R^k) \int_{R^k} f(q_i | X_i; \theta) dq_i \right\}^{-1} \end{aligned} \quad (2)$$

where q_i is subject i 's observed value of the sampling variable Q_i , $\pi(q_i)$ is subject i 's sampling probability that is based on q_i , and $\pi(R^k)$ is the sampling probability for all Q_i values in the region, R^k . If a total of N_s subjects are selected into the ODS, the ascertainment corrected log-likelihood is given by,

$$l^C(\theta; \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{N_s} \left[l_i(\theta; \mathbf{Y}_i, \mathbf{X}_i) - \log \underbrace{\left\{ \sum_{k=1}^K \pi(R^k) \int_{R^k} f(q_i | \mathbf{X}_i; \theta) dq_i \right\}}_{AC_i} \right], \quad (3)$$

where AC_i is the ascertainment correction portion of the subject specific-contribution to the log-conditional likelihood. Carroll et al. (1995) and Lawless et al. (1999) refer to this conditional likelihood as the complete data likelihood since only subjects who have complete data are included in the analysis. It may be possible to improve efficiency by also exploiting the incomplete data from subjects in whom X_{ei} was not observed but this additional likelihood contribution requires integration over the distribution of X_{ei} , and such a covariate model is not required for the complete data likelihood. Whether such additional effort is warranted is the natural next question in our line of research.

For simplicity, we will focus on common summary statistics where $Q_i = \mathbf{W}_i \mathbf{Y}_i$ is a linear function of the response profile. Since $\mathbf{Y}_i | \mathbf{X}_i \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i)$, then the distribution of $Q_i | \mathbf{X}_i$ is also normal, $N(\boldsymbol{\mu}_{qi}, \Sigma_{qi})$ where $\boldsymbol{\mu}_{qi} = \mathbf{W}_i \boldsymbol{\mu}_i$ and $\Sigma_{qi} = \mathbf{W}_i \mathbf{V}_i \mathbf{W}_i^t$. If Q_i is univariate, $\boldsymbol{\mu}_{qi} = \mu_{qi}$ and $\Sigma_{qi} = \sigma_{qi}^2$, and we can write,

$$AC_i = \sum_{k=1}^K \pi(R^k) \left\{ F_{Q_i | \mathbf{X}_i}(k_k) - F_{Q_i | \mathbf{X}_i}(k_{k-1}) \right\} \equiv \sum_{k=1}^K \pi(R^k) \left\{ \Phi \left(\frac{k_k - \mu_{qi}}{\sigma_{qi}} \right) - \Phi \left(\frac{k_{k-1} - \mu_{qi}}{\sigma_{qi}} \right) \right\}$$

where $F_{Q_i | \mathbf{X}_i}(c)$ is the cumulative distribution function $pr(Q_i < c | \mathbf{X}_i)$ and $\Phi(c)$ is the standard normal cumulative distribution function.

2.4 Ascertainment Corrected Maximum Likelihood Estimation

For parameter estimation and inference, the ascertainment corrected conditional likelihood from equation 3 must be maximized with respect to the parameter $\boldsymbol{\theta}$. This can be done with a Newton-Raphson approach. The general form of the score equation is given by,

$\frac{\partial l_i^c}{\partial \boldsymbol{\theta}} = \frac{\partial l_i}{\partial \boldsymbol{\theta}} - \frac{\partial AC_i}{\partial \boldsymbol{\theta}} \cdot [AC_i]^{-1}$ where $\frac{\partial l_i}{\partial \boldsymbol{\theta}}$ is the derivative of the standard log-likelihood. For univariate Q_i , the specific score function for the ascertainment corrected likelihood with respect to the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ is given by the equations:

$$\frac{\partial l_i^c}{\partial \boldsymbol{\beta}} = \frac{\partial l_i}{\partial \boldsymbol{\beta}} + \left[\frac{1}{\sigma_{qi}} \frac{\partial \mu_{qi}}{\partial \boldsymbol{\beta}} \sum_{k=1}^{K-1} \left\{ \pi(R^{k+1}) - \pi(R^k) \right\} \cdot \varphi \left(\frac{k_k - \mu_{qi}}{\sigma_{qi}} \right) \right] \cdot [AC_i]^{-1} \quad (4)$$

for fixed effect parameters $\boldsymbol{\beta}$ and

$$\begin{aligned} \frac{\partial l_i^c}{\partial \alpha_m} = & \frac{\partial l_i}{\partial \alpha_m} + \left[\frac{1}{\sigma_{qi}^2} \frac{\partial \sigma_{qi}}{\partial \alpha_m} \sum_{k=1}^{K-1} \left\{ \pi(R^{k+1}) - \pi(R^k) \right\} (k_k - \mu_{qi}) \right. \\ & \left. \cdot \varphi \left(\frac{k_k - \mu_{qi}}{\sigma_{qi}} \right) \right] \cdot [AC_i]^{-1} \end{aligned} \quad (5)$$

for each of the variance component parameters in $\boldsymbol{\alpha} = (\sigma_0, \sigma_1, \rho, \sigma_e)$. Here, $\frac{\partial \mu_{qi}}{\partial \boldsymbol{\beta}} = \mathbf{W}_i \mathbf{X}_i \phi(c)$ is the standard normal density at the value c , and $\frac{\partial \sigma_{qi}}{\partial \alpha_m} = \mathbf{W}_i \frac{\partial \mathbf{V}_i}{\partial \alpha_m} \mathbf{W}_i^t$. We refer readers to McCulloch and Searle, Chapter 6 Appendix for the calculations of first and second partial derivatives of the log-likelihood for the mixed effects model with a random intercept and slope. Since all parameters in $\boldsymbol{\alpha}$ are subject to constraints e.g., variance components and

variances must be positive and ρ must fall within $[-1, 1]$, for our analyses, we transform α and estimate the following parameters $\alpha = (\alpha_0, \alpha_1, \alpha_\rho, \alpha_e) = (\log(\sigma_0), \log(\sigma_1), \log\{(1 + \rho)/(1 - \rho)\}, \log(\sigma_e))$. The use of conditional likelihood to address biased sampling designs has been discussed in, for example, Glidden and Liang (2002) and Clayton (2003). While analytical Hessian (observed information) calculations are possible, we calculated them by taking numerical derivatives of the score. We then inverted the observed information to obtain variance-covariance estimates.

3. Finite Sampling Operating Characteristics

We have described an ODS sampling design and ACML estimation strategy for parameter estimation and we now examine the operating characteristics in scenarios similar to those that might be observed in practice. Our study involves generating a cohort of $N=1,000$ subjects with longitudinal continuous response data and with n_i observations for subject $i \in 1, 2, \dots, N$. Furthermore, we examine relatively small clusters ($n_i = n = 5$), moderate sized cluster ($n_i = n = 11$), and unequal sized clusters $n_i \sim U(5, 17)$ (e.g. uniform from 5 to 17). We use a random intercept and slopes mixed effects model, and at the j^{th} observation time, the data generating population model is given by, $Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_g g_i + \beta_{gt} g_i t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}$ where g_i denotes a subject-level covariate, and t_{ij} denotes time. The time vector, $\mathbf{t}_i = \{t_{i1}, \dots, t_{in_i}\}$ contains equally spaced timepoints that range from -1 to 1 with fixed cluster sizes. Under the scenario with unequal sized clusters ($n_i \sim U(5, 17)$), we generated $\mathbf{t}_i$ based on 17 equally spaced values that ranged from -1 to 1 , and we then randomly selected a missing completely at random dropout time beyond which subsequent follow-up data were removed. The exposure, g_i , is assumed to be a relatively rare, binary, group variable with $pr(g_i = 1) = 0.1$. Random effects, $\mathbf{b}_i = (b_{0i}, b_{1i})^T \sim N(\mathbf{0}, \mathbf{D})$, where $\mathbf{D}$ is a covariance matrix with variance components (σ_0^2, σ_1^2) and with correlation parameter ρ that was set to 0. We also examined $\sigma_0 = \sigma_1 = 1$ and 2, but due to the similarity of the results, we only report results from the latter. The error term $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ with σ_e set to 2. In all simulation scenarios, the fixed effect parameter $(\beta_0, \beta_1, \beta_g, \beta_{gt})$ was set to $(10, -0.25, -0.75, 0.50)$.

We compare the validity and efficiency of several designs under the assumption that we can sub-sample approximately 250 individuals out of the original cohort of 1,000 subjects. One approach is to randomly sample 250 individuals and to analyze the data using standard maximum likelihood. The twelve ODS designs that we consider can be differentiated by; 1) Q_i ; 2) the cutpoints that define the sampling regions; and 3) the probability of being sampled within each region, $\pi(R^k)$. The Q_i are all defined by the subject-specific regression of the response vector $\mathbf{Y}_i$ on the time variable $\mathbf{T}_i$. Figure 1 shows the general designs. As discussed previously, we consider three values of Q_i : the subject-specific 1) intercept (*Int*); 2) slope (*Slp*); and 3) bivariate, intercept and slope (*Biv*). For all of *Int*, *Slp*, and *Biv*, we consider strata defined by a central region (R^2 for *Int* and *Slp* and R^1 for *Biv*) and outlying regions (R^1 and R^3 for *Int* and *Slp* and R^2 for *Biv*). We defined the cutpoints for the sampling regions by generating the population model on a large sample and then searching for the cutpoints that yield the desired number of subjects inside and outside the central regions. For *Int* and *Slp* the two cutpoints, k_1 and k_2 that define the three sampling regions are based on the percentiles (*pct*) of the population distribution of Q : $(k_1, k_2) = (Q^{100-pct}, Q^{pct})$. We consider *pct* = 90 and *pct* = 80 so that the central regions of the univariate Q_i contain on average 80 and 60 percent, respectively, of all simulated participants. Similarly, for the *Biv* designs, where $Q_i = (Q_{1i}, Q_{2i})$, we searched for the parameters of a quadrant so that eighty percent and sixty percent of participants fell within the central region defined by $(Q_1^{100-pct}, Q_1^{pct})$ and $(Q_2^{100-pct}, Q_2^{pct})$. In simulations with Q_1 independent of Q_2 , the values *pct* = 94.6 and 88.0 yielded central regions containing eighty and sixty percent of all subjects, respectively. When Q_1 and Q_2 were not independent (e.g., cluster sizes were unequal) *pct* was found via a

numerical search. From within each R^k , we then considered sampling in such a way that yielded 250 subjects, on average, in the ODS. With *Int* and *Slp* we sought to sample with probability so that, on average (100, 50, 100) and (75, 100, 75) are sampled from (R^1, R^2, R^3) . Similarly, with *Biv*, we sought to sample with probability so that, on average, (50, 200) and (100, 150) were sampled from (R^1, R^2) . We uniquely identify the twelve study designs with the notation $Q_i(N_1^s/N_1, N_2^s/N_2, N_3^s/N_3)$, where $Q_i \in \{Int, Slp, Biv\}$, N_k is the expected number of subjects from a cohort of 1,000 patients that fall in R^k and N_k^s is the expected number of sampled subjects from R^k .

For each scenario we used 1,000 replicates to characterize the validity and relative efficiency of the ACML procedure. We also examined the very simple inverse probability (of being sampled) weighted (IPWL) estimator based on a weighted likelihood with robust variance estimates (Zeger and Liang, 1986). This estimator has been well studied in the survey sampling and missing data literature and is known to be valid but in some circumstances can be inefficient.

3.1 Validity

We now examine the implications of conducting the ODS designs and either properly acknowledging the biased samples with the ACML approach or naively ignoring the design and performing analysis as if data were from a standard prospective study. Table 1 shows the average estimates and the coverage probabilities for the appropriate ACML approach and the naive approach when $n_i = n = 11$ with biases that are greater than or equal to 10% in bold. Results from $n_i = n = 5$ (not shown) were qualitatively similar. The properly specified ACML approach was nearly unbiased and yielded valid inference (i.e., 95 percent coverage) in all of the designs considered. For designs with $Q_i = Int$, the naive approach yielded biased parameter estimates on the time-invariant covariate coefficient β_g and on $\log(\sigma_0)$. For example, under the *Int*(100/100, 50/800/100/100) design the average estimates of β_g and $\log(\sigma_0)$ were -1.89 and 1.20 as opposed to the true values -0.75 and 0.69 , respectively. It is worth noting that we observe little to no bias on time-varying covariate coefficients. When sampling based on the subject specific slope, with the naive estimation approach, we observe bias in time-varying covariate coefficients but not in the time invariant covariate coefficients. For example with the *Slp*(100/100, 50/800/100/100) design, the average estimates of β_t , β_{gt} and $\log(\sigma_1)$ were -0.33 , 1.28 , and 1.24 as opposed to the true values of -0.25 , 0.50 and 0.69 , respectively. Because our longitudinal design was balanced and there was no correlation between time-varying and time-invariant covariates, we found that biased sampling only led to bias in naively estimated covariate coefficients that were related to the sampling variable, Q_i . However, when we sampled based on the intercept and slope with the *Biv* designs and subsequently ignored the design for analyses, we observed bias in all parameters. For example, under the *Biv*(50/800, 200/200) design, estimates of the parameter vector $(\beta_t, \beta_g, \beta_{gt}, \log(\sigma_0), \log(\sigma_1))$ were $(-0.31, -1.57, 1.06, 1.09, 1.13)$ which is highly biased for all true parameter values. While the results are not shown, with an imbalanced design, such as when $n_i \sim U(5, 17)$, we observed at least some bias in all target parameters when ignoring the study design under each of the *Int*, *Slp*, and *Biv* designs.

Table 2 shows empirical standard deviations and average estimated standard errors (square root of the average variance) for RS with ML estimation and for ODS with ACML and IPWL estimation. While the ACML approach yielded largely unbiased standard errors, in some circumstances, the IPWL approach could yield modestly biased estimates. This occurred to a greater degree in designs that over-sampled a proportionately greater number of subjects at the tails of the Q_i distribution. For example, with the *Biv*(50/800, 200/200) design, empirical standard deviations for β_g and β_{gt} were 0.65 and 0.68 , respectively, while the estimated standard errors were 0.59 and 0.62 . We believe this bias is a small sample

result and asymptotically the estimators are valid. Finally, the simple IPWL estimators that we studied were inefficient, in part, due to highly variable inverse probability weights. However, more complex and more efficient inverse probability weighted estimators such as those described in Robins et al. (1994) could be derived, but such estimators have not been examined here.

3.2 Relative Efficiency

The primary motivation for the proposed designs and analytical procedures involves presumed efficiency improvements over standard random sampling. Figure 2 shows the relative efficiency of a range of designs coupled with ACML estimation compared to random sampling of 250 participants. Relative efficiency was defined as 100 times the empirical variance of the RS design with ML estimation divided by the empirical variance of each of ODS design with ACML estimation. We show results for small and moderate sized clusters, $n_i = n = 5$ and $n_i = n = 11$, and for studies with a missing completely at random dropout mechanism $n_i \sim U(5, 17)$. As expected, we observed substantial efficiency improvements for coefficients whose corresponding covariate was associated with Q_i . For example, relative efficiency for β_0 and β_g was very high when sampling was based on the intercept (*Int* and *Biv* designs) and it was high for β_t and β_{gt} when sampling was based on the subject-specific slope (*Slp* and *Biv* designs). Sampling based on a variable that was effectively orthogonal to the target covariate, did not lead to any efficiency gains. Because the *Biv* designs sampled based on both the intercept and the slope, we observed efficiency improvements for all coefficients. The largest efficiency gains tended to be for the designs that sampled the greatest number of subjects at the extremes of the Q_i distributions. This occurred in the $Q_i(100/100, 50/800, 100/100)$ designs and the *Biv*(50/800, 200/200) designs. For the *Int*(100/100, 50/800, 100/100) design and with $n_i = n = 5$ efficiency improved by 126 percent for β_g estimates over random sampling. Similarly with the *Slp*(100/100, 50/800, 100/100) design and $n_i = n = 5$ efficiency improvements for β_t and β_{gt} estimates were 156 and 166 percent, respectively. The *Biv*(50/800, 200/200) design led to 106, 91, and 101 percent efficiency improvements over random sampling for estimates of β_g , β_t and β_{gt} , respectively. Even though it is not the most efficient design for any single parameter, the *Biv* design may be optimal if interested in both time-varying and time-invariant covariate coefficients.

4. The Childhood Asthma Management Program

We now return to the CAMP example where our interest is in the longitudinal profiles and the comparison of longitudinal profiles between children with and without a variant allele on locus 2 of the IL-13 cytokine. Previous reports have observed that lung growth in children with asthma is lower than those without asthma (Strunk et al. 2006). One of our goals is to examine, among the children with asthma, whether the presence of the variant allele may impact lung function and lung growth. The time-varying response in this study is the post-bronchodilator FEV_1 as a percent of the predicted value based on age, height, gender, and race. We refer to the response value FEV_1^{pp} . Even though in CAMP, single nucleotide polymorphism (SNP) data were available for all 930 used in our analyses, for illustration we consider the scenario where the SNP data would not be available when the substudy was conceived and they must be ascertained retrospectively. We also assume that substudy funding permits genotyping of only 250 children.

Table 3 shows the demographic characteristics of the CAMP children on whom genetic data were available. These children represent the “original cohort” from which we subsample. Children were enrolled at eight different sites and the median age was approximately nine years. Forty percent of the children were female and 32 percent were ethnicities other than

white. The variant G allele was present in eleven percent of children. The vast majority of children completed the four years of followup, and 5th, 50th, and 95th percentiles of FEV_1^{pp} were 83, 103, and 124, respectively.

From the original cohort, we subsampled approximately 250 for inclusion in each of 250 replicated substudies. Similar to Section 3, we examined a random sampling approach (RS) and several ODS designs. The sampling variable Q_i was based on the subject-specific regression of the response on years since enrollment (i.e., time). We consider for Q_i , the regression intercept (*Int*), the slope (*Slp*), and the bivariate (*Biv*) intercept and slope. In each case, we sampled inside and outside the central region with different probabilities. We considered central regions that contain 80 and 60 percent of all Q_i data. Thus, we define six study designs in a way that is similar to how we had done in Section 3. They are given by: *Int*(93/93, 65/744, 93/93), *Int*(93/186, 65/558, 93/186), *Slp*(93/93, 65/744, 93/93), *Int*(93/186, 65/544, 93/186), *Biv*(65/745, 185/185), and *Biv*(65/558, 185/372). With the univariate designs, we used the (10th, 90th) and the (20th, 80th) percentiles of Q_i for the value of the cutpoints (k_1, k_2). For the bivariate designs we search for the central regions and found, similar to the results from the simulations, that the region that falls inside the 5.3rd and 94.7th percentiles of the *Int* and *Slp* distributions contained 80 percent of all subjects and the region that falls inside the 12.0th and 88.0th percentiles contained 60 percent of all subjects. The results we report are based on the average of 250 replicates of conducting each substudy. On average 250 children were sampled for each substudy, except the RS design where exactly 250 children were sampled for each replication.

Table 4 shows the results from the proposed substudies. Of primary interest FEV_i^{pp} falls by approximately 0.3 to 0.5 per year among those without the variant G allele on locus 2 of the IL13 cytokine (main effect of study year). Those with the variant allele began the study (i.e. study year=0) with FEV_i^{pp} values that are expected to be 4 to 7 points lower than those without the variant allele (main effect of variant allele indicator). Finally, those with the variant allele then seemed to improve relative to those without it by approximately one point per year (positive the time by variant allele interaction). There exists substantial heterogeneity in subject-specific values of lung function and per year change in lung function as the estimate of the square root of the variance components are:

$\hat{\sigma}_0 \approx \exp(2.31)=10.1$ and $\hat{\sigma}_1 \approx \exp(0.80)=2.2$, respectively. Consistent with the simulation studies, we see that ODS designs can lead to improved precision for target parameters relative to random sampling. The *Int* designs tended to lead to lower standard errors than random sampling for the variant allele effect (e.g., 1.82 and 1.56 versus 2.24), and the *Slp* designs led to lower standard errors for the time effect in those without the variant alleles (0.14 and 0.12 versus 0.18) and for the difference in the time effect between those with and without the variant allele (0.43 and 0.36 versus 0.55). However, *Int* designs did not lead to efficiency improvements for the time and difference-in-time coefficients and *Slp* did not lead to efficiency improvements in the time-invariant, variant allele coefficient. The *Biv* design led to estimation precision improvements, relative to random sampling, in all three target coefficients. For both the variant allele coefficient and for the study year by variant allele interaction, random sampling would have led to inconclusive results. Most of the ODS designs led to statistically significant effects in both coefficients.

5. Discussion

We have proposed a class of ODS study designs and an ascertainment corrected maximum likelihood estimation strategy for subsampling from an existing longitudinal cohort. The proposed designs appear to be highly efficient relative to random sampling, and we show that the choice of the sampling variable Q_i dictates the regression coefficients for which

precision is improved. As a general rule, sampling on a summary statistic that is related to the target covariate will likely improve estimation precision and sampling more subjects towards the extremes of Q_i (i.e., the informative subjects) leads to larger efficiency gains. Based on the sampling variables considered here, we would suggest that if a researcher is interested in a time-varying covariate coefficient, sampling should be based on the subject-specific slope. Alternatively, if interested in a time-invariant covariate, we recommend sampling based on the subject specific intercept (or mean). Finally, if interested in a combination of time-invariant and time varying covariates, then sampling based on the joint intercept and slope distribution is attractive.

We have only considered a small subset of possible study designs and in a number of circumstances others designs may be more efficient. For example, we believe that in certain situations, sampling that depends on the Mahalanobis distance from the center of the bivariate intercept-slope distribution may, in some circumstances lead to greater efficiency than the designs proposed here. The key challenge associated with any proposed design is that we must be able to identify the distribution of $Q_i | X_i$ to calculate the corrected log-likelihood shown in equation 3. Our choice of Q_i , based on linear functions of Y_i , makes this calculation relatively simple due to properties of the normal distribution.

Though our design appears to be highly efficient relative to random sampling, we think it is important to highlight two possible limitations of the approach that warrant exploration. The first limitation is the use of complete case data rather than attempting to exploit information available on those subjects for whom X_{ei} is not ascertained. Such subjects have all other data (Y_i, X_{oi}), and based on the results of, for example, Kalbfleisch et al. (1999) and Weaver and Zhou (2005), we believe we may gain some efficiency by explicit use of these additional data. A second issue involves robustness to model misspecification. Although the linear mixed effects model has been shown to exhibit strong robustness properties to distributional misspecification our results may be sensitive to violations of normality for random effects and/or within-subject errors. For example, by inspection of equation 3, we can see that the normality assumption is required for the standard likelihood portion and the ascertainment correction component of the conditional likelihood.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded by the NHLBI grants R01 HL94786 and R01 HL079266. The work was conducted in part using the advanced computing resources of ACCRE at Vanderbilt University, Nashville, TN. The CAMP Trial and Continuation Study were supported by contracts N01-HR-16044, 16045, 16046, 16047, 16048, 16049, 16050, 16051, and 16052 with the NHLBI and GCRC grants M01RR00051, M01RR0099718-24, M01RR02719-14, and RR00036 from the NCRR. The CAMP Genetics Ancillary Study is supported by grants U01HL075419, U01HL65899, and P01HL083069 from the NHLBI.

References

- Anderson JA. Separate sample logistic discrimination. *Biometrika*. 1972; 59:19–35.
- Buzkova P, Lumley T. Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Stat Med*. 2009; 28:987–1003. [PubMed: 19035546]
- CAMP Research Group. The Childhood Asthma Management Program (CAMP): design, rationale, and methods. Childhood Asthma Management Program Research Group. *Control Clin Trials*. 1999; 20:91–120. [PubMed: 10027502]
- CAMP Research Group. Long-term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine*. 2000; 343 (15):1054–1063. [PubMed: 11027739]

- Carroll, R. Monographs on statistics and applied probability. Chapman & Hall/CRC; 2006.
Measurement error in nonlinear models: a modern perspective.
- Chen H. Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association*. 2004; 99:1176–1189.
- Clayton D. Conditional likelihood inference under complex ascertainment using data augmentation. *Biometrika*. 2003; 90:976–981.
- Glidden D, Liang KY. Ascertainment adjustment in complex diseases. *Genetic Epidemiology*. 2002; 23:201–208. [PubMed: 12384971]
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- Lawless JF, Kalbfleisch J, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society Series B-Statistical Methodology*. 1999; 61:413–438.
- Lin D, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*. 2001; 96:103–113.
- Lindstrom M, Bates D. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*. 1988; 83:1014–1022.
- Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Gelber R, Lipshultz S. Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics*. 2002; 58:621–630. [PubMed: 12229997]
- Lyon H, Lange C, Lake S, Silverman EK, Randolph AG, Kwiatkowski D, Raby BA, Lazarus R, Weiland KM, Laird N, Weiss ST. IL10 gene polymorphisms are associated with asthma phenotypes in children. *Genet Epidemiol*. 2004; 26:155–165. [PubMed: 14748015]
- McCulloch, CE.; Searle, SR. Generalized, Linear and Mixed Models. Wiley; New York: 2001.
- Neuhaus J, Scott AJ, Wild CJ. The analysis of retrospective family studies. *Biometrika*. 2002; 89:23–37.
- Neuhaus JM, Scott AJ, Wild CJ. Family-specific approaches to the analysis of case-control family data. *Biometrics*. 2006; 62:488–494. [PubMed: 16918913]
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–412.
- 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.
- Rubin DB. Inference and Missing Data. *Biometrika*. 1976; 63:581–590.
- 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. Outcome-dependent sampling from existing cohorts with longitudinal binary response data: study planning and analysis. *Biometrics*. 2011; 67:1583–1593. [PubMed: 21457191]
- Schildcrout JS, Mumford SL, Chen Z, Heagerty PJ, Rathouz PJ. Outcome-dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable. *Stat Med*. 2011
- Schildcrout JS, Rathouz PJ. Longitudinal studies of binary response data following case-control and stratified case-control sampling: design and analysis. *Biometrics*. 2010; 66:365–373. [PubMed: 19673861]
- Strunk RC. Childhood Asthma Management Program: lessons learned. *J Allergy Clin Immunol*. 2007; 119:36–42. [PubMed: 17125822]
- Strunk RC, Weiss ST, Yates KP, Tonascia J, Zeiger RS, Szeffler SJ. Mild to moderate asthma affects lung growth in children and adolescents. *J Allergy Clin Immunol*. 2006; 118:1040–1047. [PubMed: 17088127]
- 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.
- Zeger S, Liang K. Longitudinal data-analysis for discrete and continuous outcomes. *Biometrics*. 1986; 42:121–130. [PubMed: 3719049]

- 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 semi-parametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*. 2002; 58:413–421. [PubMed: 12071415]
- Zhou H, Wu Y, Liu Y, Cai J. Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. *Biostatistics*. 2011; 12:521–534. [PubMed: 21252082]

Figure 1.

Sampling regions for study designs based on subject-specific regressions of the response on time. Each point in the plot represents a single subjects' intercept and slope. Panels a) and b) show three sampling regions with sampling based on the univariate intercept and slope estimates, respectively. Panel c) shows a two region design with sampling based on the bivariate, slope and intercept. In all designs, the centralmost region pertains to subjects whose intercept and/or slope values fall towards the center of the respective distributions. The outer regions consist of those with intercept or slope values that tend to be more outlying and therefore potentially more informative. In the examples shown here, all central regions, contain the central 80 percent of all data points. Black pluses denote sampled subjects and grey dots denote unsampled subjects.

Figure 2.

Relative efficiency of the ODS study designs against randomly sampling 250 participants from the original cohort of 1000. The relative efficiency is defined as 100 times the empirical variance from random sampling with maximum likelihood estimation divided by the empirical variance based on each of the ODS designs with ACML estimation. We show results with Q_i defined by the subject-specific intercept (*Int*), the slope (*Slp*), and the bivariate intercept and slope (*Biv*). Squares correspond to designs $Q_i(100/100, 50/800, 100/100)$ when Q_i is the intercept or slope and to design *Biv*(50/800, 200/200) when sampling based on the intercept and slope. Triangles correspond to designs $Q_i(75/100, 100/800, 75/100)$ when Q_i is the intercept or slope and to design *Biv*(100/800, 150/200) when sampling jointly on the intercept and slope. Pluses correspond to designs $Q_i(100/200, 50/600, 100/200)$ when Q_i is the intercept or slope and to design *Biv*(50/600, 200/400) when sampling based on the intercept and slope. Finally, X's correspond to designs $Q_i(75/200, 100/600, 75/200)$ when Q_i is the intercept or slope and to design *Biv*(100/600, 150/400) when sampling based on the intercept and slope.

Table 1

Finite Sample Average Estimates and Coverage Probabilities: Results shown summarize 1000 replications with cluster sizes, $n_i = n = 11$, and with $\sigma_0 = \sigma_1 = 2$. Study designs are uniquely identified by the sampling variable, Q_i , the expected number of subject per stratum k , N_k , and the expected number sampled into the ODS from stratum k , N_k^s . The design column is given by $Q_i(N_1^s/N_1, N_2^s/N_2, N_3^s/N_3)$ for univariate Q_i (Int or Slp) and by $Q_i(N_1^s/N_1, N_2^s/N_2)$ for bivariate Q_i (biv). We show results for ascertainment corrected maximum likelihood (ACML) estimation and for naive, standard maximum likelihood estimation under the ODS designs. We also show results for standard random subsampling with appropriate maximum likelihood estimation (ML). Quantities in bold correspond to estimates that were biased by at least ten percent.

Design	Estimation	$\beta_0 = 10$	$\beta_0 = -0.25$	$\beta_0 = -0.75$	$\beta_0 = 0.50$	$\log(\sigma_0) = 0.69$	$\log(\sigma_1) = 0.69$	$\log\left(\frac{1-p}{1+p}\right) = 0$	$\log(\sigma_2) = 0.69$
RS	ML	9.99 [0.96]	-0.25 [0.96]	-0.73 [0.96]	0.50 [0.95]	0.69 [0.95]	0.68 [0.94]	0.00 [0.95]	0.69 [0.94]
Int(100/100,50/800,100/100)	ACML	10.00 [0.96]	-0.25 [0.94]	-0.76 [0.97]	0.48 [0.94]	0.69 [0.96]	0.68 [0.95]	0.00 [0.95]	0.69 [0.95]
	Naive	10.12 [0.92]	-0.25 [0.94]	-1.89 [0.61]	0.48 [0.94]	1.20 [0.00]	0.68 [0.95]	0.00 [0.95]	0.69 [0.95]
Int(100/200,50/600,100/200)	ACML	10.00 [0.95]	-0.24 [0.95]	-0.78 [0.95]	0.48 [0.95]	0.69 [0.96]	0.68 [0.94]	0.00 [0.94]	0.69 [0.94]
	Naive	10.06 [0.92]	-0.24 [0.95]	-1.32 [0.84]	0.48 [0.95]	0.99 [0.00]	0.69 [0.94]	0.00 [0.94]	0.69 [0.94]
Int(75/100,100/800,75/100)	ACML	10.00 [0.96]	-0.25 [0.95]	-0.76 [0.95]	0.49 [0.96]	0.69 [0.94]	0.68 [0.94]	0.00 [0.96]	0.69 [0.95]
	Naive	10.08 [0.93]	-0.25 [0.95]	-1.52 [0.76]	0.49 [0.95]	1.08 [0.00]	0.68 [0.94]	0.00 [0.96]	0.69 [0.95]
Int(75/200,100/600,75/200)	ACML	9.99 [0.96]	-0.25 [0.95]	-0.77 [0.95]	0.49 [0.95]	0.69 [0.94]	0.68 [0.95]	0.00 [0.95]	0.69 [0.96]
	Naive	10.02 [0.95]	-0.25 [0.95]	-1.05 [0.92]	0.49 [0.95]	0.87 [0.03]	0.68 [0.95]	0.00 [0.95]	0.69 [0.96]
Slp(100/100,50/800,100/100)	ACML	9.99 [0.95]	-0.25 [0.96]	-0.74 [0.96]	0.50 [0.96]	0.69 [0.96]	0.69 [0.94]	0.00 [0.94]	0.69 [0.95]
	Naive	9.99 [0.95]	-0.33 [0.93]	-0.74 [0.96]	1.28 [0.83]	0.69 [0.96]	1.24 [0.00]	0.00 [0.94]	0.69 [0.95]
Slp(100/200,50/600,100/200)	ACML	10.00 [0.94]	-0.24 [0.95]	-0.75 [0.95]	0.49 [0.94]	0.68 [0.95]	0.69 [0.94]	0.00 [0.96]	0.69 [0.95]
	Naive	10.00 [0.94]	-0.28 [0.94]	-0.75 [0.95]	0.84 [0.90]	0.69 [0.95]	1.02 [0.00]	0.00 [0.96]	0.69 [0.95]
Slp(75/100,100/800,75/100)	ACML	10.00 [0.95]	-0.25 [0.95]	-0.74 [0.96]	0.50 [0.94]	0.69 [0.94]	0.69 [0.95]	0.00 [0.95]	0.69 [0.95]
	Naive	10.00 [0.95]	-0.30 [0.93]	-0.74 [0.96]	1.02 [0.89]	0.69 [0.95]	1.12 [0.00]	0.00 [0.95]	0.69 [0.95]
Slp(75/200,100/600,75/200)	ACML	10.00 [0.94]	-0.24 [0.95]	-0.76 [0.95]	0.50 [0.96]	0.69 [0.94]	0.69 [0.96]	0.00 [0.94]	0.69 [0.94]
	Naive	10.00 [0.94]	-0.26 [0.95]	-0.76 [0.95]	0.69 [0.95]	0.69 [0.94]	0.88 [0.02]	0.00 [0.94]	0.69 [0.94]
Biv(50/800,200/200)	ACML	10.00 [0.95]	-0.25 [0.95]	-0.77 [0.95]	0.52 [0.95]	0.69 [0.94]	0.69 [0.95]	0.00 [0.94]	0.69 [0.95]
	Naive	10.09 [0.92]	-0.31 [0.93]	-1.57 [0.75]	1.06 [0.87]	1.09 [0.00]	1.13 [0.00]	0.01 [0.98]	0.69 [0.94]
Biv(50/600,200/400)	ACML	10.00 [0.95]	-0.25 [0.94]	-0.77 [0.95]	0.52 [0.95]	0.69 [0.95]	0.69 [0.95]	0.00 [0.95]	0.69 [0.95]
	Naive	10.04 [0.95]	-0.28 [0.94]	-1.19 [0.89]	0.79 [0.92]	0.94 [0.00]	0.96 [0.00]	0.01 [0.98]	0.69 [0.95]
Biv(100/800,150/200)	ACML	10.00 [0.95]	-0.25 [0.94]	-0.77 [0.95]	0.52 [0.95]	0.69 [0.94]	0.69 [0.95]	0.00 [0.94]	0.69 [0.95]
	Naive	10.06 [0.93]	-0.29 [0.93]	-1.32 [0.84]	0.89 [0.90]	0.99 [0.00]	1.02 [0.00]	0.01 [0.98]	0.69 [0.96]

Design	Estimation	$\beta_0 = 10$	$\beta_1 = -0.25$	$\beta_2 = -0.75$	$\beta_{gt} = 0.50$	$\log(\sigma_0) = 0.69$	$\log(\sigma_1) = 0.69$	$\log\left(\frac{1-\sigma}{1+\sigma}\right) = 0$	$\log(\sigma_t) = 0.69$
Biv(100/600,150/400)	ACML	10.00 [0.94]	-0.25 [0.95]	-0.77 [0.95]	0.50 [0.95]	0.69 [0.94]	0.69 [0.95]	0.00 [0.95]	0.69 [0.95]
	Naïve	10.03 [0.94]	-0.26 [0.95]	-0.98 [0.92]	0.65 [0.95]	0.82 [0.21]	0.84 [0.18]	0.01 [0.97]	0.69 [0.95]

Table 2

Empirical standard deviations/estimated standard errors: IPWL is based on inverse probability weighted likelihood with robust, sandwich variance estimates. Details describing the simulations are described in the Table 1 caption.

Design	Estimation	β_0	β_1	β_k	$\beta_{\theta d}$	$\log(\sigma_0)$	$\log(\sigma_1)$	$\log\left(\frac{1-\rho}{1+\rho}\right)$	$\log(\sigma_e)$
RS	ML	0.14 / 0.14	0.14 / 0.15	0.44 / 0.45	0.46 / 0.47	0.05 / 0.05	0.06 / 0.06	0.15 / 0.15	0.02 / 0.01
	ACML	0.08 / 0.09	0.15 / 0.15	0.29 / 0.30	0.45 / 0.46	0.04 / 0.04	0.05 / 0.06	0.09 / 0.09	0.01 / 0.01
	IPWL	0.17 / 0.17	0.27 / 0.26	0.58 / 0.54	0.88 / 0.73	0.05 / 0.05	0.10 / 0.10	0.19 / 0.18	0.03 / 0.03
Int(100/200,50/600,100/200)	ACML	0.11 / 0.10	0.15 / 0.15	0.35 / 0.35	0.47 / 0.47	0.04 / 0.04	0.06 / 0.06	0.11 / 0.11	0.01 / 0.01
	IPWL	0.13 / 0.13	0.21 / 0.21	0.44 / 0.43	0.66 / 0.61	0.05 / 0.05	0.08 / 0.08	0.14 / 0.13	0.02 / 0.02
Int(75/100,100/800,75/100)	ACML	0.09 / 0.10	0.15 / 0.15	0.34 / 0.32	0.46 / 0.46	0.04 / 0.04	0.06 / 0.06	0.10 / 0.10	0.01 / 0.01
	IPWL	0.13 / 0.13	0.19 / 0.19	0.46 / 0.43	0.60 / 0.58	0.04 / 0.04	0.07 / 0.07	0.14 / 0.14	0.02 / 0.02
Int(75/200,100/600,75/200)	ACML	0.11 / 0.12	0.14 / 0.15	0.39 / 0.39	0.47 / 0.47	0.04 / 0.04	0.06 / 0.06	0.13 / 0.13	0.01 / 0.01
	IPWL	0.12 / 0.12	0.16 / 0.16	0.41 / 0.41	0.51 / 0.50	0.05 / 0.04	0.06 / 0.06	0.13 / 0.13	0.02 / 0.02
Slp(100/100,50/800,100/100)	ACML	0.14 / 0.14	0.09 / 0.09	0.41 / 0.44	0.30 / 0.30	0.05 / 0.05	0.04 / 0.04	0.09 / 0.09	0.02 / 0.01
	IPWL	0.26 / 0.25	0.19 / 0.18	0.75 / 0.71	0.61 / 0.56	0.09 / 0.09	0.06 / 0.06	0.18 / 0.18	0.03 / 0.03
Slp(100/200,50/600,100/200)	ACML	0.14 / 0.14	0.11 / 0.11	0.44 / 0.44	0.38 / 0.36	0.05 / 0.05	0.05 / 0.05	0.11 / 0.11	0.01 / 0.01
	IPWL	0.19 / 0.20	0.14 / 0.13	0.61 / 0.57	0.47 / 0.44	0.07 / 0.07	0.06 / 0.06	0.13 / 0.13	0.02 / 0.02
Slp(75/100,100/800,75/100)	ACML	0.14 / 0.14	0.10 / 0.10	0.43 / 0.44	0.34 / 0.33	0.05 / 0.05	0.04 / 0.04	0.10 / 0.10	0.01 / 0.01
	IPWL	0.18 / 0.18	0.15 / 0.14	0.56 / 0.55	0.46 / 0.45	0.06 / 0.06	0.05 / 0.05	0.15 / 0.14	0.02 / 0.02
Slp(75/200,100/600,75/200)	ACML	0.14 / 0.14	0.12 / 0.13	0.45 / 0.45	0.41 / 0.41	0.05 / 0.05	0.05 / 0.05	0.13 / 0.12	0.02 / 0.01
	IPWL	0.15 / 0.15	0.13 / 0.13	0.49 / 0.47	0.42 / 0.42	0.05 / 0.05	0.05 / 0.05	0.13 / 0.13	0.02 / 0.02
Biv(50/800,200/200)	ACML	0.09 / 0.10	0.10 / 0.10	0.32 / 0.32	0.34 / 0.33	0.03 / 0.03	0.04 / 0.04	0.09 / 0.09	0.01 / 0.01
	IPWL	0.20 / 0.20	0.21 / 0.21	0.65 / 0.59	0.68 / 0.62	0.06 / 0.05	0.06 / 0.06	0.18 / 0.17	0.03 / 0.03
Biv(50/600,200/400)	ACML	0.11 / 0.11	0.12 / 0.12	0.37 / 0.36	0.39 / 0.38	0.04 / 0.04	0.05 / 0.05	0.11 / 0.11	0.01 / 0.01
	IPWL	0.14 / 0.15	0.16 / 0.16	0.51 / 0.46	0.51 / 0.48	0.05 / 0.05	0.05 / 0.05	0.13 / 0.13	0.02 / 0.02
Biv(100/800,150/200)	ACML	0.11 / 0.10	0.11 / 0.11	0.35 / 0.34	0.37 / 0.36	0.04 / 0.03	0.04 / 0.04	0.10 / 0.10	0.01 / 0.01
	IPWL	0.15 / 0.15	0.16 / 0.16	0.47 / 0.47	0.51 / 0.49	0.04 / 0.04	0.05 / 0.05	0.13 / 0.13	0.02 / 0.02
Biv(100/600,150/400)	ACML	0.12 / 0.12	0.13 / 0.13	0.40 / 0.40	0.42 / 0.42	0.04 / 0.04	0.05 / 0.05	0.13 / 0.12	0.01 / 0.01
	IPWL	0.13 / 0.13	0.14 / 0.14	0.42 / 0.42	0.45 / 0.44	0.05 / 0.04	0.05 / 0.05	0.13 / 0.13	0.02 / 0.02

Table 3

Demographic Characteristics of the CAMP Study Cohort. Continuous variables are summarized with the 5th, 25th, 50th, 75th, 95th percentiles, and categorical variables with proportions.

Variable	Summary
N	930
Albuquerque	97
Baltimore	122
Boston	114
Denver	131
San Diego	106
Seattle	127
Saint Louis	124
Toronto	109
Age at randomization (years)	5.68, 7.21, 8.88, 10.64, 12.52
Female	0.40
Race	
Black	0.14
Hispanic	0.09
Other	0.09
White	0.68
At least one G allele at Locus 2	0.11
Observations per subject	8, 10, 10, 10, 10
Follow-up time (years)	3.84, 3.9, 3.99, 4.03, 4.12
Post BD FEV_1 Percent Predicted	83, 95, 103, 111, 124

Table 4

Results from CAMP: We report results from 7 design by estimation procedure combinations: Random Subsampling (RS) with maximum likelihood (ML) estimation and ODS sampling based on Q_i where Q_i the subject-specific intercept (Int), slope (Slp), or the bivariate intercept and slope (Biv). From the original cohort of CAMP participants we conducted each substudy 250 times and results reported here are the average parameter estimates across replications and the square root of the average estimated variance (in parentheses).

N_k^s / N_k	(250/930)	(93/186, 64/558, 93/186)	(65/558, 185/372)	(93/93, 64/744, 93/93)	(65/745, 185/185)		
Q_i	RS	Int	Slp	Biv	Int	Slp	Biv
Intercept	102.42 (2.07)	102.41 (1.76)	102.56 (2.07)	102.3 (1.78)	102.95 (1.49)	104.39 (2.11)	102.86 (1.78)
Study year	-0.38 (0.18)	-0.43 (0.18)	-0.33 (0.14)	-0.31 (0.14)	-0.5 (0.18)	-0.31 (0.12)	-0.35 (0.13)
IL13 Variant Allele	-3.87 (2.24)	-4.68 (1.82)	-4.54 (2.14)	-4.76 (1.96)	-5.1 (1.56)	-6.95 (2.11)	-6.82 (1.81)
Year by Variant Allele	0.86 (0.55)	0.91 (0.54)	0.81 (0.43)	0.92 (0.46)	1.38 (0.54)	0.98 (0.36)	1.11 (0.4)
Female (versus male)	3.8 (1.29)	3.85 (1.09)	3.04 (1.3)	3.47 (1.11)	3.02 (0.92)	3.98 (1.31)	4.45 (1.03)
Age (per year)	-0.56 (0.32)	-0.82 (0.26)	-0.79 (0.32)	-0.71 (0.27)	-0.89 (0.23)	-1.24 (0.32)	-0.78 (0.24)
Study Medication (vs Budesonide)							
Nedocromil	-2.53 (1.62)	-2.74 (1.37)	-2.72 (1.65)	-2.6 (1.38)	-2.26 (1.15)	-2.71 (1.69)	-1.2 (1.33)
Budesonide placebo	-2.19 (1.82)	-2.1 (1.51)	-1.52 (1.87)	-1.62 (1.57)	-1.18 (1.27)	-2.93 (1.86)	-0.81 (1.43)
Nedocromil placebo	-2.55 (1.78)	-3.02 (1.49)	-2.63 (1.78)	-2.51 (1.56)	-2.29 (1.31)	-2.86 (1.84)	-2.39 (1.46)
Study site (vs. Baltimore)							
Albuquerque	4.14 (2.59)	4.43 (2.21)	4.68 (2.51)	5.13 (2.2)	4.85 (1.98)	3.14 (2.58)	4.71 (2.12)
Boston	0.93 (2.49)	1.09 (2.1)	0.54 (2.55)	0.82 (2.16)	0.76 (1.78)	-1.53 (2.4)	-1.77 (2.06)
Denver	2.9 (2.39)	3.04 (1.98)	3.69 (2.33)	3.42 (2.00)	2.26 (1.69)	2.47 (2.33)	3.04 (1.89)
San Diego	-2.85 (2.53)	-2.37 (2.14)	-1.75 (2.49)	-1.45 (2.19)	-0.9 (1.86)	-2.97 (2.53)	-0.9 (2.01)
Seattle	0.03 (2.41)	-0.13 (2.06)	-0.42 (2.38)	-0.15 (2.02)	-0.43 (1.7)	-1.82 (2.43)	-2.3 (1.95)
St. Louis	3.87 (2.42)	4.42 (2.12)	3.57 (2.43)	4 (2.07)	2.67 (1.74)	3 (2.34)	3.5 (2.04)
Toronto	4.17 (2.51)	4.38 (2.24)	4.57 (2.55)	4.66 (2.19)	4.01 (1.88)	5.02 (2.59)	4.04 (2.12)
Variance components							
$\log(\sigma_i)$	2.31 (0.05)	2.32 (0.04)	2.3 (0.05)	2.3 (0.04)	2.32 (0.03)	2.29 (0.05)	2.31 (0.04)
$\log(\sigma_i)$	0.8 (0.06)	0.84 (0.06)	0.79 (0.06)	0.76 (0.06)	0.9 (0.06)	0.76 (0.05)	0.77 (0.05)
$\log((1 + \rho)/(1 - \rho))$	-0.72 (0.17)	-0.7 (0.13)	-0.65 (0.14)	-0.68 (0.13)	-0.72 (0.11)	-0.52 (0.13)	-0.72 (0.11)