Response-Dependent Sampling with Clustered and Longitudinal Data

Michael A. McIsaac and Richard J. Cook

Abstract Prospective cohort studies typically involve repeated assessment of individuals to determine whether they have a particular health condition. The usual goal in such studies is to relate the presence of the condition to disease markers or exposure variables. Disease markers are often too difficult or costly to measure for all individuals in a sample. In such settings, two- and multi-phase sampling designs are routinely adopted to enable researchers to select individuals on whom these expensive markers are to be assessed. In this article we review the rationale and format of two-phase sampling designs in retrospective and cross-sectional studies. We then develop frameworks for multi-phase designs in the context of studies with clustered or longitudinal responses. Model-based and semi-parametric methods are discussed for estimation and inference.

1 Introduction

Two-phase sampling designs have proven useful in epidemiology for ensuring efficient use of resources when estimating the effect of expensive or otherwise difficult to measure exposure variables on a response. Under such designs, a regression model is often specified with a binary response indicating disease status and a covariate vector recording the exposure variable of interest along with possible auxiliary covariates. The first phase of sampling generates data on the response and auxiliary covariates. A sub-sample of these individuals is chosen at a second phase of sampling, and the expensive exposure variable is measured for these individuals. Viewed as a whole, the full sample features missing exposure data in individuals

M.A. McIsaac (✉) • R.J. Cook
Department of Statistics and Actuarial Science, University of Waterloo,
Waterloo, Ontario, Canada, N2L 3G1
e-mail: mamcisaa@uwaterloo.ca; rjcook@uwaterloo.ca

B.C. Sutradhar (ed.), *ISS-2012 Proceedings Volume On Longitudinal Data Analysis Subject to Measurement Errors, Missing Values, and/or Outliers*, Lecture Notes in Statistics 211,
DOI 10.1007/978-1-4614-6871-4_8, © Springer Science+Business Media New York 2013

selected in phase I but not selected in the phase II sub-sample, with the missing data mechanism determined by the nature of the phase-II sampling probabilities.

There is a wide range of statistical approaches for regression with incomplete covariate data, including methods based on maximum likelihood (Lawless et al. 1999), mean score equations (Reilly and Pepe 1995; Reilly 1996), inverse probability weighted estimating functions, and augmented inverse probability weighted estimating functions (Robins et al. 1994; Tsiatis 2006). These approaches differ in the nature of the assumptions required and the extent to which data from individuals with incomplete exposure data are utilized. Maximum likelihood, while potentially optimally efficient, requires one to model the distribution of the exposure variable given any auxiliary variables, and misspecification of this model can lead to an inconsistent estimator (Horton and Laird 2001). The mean score method involves specification of unbiased estimating functions by nonparametrically estimating the conditional distribution of the exposure variable given the response and auxiliary variables based on the phase-II sample (Reilly and Pepe 1995). In their simplest form, inverse probability weighted estimating equations restrict attention to individuals in the phase-II sample and hence do not require modelling of the covariate distribution. The resulting estimates are consistent provided the weights are correctly specified, but they are typically less efficient than maximum likelihood estimates (Lawless et al. 1999). Augmented inverse probability weighted estimating equations aim to improve efficiency by exploiting information in the individuals who only provide information in the phase-I sample (Robins et al. 1994; Tsiatis 2006).

When planning studies, the challenge is to specify the phase-II selection model which will lead to the most efficient estimators of the parameters of interest; this is typically the coefficient of the exposure variable. To do this one must adopt a response model and a framework for inference which accommodate the incomplete exposure data. Factors influencing the choice of the framework for inference include the kinds of assumptions one is willing to make, the degree of importance placed on robustness, and efficiency. Given any particular framework, the asymptotic distribution of the resulting estimators is then required to inform the design (i.e. specification of the phase-II sampling probabilities).

Much of the work to date on two-phase designs involves univariate outcomes reflecting disease status. The purpose of this article is to consider statistical issues in two-phase designs with more complex disease outcomes, motivated by our involvement in the following two studies.

Example 1 (A Study of Genetic Risk in Psoriatic Arthritis). The Centre for Prognosis Studies in the Rheumatic Diseases maintains a clinical registry of patients at the Toronto Western Hospital with psoriatic arthritis. Patients have been recruited and followed since its inception in 1976 and it is now the largest cohort of patients with PsA in the world. Upon entry to the clinic patients undergo a detailed clinical and radiological examination and provide serum samples which are subsequently stored. Follow-up clinical and radiological assessments are scheduled annually and biannually, respectively, in order to track changes in joint damage. Disease progression can be modelled in a number of ways including the development of newly damaged joints (Sutradhar and Cook, 2009), the involvement of new types of joints (Tolusso and Cook 2009; Chandran et al. 2010), and the onset of a particular condition. These approaches, however, involve composite outcomes because they aggregate information over multiple joints. We consider analyses based on models for the onset of damage in the sacroiliac (SI) joints, which signals the onset of spondyloarthritis. Damage of the SI joints is determined by radiological examination with the extent of damage in each joint graded using a standardized scale (Rahman et al. 1998). Serum biomarkers and genetic factors can play important roles in identifying patients at high risk for developing psoriatic spondyloarthritis (Rahman et al. 1998), and as a consequence, biomarker studies are of considerable importance.

We consider data from patients from the first assessment at which serum samples are taken which can be used for genetic testing. We restrict attention to individuals who have not experienced damage in their sacroiliac joints as of this assessment and a clustered (paired) response is based on the onset of damage in the left and right sacroiliac joints between the baseline and a follow-up assessment. The candidate genetic risk factor in this setting is the human leukocyte antigen B27, a factor known to be associated with progression of other diseases involving connective tissue and joints, and the auxiliary variable is a marker of inflammation called C-reactive protein (CRP) (del Rincon et al. 2003). Genetic typing is costly and it is desirable to carry this out for a subset of individuals in the cohort.

Example 2 (The Canadian Longitudinal Study of Aging). The Canadian Longitudinal Study on Aging (CLSA) involves the establishment of a pan-Canadian cohort to enable estimation of the incidence rates of several chronic diseases and to study associated risk factors. It involves 50,000 individuals aged 45 to 85 years old who are to be followed for 20 years or until the time of death. All participants in the CLSA will provide some information to the study, while a subset of 30,000 will be chosen for additional, in-depth examination. This sub-cohort will undergo a more intensive clinical examination, provide imaging data, and give biological specimens every three years; specimens will be stored in biobanks in a controlled environment to facilitate subsequent testing. Thus the biobank will serve as a valuable resource for affiliated investigators to study risk factors predictive of disease onset and progression. Samples will be too expensive to process for all 30,000 individuals in the cohort undergoing intensive follow-up, so it will be of central importance to determine how individuals should be selected for testing of stored specimens (Raina et al. 2009). We therefore explore the extension of the two-phase sampling problem to longitudinal data. Since interest lies in the onset of disease, we focus on transitional models and formulate the exposure effects on transition probabilities. We study various designs for sampling and analysis to investigate how optimal selection procedures can be derived at a particular time point given the available partial histories. Specifically, we examine the improved precision in estimation that can result when more information is used in deriving optimal selection probabilities.

The remainder of this article is organized as follows. Notation is defined and the format of two-phase response-dependent sampling schemes is described in Sect. 2. In Sect. 3 we consider the setting of clustered responses with cluster-level exposure and auxiliary variables. Marginal models (Liang and Zeger 1986) are adopted in this setting with analysis frameworks based on maximum likelihood, mean score estimating functions, and inverse probability weighted pseudolikelihood. In Sect. 4 we give a framework for two-phase designs in longitudinal studies where interest lies in modelling the effect of an exposure variable on the onset of disease under a first-order Markov model. Asymptotic theory and optimal designs are provided for each setting. Concluding remarks and topics for further research are given in Sect. 5.

2 Response-dependent Sampling with Correlated Data

2.1 Notation and Study Design

Two-phase sampling has been widely used to enhance precision of estimators of key parameters with resource constraints (Chatterjee et al. 2003; Pickles et al. 1995; Whittemore and Halpern 1997). This sampling framework is particularly appealing whenever the measurement of a covariate of central importance incurs considerable cost relative to the cost of associated auxiliary variables. Two-phase sampling involves the collection of outcome and inexpensive auxiliary data in a large phase-I sample, which is exploited to determine how individuals should be selected into a phase-II subsample for measurement of the expensive covariate (Reilly and Pepe 1995; Zhao et al. 2009). The efficiency gain that comes from such a two-phase sampling framework depends on the parameter of interest, the method of analysis, and the way in which the phase-I data are exploited in the design of the phase-II selection probabilities (Reilly 1996).

We begin with a discussion of likelihood-based inference which requires full model specification but enables optimal efficiency. To cover the case of clustered and longitudinal data simultaneously, we adopt a general formulation whereby $Y_i = (Y_{i1}, \dots, Y_{iK})'$ denotes a $K \times 1$ response vector for individual i; we let X_i and V_i denote the expensive exposure variable and the auxiliary variable, respectively. Let $f(Y_i|X_i, V_i; \beta)$ denote the conditional joint density or mass function for Y_i given (X_i, V_i) indexed by a $p \times 1$ parameter β. Let $g(X_i|V_i; \alpha)$ denote the conditional distribution of $X_i|V_i$ indexed by a $q \times 1$ parameter α and let the $r \times 1$ parameter γ index the marginal distribution of V. The random variables are governed by the joint model $f(Y, X, V; \beta, \alpha, \gamma) = f(Y|X, V; \beta)g(X|V; \alpha)h(V; \gamma)$, but (α, γ) are nuisance parameters which are routinely eliminated by conditioning on (X, V) when data are complete.

In a two-phase study, $\{(Y_i, V_i), i = 1, 2, \dots, N\}$ are observed for all N individuals selected in the phase-I sample and X_i is observed in the n individuals selected for inclusion in the phase-II sample. If $R_i = I(X_i$ is observed $)$, then selection into the phase-II sample is governed by the phase-II selection probabilities $\pi(Y,V;\delta) = P(R=1|Y,V;\delta)$, where δ indexes this distribution. Note that within this two-phase sampling framework, we consider missingness by design, so we can be confident that data are *missing at random* (MAR)—i.e. $P(R=1|Y,X,V;\delta) = P(R=1|Y,V;\delta)$ (Little and Rubin 2002). If the phase-II selection probabilities do not exploit the phase-I data—i.e. $P(R=1|Y,V;\delta) = P(R=1;\delta)$—then individuals are selected for the phase-II sample by simple random sampling and the expensive exposure variable will be *missing completely at random* (MCAR). Phase-two selection probabilities which exploit phase-I data can result in more efficient estimators.

2.2 Methods of Analysis

A variety of frameworks are available for the analyses of clustered data (Neuhaus 1992). Mixed-effect models (Laird and Ware 1982; Stiratelli et al. 1984) are effective when one wishes to assess the effects of within-cluster covariates. These models account for the dependence of responses within clusters by introducing unobservable, cluster-specific latent variables. When one wishes to explore the effects of cluster-level covariates on marginal means, analyses are often more naturally carried out via population-average approaches which may involve full model specification (Heagerty and Zeger 2000; Heagerty 2002); first order generalized estimating equations can also be adopted (Liang and Zeger 1986) or second order generalized estimating equations could be used, the latter being most often considered for clustered binary responses (Prentice 1988; Zhao and Prentice 1990). Autoregressive models are appropriate when response data arise serially and it is of interest to determine how changes occur over time (Zeng and Cook 2007; Sutradhar 2008). These methods of analyses can be extended in different ways to account for data which are incomplete (Lawless et al. 1999; Robins et al. 1995; Troxel et al. 1997).

We consider three likelihood-based methods for estimation of regression coefficients in marginal mean models and conditional means when covariate data are incomplete due to a MAR mechanism.

2.2.1 Maximum Likelihood

The full likelihood for these data is

$$\begin{aligned} L_F(\beta,\alpha,\gamma,\delta) = \prod_{i=1}^{N} & \left[f(Y_i,X_i,V_i;\beta,\alpha,\gamma)\; P(R_i=1|Y_i,V_i;\delta)\right]^{R_i} \\ & \times \left[f(Y_i,V_i;\beta,\alpha,\gamma)\; P(R_i=0|Y_i,V_i;\delta)\right]^{1-R_i}. \end{aligned}$$

One may restrict attention to the partial likelihood

$$L(\beta,\alpha,\gamma)=\prod_{i=1}^{N}\left[f(Y_i,X_i,V_i;\beta,\alpha,\gamma)\right]^{R_i}\left[f(Y_i,V_i;\beta,\alpha,\gamma)\right]^{1-R_i}$$

provided δ is functionally independent of $(\beta',\alpha',\gamma')'$. In the special case that data are complete, orthogonality of the parameters enables focus on the partial likelihood

$$L(\beta)=\prod_{i=1}^{N}f(Y_i|X_i,V_i;\beta) \tag{1}$$

(Breslow and Chatterjee 1999; Lawless et al. 1999). More generally however, if X_i is not observed for some clusters and the missing data mechanism is MAR, then the observed data partial likelihood is

$$L(\theta)=\prod_{i=1}^{N}\left[f(Y_i|X_i,V_i;\beta)g(X_i|V_i;\alpha)\right]^{R_i}\left[E_{X|V}\{f(Y_i|\mathrm{X},V_i;\beta)\}\right]^{1-R_i}, \tag{2}$$

where it can be seen that estimation of the parameters of interest, β, must occur jointly with the estimation of the nuisance parameter α in $\theta=(\beta',\alpha')'$.

Parameter estimates can be found by maximizing the likelihood in (2) directly, or by implementing an EM algorithm (Dempster et al. 1977) and iteratively maximizing the complete-data likelihood

$$L_c(\theta)=\prod_{i=1}^{N}[f(Y_i|X_i,V_i;\beta)g(X_i|V_i;\alpha)]^{R_i}[f(Y_i|X_i,V_i;\beta)g(X_i|V_i;\alpha)]^{1-R_i} \tag{3}$$

(Little and Rubin 2002). The expectation step involves computing $Q(\theta;\theta^k)=E_{X|Y,V}[\log L_c(\theta);\theta^k]$, where θ^k is the estimate of θ at the kth iterations and $E_{X|Y,V}[\log L_c(\theta);\theta^k]$ is

$$\sum_{i=1}^{N}\Bigg\{R_i\left[\log\ f(Y_i|X_i,V_i;\beta)+\log\ g(X_i|V_i;\alpha)\right]$$
$$+\ (1-R_i)\left[E_{X|Y,V}\{\log\ f(Y_i|X,V_i;\beta);\theta^k\}+E_{X|Y,V}\{\log\ g(X|V_i;\alpha);\theta^k\}\right]\Bigg\}.$$

The maximization step yields updated estimates $\theta^{(k+1)}$ obtained by solving

$$\frac{\partial Q(\theta;\theta^k)}{\partial\beta}=\sum_{i=1}^{N}\left\{R_iU_\beta(Y_i|X_i,V_i)+(1-R_i)E_{X|Y,V}[U_\beta(Y_i|X,V_i);\theta^k]\right\}=0 \tag{4}$$

and

$$\frac{\partial Q(\theta;\theta^k)}{\partial\alpha}=\sum_{i=1}^{N}\left\{R_iU_\alpha(X_i|V_i)+(1-R_i)E_{X|Y,V}[U_\alpha(X|V_i);\theta^k]\right\}=0,$$

where $U_\alpha(X_i|V_i)=\partial\log g(X_i|V_i;\alpha)/\partial\alpha$, and $U_\beta(Y_i|X_i,V_i)=\partial\log f(Y_i|X_i,V_i;\beta)/\partial\beta$.

Note that this method simultaneously estimates β and α and hence if $X|V$ is not correctly modelled, estimates of β will be inconsistent. We therefore consider alternative methods of analysis which, while motivated by the likelihood approach, do not require specification of the model for $X|V$. These pseudolikelihood approaches are potentially less efficient, but can provide consistent estimators of β without making any model assumptions about the covariate distributions.

2.2.2 The Mean Score Method

Each step in the iterative EM procedure involves using (4) to update β by estimating the conditional expectation of the pseudoscore function $U_\beta(Y|X,V)$ for individuals with incomplete data. This expectation can alternatively be estimated empirically in a single step (Lawless et al. 1999) rendering the so-called mean score equation of Reilly and Pepe (1995):

$$\overline{U}(\beta) = \sum_{i=1}^{N} \left\{ R_i U_\beta(Y_i|X_i,V_i) + (1-R_i)\hat{E}_{X|Y,V}[U_\beta(Y_i|X,V_i)] \right\} = 0. \qquad (5)$$

The problem then reduces to obtaining a robust nonparametric estimate of $g(X|Y,V)$ in order to compute $\hat{E}_{X|Y,V}(\cdot)$. When data are MAR and (Y,V) is discrete, the conditional distribution can be consistently estimated nonparametrically using the phase-II sample since $g(X|Y,V,R=1) = g(X|Y,V)$.

2.2.3 Weighted Pseudolikelihood

Recall that with complete data on all individuals we would want to maximize the likelihood function (1) or, equivalently, solve the score equations

$$U(\beta) = \sum_{i=1}^{N} U_\beta(Y_i|X_i,V_i) = \sum_{i=1}^{N} \partial \log f(Y_i|X_i,V_i;\beta)/\partial\beta = 0. \qquad (6)$$

When the data are incomplete, rather than making auxiliary distributional assumptions, we may wish to restrict attention to individuals who provide complete information. Such complete-case estimators often induce bias when data are not MCAR (Little and Rubin 2002), but if contributions to (6) are weighted by the inverse of the probability X_i is observed, the resultant estimators will be consistent (Lawless et al. 1999; Robins et al. 1994). That is, we can maximize the weighted log-pseudolikelihood or, equivalently, solve the weighted pseudoscore equations

$$\overline{\overline{U}}(\beta) = \sum_{i=1}^{N} \overline{\overline{U}}_i(\beta) = \sum_{i=1}^{N} \frac{R_i}{\pi(Y_i,V_i;\delta)} U_\beta(Y_i|X_i,V_i) = 0. \qquad (7)$$

Solving $\overline{\overline{U}}(\beta)=0$ yields a consistent estimator of β since (7) is an unbiased estimating function. To see this, we take the expectation of a single term in the sum and drop the subscript i for convenience to get

$$E_{Y|X,V}\left[\frac{E_{R|Y,X}\{R\}}{\pi(Y,V;\delta)}U_{\beta}(Y|X,V)\right]=E_{Y|X,V}\left[U_{\beta}(Y|X,V)\right]=0\,,$$

since $R\perp X|(Y,V)$ if X is MAR and $U_{\beta}(Y|X,V)$ is an unbiased estimating function.

We now turn our attention to the particular problems of optimal two-phase design with clustered and longitudinal data and restrict attention to the case of binary responses.

3 Response-dependent Sampling with Clustered Binary Data

3.1 The Response Model for Clustered Data

Let $Y_i=(Y_{i1},Y_{i2})'$ denote the bivariate binary response for cluster i, and let X_i and V_i be the univariate expensive and auxiliary covariates, respectively, defined at the cluster level (i.e. all subjects in a given cluster have the same values of these covariates). In the context of the study from the University of Toronto Psoriatic Arthritis Clinic, the responses correspond to the status of the left and right sacroiliac joints. The expensive covariate represents the human leukocyte antigen (HLA) marker B27 and the auxiliary variable is the inexpensive marker of inflammation, CRP, measured at the baseline visit. We consider a regression model for the marginal mean and let $\mu_{ij}=E[Y_{ij}|X_i,V_i]=P(Y_{ij}=1|X_i,V_i)$. Specifically we adopt the logistic model

$$\text{logit }\ \mu_{ij}=\beta_0+\beta_x X_i+\beta_v V_i, \tag{8}$$

where the covariates are assumed to have a common affect on both responses. We adopt the model of Lipsitz et al. (1991) and so account for the association between Y_{i1} and Y_{i2} given (X_i,V_i) via a common conditional odds ratio. That is, we let $\mu_{ikl}=P(Y_{i1}=k,Y_{i2}=l|X_i,V_i;\beta)$, where $\beta=(\beta_0,\beta_x,\beta_v,\psi)'$, with

$$\psi=\frac{P(Y_{i1}=1,Y_{i2}=1|X_i,V_i)/P(Y_{i1}=0,Y_{i2}=1|X_i,V_i)}{P(Y_{i1}=1,Y_{i2}=0|X_i,V_i)/P(Y_{i1}=0,Y_{i2}=0|X_i,V_i)}=\frac{\mu_{i11}/\mu_{i01}}{\mu_{i10}/\mu_{i00}}$$

the odds of subunit 1 in cluster i responding given X_i and V_i when subunit 2 responds, versus the respective odds when subunit 2 doesn't respond, Then

$$P(Y_{i1}=1,Y_{i2}=1|X_i,V_i;\beta)=\begin{cases}\frac{c_i-[c_i^2-4\psi(\psi-1)\mu_{i1}\mu_{i2}]^{1/2}}{2(\psi-1)} & \text{if } \psi\neq 1\\ \mu_{i1}\mu_{i2} & \text{if } \psi=1,\end{cases}$$

where $c_i = 1 - (1 - \psi)(\mu_{i1} + \mu_{i2})$. The marginal means and the odds ratio completely specify the bivariate distribution of the clustered binary responses. We consider binary covariates X and V which arise so that

$$\operatorname{logit} P(X_i = 1 | V_i; \alpha) = \alpha_0 + \alpha_v V_i$$

and

$$\operatorname{logit} P(V_i = 1; \gamma) = \gamma.$$

The discrete nature of the covariates (X, V) means there is no issue of misspecification in this part of the model.

3.2 The Selection Model

We specify the second-phase sampling design for these bivariate data through the choice of selection parameters δ in the probabilities $\pi(Y_i, V_i; \delta) = P(R_i = 1 | Y_i, V_i; \delta)$, where we consider the selection model

$$\operatorname{logit} \pi(Y_i, V_i; \delta) = \delta_0 + \delta_1 Y_{i1} + \delta_2 Y_{i2} + \delta_3 V_i + \delta_4 Y_{i1} Y_{i2} + \delta_5 Y_{i1} V_i + \delta_6 Y_{i2} V_i + \delta_7 Y_{i1} Y_{i2} V_i.$$

Note that since the covariate V and the responses Y_1 and Y_2 are binary, the use of this saturated selection model is equivalent to specifying stratum-specific sampling probabilities which indicate the selection probabilities that should be used within each of the eight strata defined by the phase-I data (Y_1, Y_2, V).

3.3 Mean Score Method with Discrete Phase-One Data

When both Y and V are discrete variables and $g(X|Y,V) = g(X|Y,V,R=1)$, a natural estimate of the conditional distribution $g(X|Y,V)$ is

$$\hat{g}(X|Y,V) = \frac{n^{(1)}_{X,Y,V}}{n^{(1)}_{Y,V}},$$

where $n^{(1)}_{Y,X,V} = \sum_{i:R_i=1} I(Y_i = Y, X_i = X, V_i = V)$ and $n^{(1)}_{Y,V} = \sum_{i:R_i=1} I(Y_i = Y, V_i = V)$. The conditional expectation of the pseudoscore is then estimated as

$$\hat{E}_{X|Y,V}[U_\beta(Y_j|X,V_j)] = \sum_x U_\beta(Y_j|X,V_j) \frac{n^{(1)}_{Y_j,X,V_j}}{n^{(1)}_{Y_j,V_j}}$$

so

$$\sum_{j:R_j=0} \hat{E}_{X|Y,V}[U_\beta(Y_j|X,V_j)] = \sum_{i:R_i=1} U_\beta(Y_j|X_i,V_j)\frac{n^{(0)}_{Y_j,V_j}}{n^{(1)}_{Y_j,V_j}},$$

where $n^{(0)}_{Y,V} = \sum_{i:R_i=0} I(Y_i = Y, V_i = V)$. Therefore the mean score estimating equations (5) reduce to

$$\overline{U}(\beta) = \sum_{i=1}^{N} R_i \left(1 + \frac{n^{(0)}_{Y_i,V_i}}{n^{(1)}_{Y_i,V_i}}\right) U_\beta(Y_i|X_i,V_i) = 0,$$

which can be seen to be a weighted pseudolikelihood approach (7) where the selection probabilities are estimated empirically using

$$\pi(Y,V;\hat{\delta}) = \left(1 + \frac{n^{(0)}_{Y,V}}{n^{(1)}_{Y,V}}\right)^{-1} = \frac{\sum_i I(R_i = 1, Y_i = Y, V_i = V)}{\sum_i I(Y_i = Y, V_i = V)}$$

(Lawless et al. 1999; Zhao 2005). The weighted pseudolikelihood approach will remain consistent if known weights are replaced with consistently estimated weights, as is done here with the mean score method. In fact, it is often advantageous to utilize estimated weights even when the true weights are known since the estimation of weights in (5) incorporates information from all individuals available at the first phase of sampling, while (7) only considers the completely observed individuals selected at phase two; therefore, this mean score approach will generally be more efficient than the weighted pseudolikelihood approach that incorporates the known selection probabilities (Lawless et al. 1999; Robins et al. 1994).

3.4 Frameworks for Analysis and Design Criteria

Different designs can exploit phase-I data in different ways. The different second-phase sampling designs will result in different levels of efficiency of the resultant estimators, and the optimal designs will depend on the chosen method of analysis. We consider five sampling designs: simple random sampling, balanced sampling, optimal maximum likelihood sampling, optimal weighted pseudolikelihood sampling, and optimal mean score sampling. These designs (which are described in more depth below) require different amounts of information at phase-I. Simple random sampling ignores all phase-I data. Balanced sampling designs require only the size of the phase-I strata. The optimal designs are derived to minimize the asymptotic variance of the estimator of β_x and they require knowledge of the parameter values at the design stage. In practice, these parameter values would be unknown; however, it would be possible to base these optimal design derivations on initial parameter estimates found using a small pilot study. This process has been shown to work well in several settings (Reilly and Pepe 1995; Reilly 1996; Pepe et al. 1994; Whittemore and Halpern 1997).

We consider the problem where N, the size of the phase-I sample is fixed and budgetary constraints require that the expected number of individuals selected at phase-II, $P(R = 1) * N$, is also fixed. Optimal designs aim to minimize the variance of the estimator of β_x subject to this budgetary constraint. We consider Bernoulli sampling (Lawless et al. 1999) wherein all N individuals are observed at phase-I and selection decisions for inclusion in phase-II are made independently and according to pre-specified selection probabilities $\pi(Y,V;\delta)$.

Truly optimal designs are not always feasible as they may sometimes result in selection probabilities that exceed one (Reilly and Pepe 1995) and may degenerate and result in selection probabilities that are near zero for some strata (Breslow and Cain 1988). In general, small selection probabilities are problematic as they may preclude testing of certain interactions, and both the mean score method and the weighted pseudolikelihood require selection probabilities be bounded away from zero. We, therefore, constrain all of our selection probabilities to be in the range $(0.05,1)$. As in Reilly and Pepe (1995), when optimal selection probabilities fall outside of this range, we fix the offending selection probability at the boundary and optimize the remaining selection probabilities. The balanced design can suffer from a similar problem in that a truly balanced design can often require selection probabilities that are larger than 1 in smaller strata. In this situation, we fix the offending selection probabilities at 1 and select the remaining individuals in a balanced way from the other strata.

3.4.1 Simple Random Sampling

Simple random sampling uses phase-II selection probabilities that are the same for all individuals irrespective of their phase-I data: i.e. $\pi(Y,V;\delta) = P_R$ for some constant P_R. The data that arise from this design are MCAR. This naive sampling scheme does not exploit information available in the phase-I data and so it will be used as a baseline to assess the efficiency gains of more sophisticated designs.

3.4.2 Balanced Sampling

Breslow and Cain (1988) and Breslow and Chatterjee (1999) advocate a balanced sampling design. Phase-I data are used to stratify the sample, and the phase-II sample is chosen to contain the same number of individuals from each stratum. This design is not optimally efficient but is thought to offer a “reasonable compromise between the competing demands of efficiency and the need to check model assumptions” (Breslow and Chatterjee 1999).

It is not always clear how the phase-I data should be used to stratify the sample. For the clustered data problem, we will consider two balanced sampling designs. In the first balanced sampling design, the phase-I sample will be divided into the eight classes defined by all possible values of (Y_1, Y_2, V). However, since we are defining efficiency in terms of the variance of the estimator of β_x, and (8) assumes a common effect of X on either response, it may be more in the spirit of the balanced design to sample equally from the six strata defined by $(Y_1 + Y_2, V)$; therefore, we also consider this second balanced design when analysing the clustered data. Note that in our asymptotic calculations, these designs are based on expected phase-I stratum sizes, which come from having knowledge of the true parameters at the design stage.

3.4.3 Optimal Likelihood Sampling

If $\hat{\theta}$ is the estimator of $\theta = (\alpha', \beta')'$ which maximizes (2) and is estimated from data obtained with phase-II selection probabilities defined by δ, then asymptotically

$$\sqrt{N}(\hat{\theta} - \theta) \sim N(0, \mathscr{I}_{\theta;\delta}^{-1} \Gamma_{\theta;\delta} \mathscr{I}_{\theta;\delta}^{-1}),$$

where $\mathscr{I}_{\theta;\delta} = E\big[-\partial S_i(\theta)/\partial\theta'\big]$, $\Gamma_{\theta;\delta} = E\big[S_i(\theta)S_i'(\theta)\big]$ and $S_i(\theta)$ is the score function corresponding to the observed-data likelihood in (2). We say that this estimator has asymptotic variance $\mathscr{I}_{\theta;\delta}^{-1}$, since $\mathscr{I}_{\theta;\delta} = \Gamma_{\theta;\delta}$ (Cox and Hinkley 1974).

The expected information is affected by the choice of the phase-II selection parameter, δ, so optimal maximum likelihood designs, $\pi(Y, V; \delta^{opt})$, can be found for any specified θ. This is done here by numerically identifying the phase-II selection probabilities that minimize the asymptotic variance of the maximum likelihood estimator of β_x subject to the budgetary constraints. The budget limits how many individuals can be sampled in the second phase; we set

$$P(R = 1) = \sum_{Y,V} \pi(Y, V; \delta) P(Y, V) = P_R \tag{9}$$

so that given the size of the phase-I sample, N, the expected phase-II sample size is fixed at $N \cdot P_R$, for some prespecified sampling fraction P_R. This sampling design will be optimally efficient for maximum likelihood estimation of β_x whenever the covariate model and the parameters used in the design are correctly specified, as they are in the asymptotic calculations in the next section.

3.4.4 Optimal Mean Score Sampling

Reilly and Pepe (1995) show that the mean score estimator is asymptotically normal with an asymptotic variance that can be written as

$$\mathscr{I}_{\beta}^{-1} + \mathscr{I}_{\beta}^{-1} \Omega_{\beta;\delta}\, \mathscr{I}_{\beta}^{-1},$$

where

$$\mathscr{I}_\beta = E\big[-\partial U_\beta(Y|X,V)/\partial\beta'\big],$$

and

$$\Omega_{\beta;\delta} = \sum_{Y,V} P(Y,V)[\pi(Y,V;\delta)^{-1} - 1] \cdot var_{X|Y,V}[U_\beta(Y|X,V)],$$

with $var_{X|Y,V}[U_\beta(Y|X,V)]$ given by

$$\big\{E_{X|Y,V}\big[U_\beta(Y|X,V)U'_\beta(Y|X,V)\big] - E_{X|Y,V}\big[U_\beta(Y|X,V)\big]E_{X|Y,V}\big[U'_\beta(Y|X,V)\big]\big\}.$$

Therefore, the optimal second-phase selection probabilities, which give the greatest precision in estimating β_x subject to the budgetary constraint (9), can be written as

$$\pi(Y,V;\delta^{opt}) = \frac{P_R\big\{\mathscr{I}_\beta^{-1} var_{X|Y,V}[U_\beta(Y|X,V)]\mathscr{I}_\beta^{-1}\big\}_{[k,k]}^{1/2}}{\sum_{Y,V} P(Y,V)\big\{\mathscr{I}_\beta^{-1} var_{X|Y,V}[U_\beta(Y|X,V)]\mathscr{I}_\beta^{-1}\big\}_{[k,k]}^{1/2}},$$

where $\{A\}_{[k,k]}$ refers to the entry of the asymptotic variance matrix corresponding to β_x (Pepe et al. 1994; Reilly and Pepe 1995; Reilly 1996).

3.4.5 Optimal Weighted Pseudolikelihood Sampling

Asymptotically, the weighted pseudolikelihood estimator, $\tilde{\beta}$, is distributed as

$$\sqrt{N}(\tilde{\beta} - \beta) \sim N(0, \mathscr{I}_\beta^{-1}\Gamma_{\beta;\delta}\ \mathscr{I}_\beta^{-1}),$$

where

$$\Gamma_{\beta;\delta} = E\left[\frac{R_i}{\pi(Y,V;\delta)^2} U_\beta(Y|X,V)U'_\beta(Y|X,V)\right]$$

(Lawless et al. 1999; Robins et al. 1994). The asymptotic variance of the weighted pseudolikelihood estimator, $\mathscr{I}_\beta^{-1}\Gamma_{\beta;\delta}\ \mathscr{I}_\beta^{-1}$, can be written explicitly as a function of the selection probabilities by noting that $\mathscr{I}_\beta$ is functionally independent of δ, and

$$\Gamma_{\beta;\delta} = \sum_{Y,V} P(Y,V)\pi(Y,V;\delta)^{-1} E_{X|Y,V}\big[U_\beta(Y|X,V)U'_\beta(Y|X,V)\big].$$

Therefore, as in Reilly and Pepe (1995), a Lagrange multiplier approach can be taken to minimize the asymptotic variance matrix entry corresponding to the estimator of β_x subject the budgetary constraint in (9). These optimal second-phase selection probabilities are

$$\pi(Y,V;\delta^{opt}) = \frac{P_R\{\mathscr{I}_\beta^{-1}E_{X|Y,V}\left[U_\beta(Y|X,V)U'_\beta(Y|X,V)\right]\mathscr{I}_\beta^{-1}\}^{1/2}_{[k,k]}}{\sum_{Y,V}P(Y,V)\{\mathscr{I}_\beta^{-1}E_{X|Y,V}\left[U_\beta(Y|X,V)U'_\beta(Y|X,V)\right]\mathscr{I}_\beta^{-1}\}^{1/2}_{[k,k]}},$$

where again $\{A\}_{[k,k]}$ refers to the entry of the asymptotic variance matrix corresponding to β_x.

3.5 Asymptotic Relative Efficiencies

In order to assess the efficiency gain that can result from exploiting available auxiliary data in the selection of individuals for measurement of expensive covariate information, balanced and optimal phase-II sampling designs were derived for a range of parameter values. The asymptotic efficiencies of the estimators resulting from these designs were calculated relative to the asymptotic efficiency of a simple random sampling design. We considered the three methods of analysis: maximizing the observed data likelihood (ML), the mean score method (MS), and maximizing the weighted pseudolikelihood (WP). For each of these methods of analysis we considered four designs: simple random sampling (SRS), balanced sampling over all eight strata defined by (Y_1,Y_2,V) (BAL 8), balanced sampling over the six strata defined by (Y_1+Y_2,V) (BAL 6), and the sampling design which is asymptotically optimal for precise estimation of β_x with the given method of analysis (OPT).

The efficiency of each design D was calculated relative to simple random sampling through

$$RE_x(D,A) = \frac{asvar_D(\hat{\beta}_x^A)}{asvar_{SRS}(\hat{\beta}_x^A)}, \tag{10}$$

where, for example, $asvar_{BAL8}(\hat{\beta}_x^{ML})$ represents the asymptotic variance of the estimator of β_x that comes from using ML analysis with the BAL 8 design. We also consider the relative efficiency of the designs for estimating the effect of the auxiliary variable β_v and ψ. Note that the "optimal" design will not necessarily be efficient for estimation of parameters other than β_x, although Reilly (1996) reported that in their examples optimal designs for one parameter "achieved an improvement in the precision of almost all parameters".

The asymptotic relative efficiencies of the different sampling designs is presented in Figs. 1, 2, and 3, for estimation of β_x, β_v, and ψ, respectively. The relative efficiencies are presented for a range of values of the association parameter ψ while the other parameters were chosen so that $E[Y_1]=E[Y_2]=0.2; E[X]=0.25; E[R]=0.25; \beta_x=\log(1.25); \beta_v=\log(1.25); \alpha_v=\log(1.25)$. The fourth panel in each of these figures presents the asymptotic variance of the estimators that result from

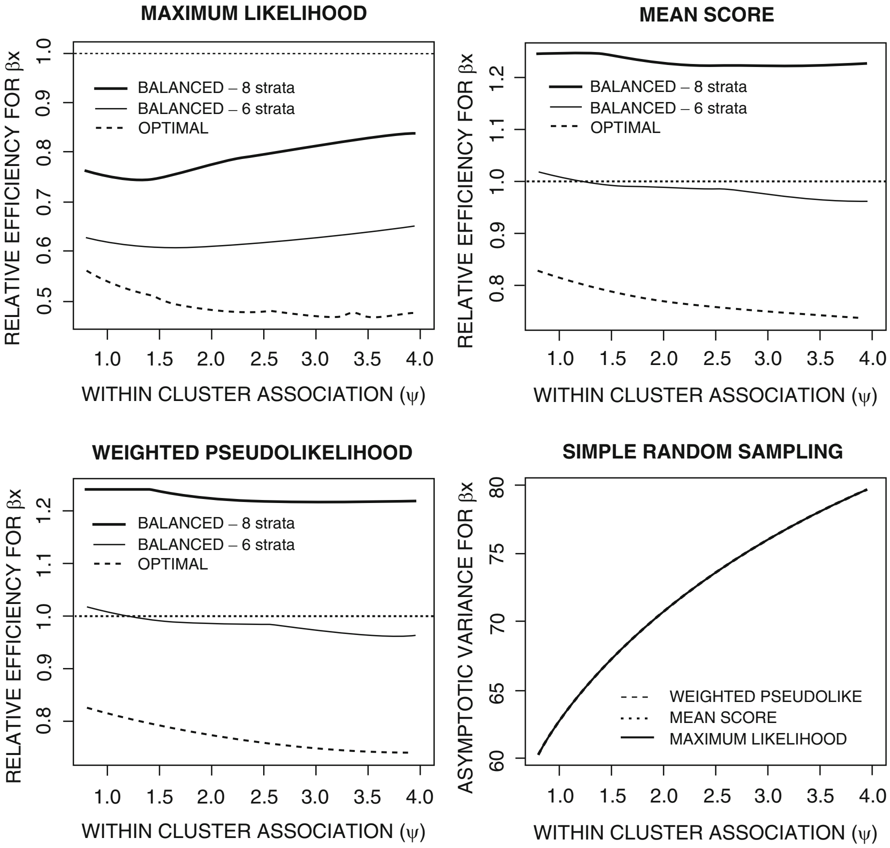

Fig. 1 Asymptotic efficiency of estimators of β_x under balanced and optimal designs relative to simple random sampling when using ML, MS, and WP. The asymptotic efficiencies are shown relative to the asymptotic variance of the SRS estimators which are shown in the fourth panel. $E[Y_1] = E[Y_2] = 0.2; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$

using SRS. Therefore, each of the first three panels within Figs. 1–3 leads to a comparison of efficiency amongst phase-II sampling designs for the specified method of analysis, while the fourth panel allows for a comparison of efficiency between methods of analysis.

It can be seen that the optimal design allows for a great increase in the efficiency of estimation of β_x; implementing an optimal phase-II sampling strategy can result in efficiency gains of 30–50 % over SRS, depending on the method of analysis (Fig. 1). In fact, for all methods of analysis, the optimal design results in more efficient estimators than SRS for all parameters, not just β_x (Figs. 1–3). This is similar to that which was reported by Reilly (1996), where optimizing for efficient estimation of one parameter led to efficiency gains everywhere.

The balanced designs sometimes result in efficiency gains and sometimes result in a loss of efficiency. In the estimation of ψ using WP analysis (Fig. 3, panel 3),

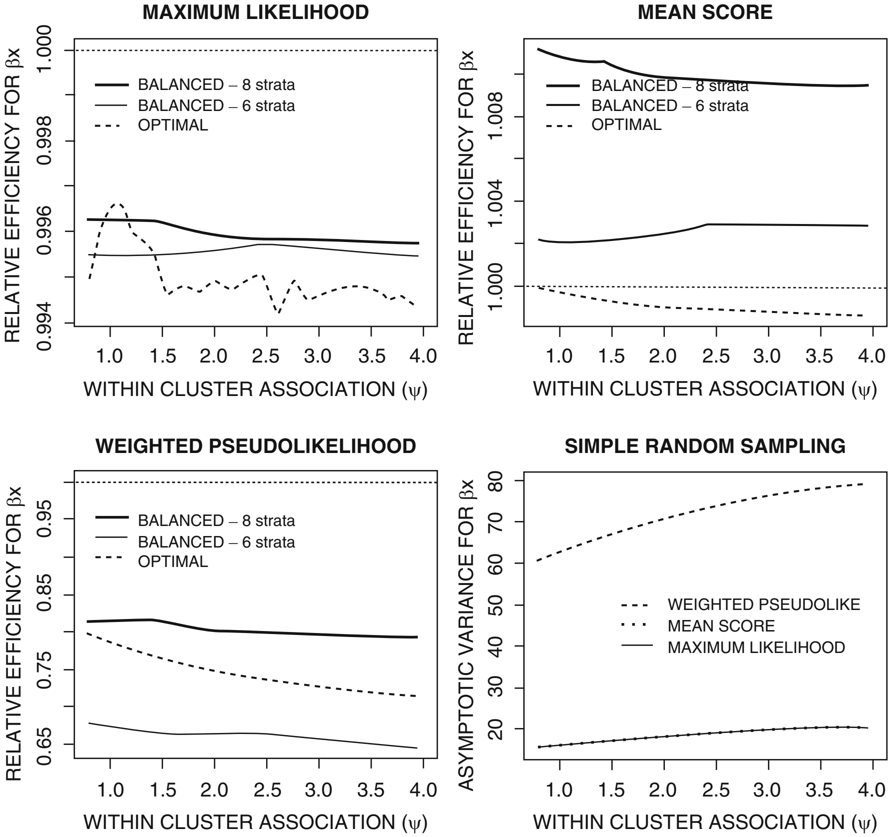

Fig. 2 Asymptotic efficiency of estimators of β_v under balanced and optimal designs relative to simple random sampling when using ML, MS, and WP. The asymptotic efficiencies are shown relative to the asymptotic variance of the SRS estimators which are shown in the fourth panel. $E[Y_1] = E[Y_2] = 0.2; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$

the balanced designs are both more efficient than the optimal design. However, for estimation of β_x (Fig. 1), a balanced design can be seen to be much less efficient than the naive SRS for both MS and WP analysis. The BAL 6 design is generally more efficient than the BAL 8 design, but neither design is consistently more efficient than SRS.

Estimators of β_x from SRS designs are very similar for ML, MS, and WP analysis (Fig. 1, panel 4), but use of WP is very inefficient for estimation of the other parameters (Figs. 2 and 3, panel 4).

It can also be seen that there is little difference amongst the designs for estimating ψ or β_v when using ML or MS (Figs. 2 and 3, panels 1 and 2). However, SRS is severely inefficient for estimating these parameters with WP analysis (Figs. 2 and 3, panel 3). So, the efficiency of estimators of ψ and β_v using WP analysis is greatly affected by the choice of sampling design, but even with the most efficient phase-II sampling design, WP estimators will still be less efficient than ML or MS estimators.

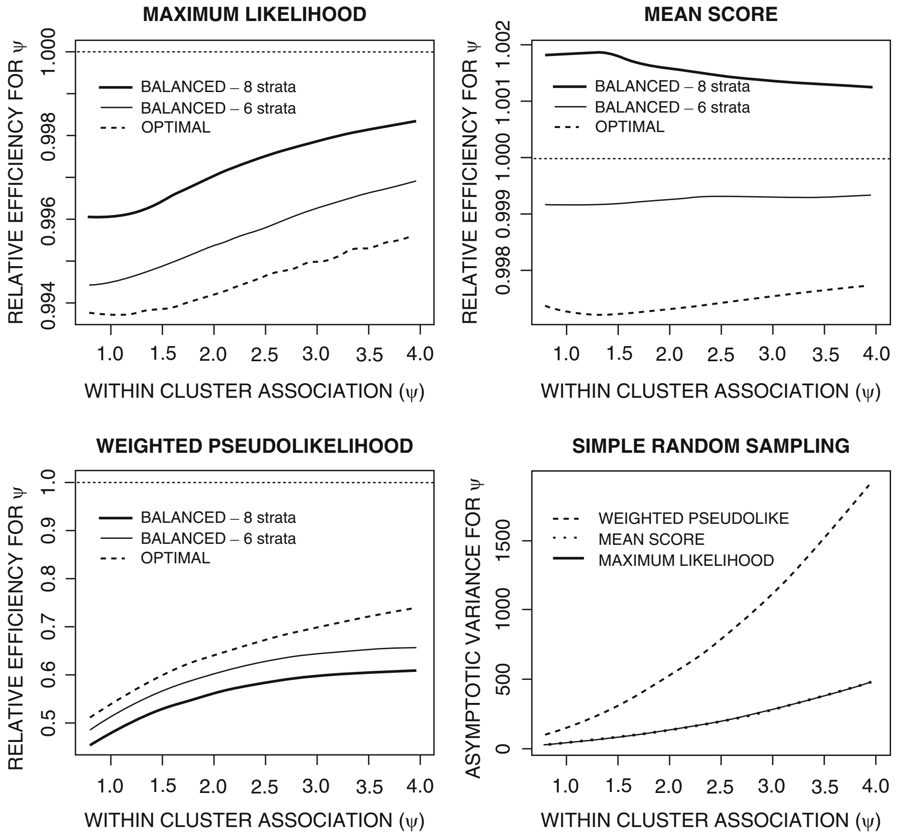

Fig. 3 Asymptotic efficiency of estimators of ψ under balanced and optimal designs relative to simple random sampling when using ML, MS, and WP. The asymptotic efficiencies are shown relative to the asymptotic variance of the SRS estimators which are shown in the fourth panel.
$E[Y_1] = E[Y_2] = 0.2; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$

4 Response-dependent Sampling with Longitudinal Binary Data

4.1 The Response Model for Longitudinal Data

Here we consider the analysis of binary data arising from a longitudinal study where the binary response variable is measured at baseline and at each of two prespecified follow-up timepoints. We assume that $Y_{i0} = 0$ and denote the response vector for individual i as $Y_i = (Y_{i0}, Y_{i1}, Y_{i2})'$. We again consider binary covariates X_i and V_i, where V_i is known for all individuals at time 0, but X_i will only be collected for individuals selected into a phase-II sample. This setting is a simplified version of the kind of data collected in the CLSA when interest lies in estimating the effect of a risk factor for the onset of a disease. Here Y_{ik} indicates the presence of disease at the assessment k, $k = 1,2$, and interest lies in the effect of covariates on $P(Y_{ik} = 1|Y_{i,k-1} = 0)$, the probability of disease developing between the $(k-1)$st and kth assessments, where we assume an irreversible disease process with $P(Y_{ik} = 0|Y_{i,k-1} = 1) = 0$. Specifically, here it is of interest to examine how the change in disease status (e.g. onset of diabetes) is affected by a time-invariant and expensive binary covariate X_i (e.g. a genetic factor), after accounting for an available baseline auxiliary covariate V_i.

We again consider analyses through maximization of the observed data likelihood (ML), the mean score method (MS), and weighted pseudolikelihoods (WP). For these data, we are not interested in estimating marginal parameters as in (8), rather we are primarily interested in the transitional effect of the covariate X in the response model

$$\operatorname{logit} P(Y_{ik} = 1|Y_{i,k-1} = 0, X_i, V_i; \beta) = \beta_0 + \beta_1 I(k=2) + \beta_x X_i + \beta_v V_i, \qquad k = 1,2.$$

Due to the irreversible nature of the disease process, the joint response model on which the likelihood methods are based is

$$\begin{aligned} P(Y_i|X_i,V_i;\beta) &= I(Y_{i,1}=1)P(Y_{i,1}=1|Y_{i,0}=0,X_i,V_i;\beta)I(Y_{i,1}=0)I(Y_{i,2}=0) \\ &\times [1-P(Y_{i,1}=1|Y_{i,0}=0,X_i,V_i;\beta)][1-P(Y_{i,2}=1|Y_{i,1}=0,X_i,V_i;\beta)] \\ &\times I(Y_{i,1}=0)I(Y_{i,2}=1)[1-P(Y_{i,1}=1|Y_{i,0}=0,X_i,V_i;\beta)] \\ &\times P(Y_{i,2}=1|Y_{i,1}=0,X_i,V_i;\beta). \end{aligned}$$

4.2 The Selection Model

Here we consider balanced and optimal designs for the selection of a phase-II sample at each of the three timepoints. This allows us to examine how the efficiency of designs is affected by the amount of auxiliary information available at phase-I for choosing the phase-II sample. Note that simple random sampling is not affected by the time at which the phase-II sample is chosen as this design does not exploit the data available at phase-I.

The selection model at time t can be expressed as $P(R_i = 1|Y_{i1}, \dots, Y_{it}, V_i; \delta^{(t)})$. At each progressive timepoint, more phase-I information is available for exploitation in deriving efficient phase-II selection probabilities. At timepoint 0, the phase-I sample can be divided into two strata based on the available information on V, so $\pi(Y,V;\delta^{(0)}) = \pi(V;\delta^{(0)})$; at timepoint 1, the phase-I sample can be stratified into four classes based on the available information on V and Y_1, so $\pi(Y,V;\delta^{(0)}) = \pi(Y_1,V;\delta^{(0)})$; at timpoint 2, the phase-I sample can be stratified into six classes based on the available information on V, Y_1, and Y_2, where it is known that $P(Y_2 = 0|Y_1 = 1) = 0$.

Simple random sampling is the same at each timepoint, but the efficiency of the balanced and optimal designs will be affected by the amount of information available at phase-I. Therefore, for this study of transitional effects, we consider 7 designs for each method of analysis: simple random sampling (SRS), balanced sampling using the phase-I data available at each timepoint (call these BAL 0, BAL 1, and BAL 2 at timepoints 0, 1, and 2, respectively) and the sampling designs which are optimal for estimating β_x given the specified method of analysis and the data that are available at the time of selection (call these OPT 0, OPT 1, and OPT 2). We will again present the efficiencies of the designs relative to simple random sampling, as calculated in (10).

The asymptotic variances and optimal designs can be found as in the previous section; however, summations are no longer over strata defined by (Y_1, Y_2, V), but rather over strata defined by the data that are available at the time of selection. This decrease in phase-I data essentially places added constraints on the optimal sampling designs derived in the previous section; for example, at timepoint 0, when only V is available for phase-II sampling decisions, then $\pi(Y,V;\delta) = \pi(V;\delta)$ for all $Y = (Y_1, Y_2) \in \{(0,0),(0,1),(1,1)\}$.

4.3 Asymptotic Relative Efficiencies

We derived optimal designs for a range of values of P_R, which defines the budgetary constraint in (9). Other parameters were chosen so that $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[R] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$. The relative efficiencies of the different sampling designs is presented in Figs. 4, 5, and 6, for ML analysis, MS analysis, and WP analysis, respectively. We consider the relative efficiency of each of the considered designs for estimating β_0, β_1, β_x, and β_v.

As expected, the optimal sampling design offered large efficiency gains over simple random and balanced designs when estimating β_x. As before, these optimal designs also added efficiency to the estimation of other parameters (Figs. 4, 5, and 6). Having more information at the time of sampling increased the efficiency of the optimal design for the estimation of all parameters. However BAL 2, the balanced design at timepoint 2, was generally less efficient than BAL 1, the balanced design which was based only on the auxiliary information available at timepoint 1. This indicates that, as was seen in the comparison of BAL 6 and BAL 8 in the previous section, having more phase-I information does not necessarily improve the efficiency of balanced designs.

The asymptotic variance of the ML and MS estimators under SRS was very similar; however, the optimal design offered a greater increase in efficiency for the ML estimator of β_x than for the MS estimator (Figs. 4 and 5, panel 3). The balanced designs were often less efficient than the naive simple random sampling approach to gathering data for estimation of β_x (Figs. 4, 5, and 6). The use of a balanced design appears to be particularly inefficient when analysis is to be carried out through the

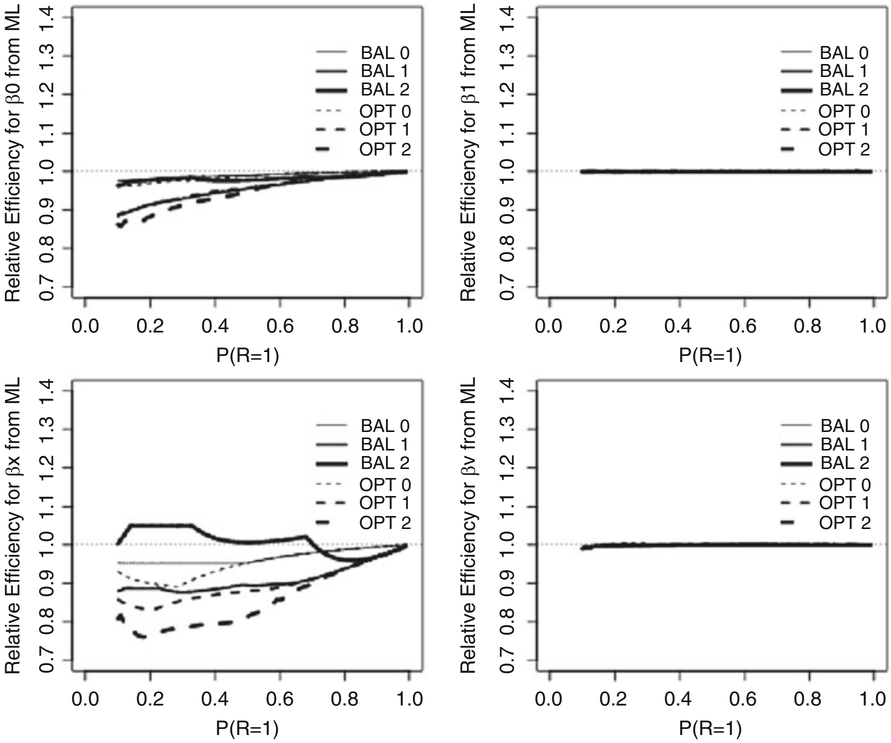

Fig. 4 Asymptotic efficiency of estimators under balanced and optimal designs relative to simple random sampling when using maximum likelihood analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[V] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$ Note: the asymptotic variance of the ML estimators of β_0, β_1, β_x, and β_v under SRS with $P(R = 1) = 0.5$ are, respectively, $9.86, 12.99, 33.00$, and 16.58

mean score method or the weighted pseudolikelihood (Figs. 5 and 6). Note that as the sampling fraction increases, smaller strata are selected in their entirety by the balanced designs (the selection probabilities must be capped at 1, as discussed previously); this accounts of the lack of smoothness in the change in asymptotic efficiency of the balanced designs. Some lack of smoothness can also be seen in the plot of the optimal ML designs; this occurs because these optimal ML designs are found numerically.

5 Discussion

To our knowledge this article was among the first to study the two-phase sampling designs involving clustered or longitudinal data. Given the increased interest in studies involving cross-sectionally clustered data and the recent trend towards the

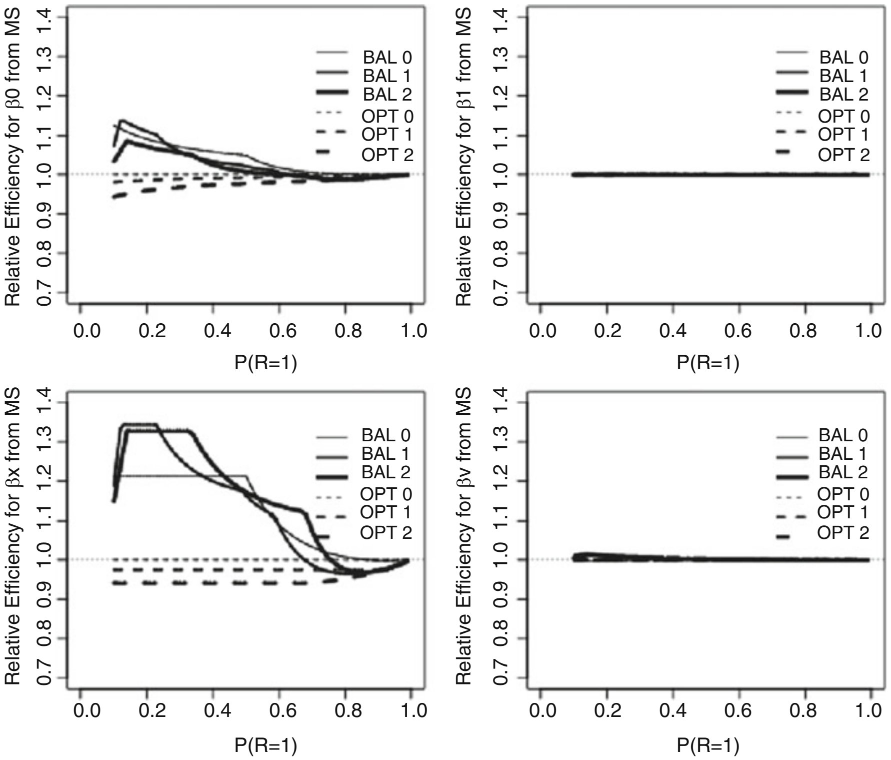

Fig. 5 Asymptotic efficiency of estimators under balanced and optimal designs relative to simple random sampling when using the mean score method for analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[V] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$ Note: the asymptotic variance of the MS estimators of β_0, β_1, β_x, and β_v under SRS with $P(R = 1) = 0.5$ are, respectively, $9.86, 12.99, 33.00$, and 16.58

design of massive cohort studies of health and disease, the insights that result from this work are important.

For the setting of clustered data, the first decision to make is typically on the method of analysis and there are a variety of frameworks one can adopt. We restricted attention to bivariate response data and marginal models for characterizing the effects of exposure. In this setting, maximum likelihood and the mean-score methods can be more efficient than weighted pseudolikelihood for estimation of the exposure effect (with maximum likelihood generally being the superior of the two) but this comes at the expense of making assumptions and modelling the covariate distribution. Interestingly, the three analysis methods have approximately the same efficiency when using simple random sampling. When covariates change within clusters an alternative model formulation could be based on random effects models. To our knowledge, there has been no work on two-phase designs within the framework of random effect models for clustered data and this is an area of current interest.

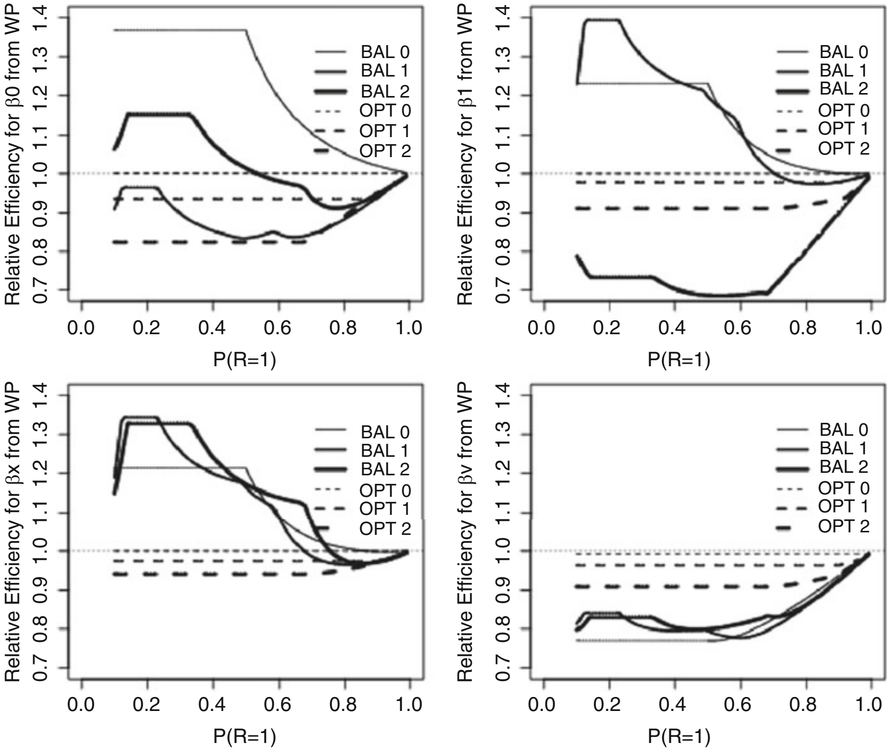

Fig. 6 Asymptotic efficiency of estimators under balanced and optimal designs relative to simple random sampling when using a weighted pseudolikelihood analysis to estimate transitional effects. $E[Y_1] = 0.2; E[Y_2] = 0.4; E[X] = 0.25; E[V] = 0.25; \beta_x = \log(1.25); \beta_v = \log(1.25); \alpha_v = \log(1.25)$ Note: the asymptotic variance of the WP estimators of β_0, β_1, β_x, and β_v under SRS with $P(R = 1) = 0.5$ are, respectively, $17.37, 25.97, 33.00$, and 33.00

We have adopted a very simple response model with a binary X and binary auxiliary variable. When the exposure variable is continuous, a robust implementation of the mean score method may be more appealing, and weighted pseudolikelihood would also have more appeal since no modelling of exposure is required. In ongoing work (not reported here) we found that optimal designs based on maximum likelihood analyses may be more sensitive to small changes in the parameters used at the design stage than optimal mean score designs. So, if models for exposure variable are difficult to formulate with confidence, the robustness of the mean score and weighted pseudolikelihood approaches may be more appealing. When the auxiliary variable is continuous, discretizing seems the most practical approach to addressing the curse of dimensionality and this has been recommended by several authors (Lawless et al. 1999).

When comparing the effect of different frameworks for analysis and design, it is interesting to note that the conclusions about optimality bear only on the criteria adopted for the optimal design. The intercept, effect of the auxiliary variable and association parameters do not necessarily behave in the same way.

The pragmatic approach of using balanced sampling designs as a compromise between robustness and efficiency does not yield clear and consistent recommendations; the resulting estimators sometimes perform well and sometimes perform poorly. It is therefore unclear what auxiliary information should be considered when implementing a balanced design in the more complex settings we consider here.

There are several directions of future research that are natural to consider. We focus on clusters of size two because of interest in the two sacroilliac joints among patients with psoriatic arthritis. However, clusters can naturally be much larger as would be the case if all joints were to be modelled. Dealing with larger cluster sizes is in principle straightforward but may suggest the use of second-order generalized estimating functions rather than likelihood analyses. One may elect to retain the robustness of a first order analysis by refraining from higher order assumptions, or invoke fourth moment assumptions to try to optimize efficiency at the expense of robustness in the estimating equation framework.

We have also restricted attention to a first order Markov model in the longitudinal context with only three assessments. Longer term follow-up, as is planned for the Canadian Longitudinal Study in Aging (Raina et al. 2009), raises questions about the need for more elaborate response models, the need for greater collapsing of strata, and issues surrounding time-varying covariates. These and other issues are subject to further research.

Acknowledgements Michael McIsaac's research was supported by an Alexander Graham Bell Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Discovery Grants to Richard Cook from NSERC (RGPIN 155849) and the Canadian Institutes for Health Research (FRN 13887). Richard Cook is a Canada Research Chair in Statistical Methods for Health Research. The authors thank Dr. Dafna Gladman and Dr. Vinod Chandran for collaboration and helpful discussions regarding the research at the Centre for Prognosis Studies in Rheumatic Disease at the University of Toronto. The authors gratefully acknowledge the careful review and comments from a referee and Dr. Brajendra Sutradhar.

References

Breslow, N.E., Cain, K.C.: Logistic regression for two-stage case-control data. Biometrika. **75**(1), 11–20 (1988)

Breslow, N.E., Chatterjee, N.: Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. Appl. Stat. **48**(4), 457–468 (1999)

Chandran, V., Tolusso, D.C., Cook, R.J., Gladman, D.D.: Risk factors for axial inflammatory arthritis in patients with psoriatic arthritis. J. Rheumatol. **37**(4), 809–815 (2010)

Chatterjee, N., Chen, Y., Breslow, N.E.: A pseudoscore estimator for regression problems with two-phase sampling. J. Am. Stat. Assoc. **98**(461), 158–168 (2003)

Cox, D.R., Hinkley, D.V.: Theoretical Statistics. Chapman & Hall, London (1974)

del Rincon, I., Williams, K., Stern, M.P., Freeman, G.L., O'Leary, D.H., Escalante, A.: Association between carotid atherosclerosis and markers of inflammation in rheumatoid arthritis patients and healthy subjects. Arthritis Rheum. **48**(7), 1833–1840 (2003)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39**(1), 1–38 (1977)

Heagerty, P.J., Zeger, S.L.: Marginalized multilevel models and likelihood inference. Stat. Sci. **15**, 1–26 (2000)

Heagerty, P.J.: Marginalized transition models and likeliood inference for longitudinal categorical data. Biometrics **58**(2), 342–351 (2002).

Horton, N.J., Laird, N.M.: Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. Biometrics **57**, 34–42 (2001).

Lawless, J.F., Kalbfleisch, J.D., Wild, C.J.: Semiparametric methods for response-selective and missing data problems in eegression. J. Roy. Stat. Soc. B **61**(2), 413–438 (1999)

Laird, N., Ware, J.H.: Random-effects models for longitudinal data. Biometrics **38**(4), 963–974 (1982)

Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**(1), 13–22 (1986)

Lipsitz, S.R., Laird, N.M., Harrington, D.P.: Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. Biometrika **78**(1), 153–160 (1991)

Little, R.J.A., Rubin, D.B.: Statistical analysis with missing data, 2nd edn. Wiley, New York (2002)

Neuhaus, J.M.: Statistical methods for longitudinal and clustered designs with binary responses. Stat. Meth. Med. Res. **1**, 249–273 (1992)

Pepe, M.S., Reilly, M., Fleming, T.R.: Auxiliary outcome data and the mean-score method. J. Stat. Plann. Infer. **42**, 137–160 (1994)

Pickles, A., Dunn, G., Vazquez-Barquero, J.L.: Screening for stratification in two-phase ("two-stage") epidemiological surveys. Stat. Meth. Med. Res. **4**, 73–89 (1995)

Prentice, R.L.: Correlated binary regression with covariates specific to each binary observation. Biometrics **44**(4), 1033–1048 (1988)

Rahman, P., Gladman, D.D., Cook, R.J., Zhou, Y., Young, G., Salonen, D.: Radiological assessment in psoriatic arthritis. Rheumatology **37**(7), 760–765 (1998)

Raina, P.S, Wolfson, C., Kirkland, S.A., Griffith, L.E., Oremus, M., Patterson, C., Tuokko, H., Penning, M., Balion, C.M., Hogan, D., Wister, A., Payette, H., Shannon, H., Brazil, K.: The Canadian longitudinal study on aging (CLSA). Can. J. Aging **28**(3), 221–229 (2009)

Reilly, M.: Optimal sampling strategies for two phase studies. Am. J. Epidemiol. **143**, 92–100 (1996)

Reilly, M., Pepe, M.S.: A mean score method for missing and auxiliary covariate data in regression models. Biometrika **82**(2), 299–314 (1995)

Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. **89**(427), 846–866 (1994)

Robins, J.M., Rotnitzky, A., Zhao, L.P.: Analysis of semiparametric regression models for repeated outcomes in the presence of Missing Data. J. Am. Stat. Assoc. **90**(429), 106–121 (1995)

Stiratelli, R., Laird, N., Ware, J.H.: Random-effects models for serial observations with binary response. Biometrics **40**(4), 961–971 (1984)

Sutradhar, B.C.: On auto-regression type dynamic mixed models for binary panel data. Metron **66**(2), 209–221 (2008)

Sutradhar, R., Cook, R.J.: A bivariate mover-stayer model for interval-censored recurrent event data: application to joint damage in rheumatology. Comm. Stat. Theor. Meth. **18**, 3389–3405 (2009)

Tolusso, D.C., Cook, R.J.: Robust estimation of state occupancy probabilities for interval-censored multistate data: an application involving spondylitis in psoriatic arthritis. Comm. Stat. Theor. Meth. **38**(18), 3307–3325 (2009)

Troxel, A.B., Lipsitz, S.R., Brennan, T.A.: Weighted estimating equations with nonignorable nonresponse data. Biometrics **53**(3), 857–869 (1997)

Tsiatis, A.A.: Semiparametric Theory and Missing Data. Springer, New York (2006)

Whittemore, A.S., Halpern, J.: Multi-stage sampling in genetic epidemiology. Stat. Med. **16**, 153–167 (1997)

Zeng, L., Cook, R.J.: Transition models for multivariate longitudinal binary data. J. Am. Stat. Assoc. **102**, 211–223 (2007)

Zhao, L.P., Prentice, R.L.: Correlated binary regression using a quadratic exponential model. Biometrika **77**(3), 642–648 (1990)

Zhao, Y.: Design and efficient estimation in regression analysis with missing data in two-phase studies. PhD thesis, University of Waterloo (2005)

Zhao, Y., Lawless, J.F., McLeish, D.L.: Likelihood methods for pegression models with expensive variables missing by design. Biom. J. **51**(1), 123–136 (2009)