Published in final edited form as:
The Canadian Journal of Statistics (2014), **42**(2): 268–284
DOI: 10.1002/cjs.11207

Response-dependent two-phase sampling designs for biomarker studies

MICHAEL A. MCISAAC

Department of Public Health Sciences,

Queen's University, Kingston, ON, K7L 3N6, Canada

E-mail: mcisaacm@queensu.ca

RICHARD J. COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

Summary

Two-phase sampling designs are developed and investigated for use in the context of a rheumatology study where interest lies in the association between a biomarker with an expensive assay and disease progression. We derive optimal phase-II stratum-specific sampling probabilities for analyses from parametric maximum likelihood (ML), mean score (MS), inverse probability weighted (IPW), and augmented IPW (AIPW) estimating equations. The easy-to-implement optimally efficient design for the MS estimator is found to be asymptotically optimal for the IPW and AIPW estimators we consider, and is shown to result in efficiency gains over balanced and simple random sampling even when analyses are likelihood-based. We further demonstrate the robustness of this optimal design and show that it results in very efficient estimation even when the model or parameters used in its derivation are misspecified.

Keywords: Asymptotic relative effciency; augmented inverse probability weighted estimating functions; inverse probability weighting; maximum likelihood estimation; response-dependent sampling; two-phase design.

This is the peer reviewed version of the following article: McIsaac, M. A. and Cook, R. J. (2014), Response-dependent two-phase sampling designs for biomarker studies. Can J Statistics, 42: 268-284. doi: 10.1002/cjs.11207, which has been published in final form at http://onlinelibrary.wiley.com/doi/10.1002/cjs.11207/references. This article may be used for non-commercial purposes in accordance With Wiley Terms and Conditions for Self-Archiving: http://olabout.wiley.com/WileyCDA/Section/id-820227.html#terms.

1 INTRODUCTION

Two-phase designs involve the collection of inexpensive auxiliary data in a phase-I sample. These data are exploited to help inform the selection of individuals for inclusion in a phase-II subsample in which expensive covariates are measured (Chatterjee, Chen, & Breslow, 2003). This sampling framework can be practically efficient whenever the cost of measuring a specific covariate of interest is high relative to the cost of measuring the response and associated auxiliary covariates (Reilly & Pepe,

1995), and it has been widely used to ensure good precision of estimates in studies with limiting budgetary constraints (Zhao, Lawless, & McLeish, 2009). The extent of the efficiency gain in two-phase designs depends on the values of interesting and nuisance parameters, the method of analysis and the way in which the phase-I data are exploited in the models for the phase-II selection probabilities (Tosteson & Ware, 1990; Reilly, 1996; Breslow & Chatterjee, 1999).

Reilly & Pepe (1995) derived optimal phase-II sampling designs for asymptotically efficient parameter estimation using their so-called mean score (MS) method. Whittemore & Halpern (1997) suggested that the simple form of this approach makes it attractive as a basis for design in a wide range of situations. Breslow and co-authors, however, criticize the generality of optimal designs which guarantee efficiency only for the chosen method of analysis and which may result in degenerate designs (Breslow & Cain, 1988; Breslow & Chatterjee, 1999). Further practical issues associated with the implementation of such optimal designs include the facts that (i) asymptotically optimal designs may not result in efficiency gains in finite samples, and (ii) derivation of optimal designs requires a priori knowledge of certain parameters. In this paper, we explore these issues in the context of a rheumatology study with the goal of characterizing the association between an expensive biomarker and disease progression.

Psoriatic Arthritis (PsA) is an autoimmune disease that causes considerable joint pain and inflammation, which can ultimately lead to serious disability and poor quality of life (Chandran et al., 2010). The disease course is complex and heterogeneous; some patients experience rapid joint destruction, and some exhibit little evidence of progression even after considerable follow-up (Gladman & Chandran, 2011). Identification of patients at high risk of progression is critical to ensure timely intervention for those who need it, and to avoid unnecessary use of expensive, powerful, but potentially toxic biologic therapies. A highly promising biomarker for this purpose is matrix metalloproteinase 3 (MMP-3), a plasma biomarker which has a role in the formation of bone, cartilage, and synovium (Okada et al., 1992). Interest lies in studying the association between MMP-3 and disease progression in PsA while controlling for erythrocyte sedimentation rate (ESR), a traditional marker of inflammation which is relatively inexpensive and easy to measure (Gladman and Chandran, 2011). Levels of MMP-3 can be measured by the assay of stored patient blood samples, but this is an expensive undertaking and cannot be carried out for all patients in the clinic. Using data available on ESR and damage progression, we consider a two-phase sampling design and derive optimal sampling probabilities for the selection of a subset of patients on whom MMP-3 measurement will be most informative.

The remainder of this paper is organized as follows. In Section 2, we introduce notation and formalize the problem of interest. In Section 3 we describe several methods for fitting regression models with incomplete covariate data and give large-sample properties of associated estimators. Section 4 contains guidelines for the derivation of optimal designs for various methods of analysis based on minimizing asymptotic variances, and simulation studies are presented to demonstrate the empirical efficiencies of the designs. In Section 5, we explore the sensitivity of optimal designs to misspecification of key parameters and misspecification of the nuisance covariate model. Further, we examine the utility of optimal designs when necessary parameters are not known a priori and must be estimated using external pilot studies. Concluding remarks and summary recommendations are made for the motivating PsA study in Section 6.

2 Design of Studies Using Two-Phase Sampling

Consider the setting where scientific interest lies in detecting and quantifying the effect of a new biomarker X on the mean of a response Y while adjusting for a known prognostic variable V. This response model of interest is denoted

$$\mu(X, V; \alpha) = E[Y|X, V; \alpha]. \tag{1}$$

We consider the case in which the response and prognostic variables are discrete and let

$$g(X|V;\beta) \tag{2}$$

denote the the conditional density of X given V, let $P(V;\gamma)$ denote the marginal probability mass function of V, let $\theta = (\alpha', \beta')'$ and let $\Psi = (\alpha', \beta', \gamma')'$. We suppose that Y and V are known for all N individuals in a phase-I sample $\{(Y_i, V_i), i = 1, \ldots, N\}$, but the covariate X can only be observed for a subset of individuals due to budgetary constraints. Let $R_i = 1$ if individual i is selected for inclusion in the phase-II sample (and hence for measurement of X_i), and let $R_i = 0$ otherwise. Thus, the data ultimately consist of N individuals: $n = \sum_{i=1}^{N} R_i$ of whom provide complete data (Y, X, V), and $(N - n)$ of whom provide information only on (Y, V).

We consider Bernoulli sampling wherein all sampling decisions are independent. The key feature of the two-phase design is that the researcher can define the sampling probabilities at the second phase in terms of the phase-I data through specification of the selection model

$$\pi(Y, V; \delta) = P(R = 1|Y, V; \delta). \tag{3}$$

Note that the covariate X is *missing at random* (Little & Rubin, 2002) when $P(R = 1|Y, X, V) = P(R = 1|Y, V)$, as assumed in the framework of this two-phase design. With discrete (Y, V), individuals in the phase-I sample can simply be divided into strata defined by (Y, V), where (3) will give stratum-specific selection probabilities. We consider optimal two-phase designs which select individuals in phase-II so that the asymptotic variance of the estimator of a particular component of α is minimized.

3 Frameworks for Analysis

If the data $\{(Y_i, X_i, V_i), i = 1, 2, \ldots, N\}$ were available for a random sample of size N from a population, the corresponding complete-data conditional likelihood would be

$$L_C = \prod_{i=1}^{N} P(Y_i, X_i|V_i) = \prod_{i=1}^{N} P(Y_i|X_i, V_i; \alpha) \cdot g(X_i|V_i; \beta).$$

Provided β is functionally independent of α, the solution to the score equation

$$\sum_{i=1}^{N} S(Y_i|X_i, V_i; \alpha) = \sum_{i=1}^{N} \partial \log P(Y_i|X_i, V_i; \alpha)/\partial\alpha = 0, \tag{4}$$

yields the maximum likelihood estimator $\widehat{\alpha}$.

3.1 Analysis via maximum likelihood

When X_i is known only if $R_i = 1$, the observed-data conditional likelihood is

$$\prod_{i=1}^{N} \left[P(Y_i, X_i|V_i; \alpha, \beta)\pi(Y_i, V_i; \delta)\right]^{R_i} \left[E_{X|V_i}\left[P(Y_i|X, V_i; \alpha)\right](1 - \pi(Y_i, V_i; \delta))\right]^{1-R_i}$$

(Robins, Rotnitzky, & Zhao, 1994). Since δ is functionally independent of θ, we need only consider the observed-data partial likelihood

$$L(\theta) = \prod_{i=1}^{N} \left[P(Y_i|X_i, V_i; \alpha)g(X_i|V_i; \beta)\right]^{R_i} \left[E_{X|V_i}\left[P(Y_i|X, V_i; \alpha)\right]\right]^{1-R_i}, \tag{5}$$

which requires specification of both the response model (1) and the nuisance covariate model (2) (Lawless, Kalbfleisch, & Wild, 1999). The ML estimate $\widehat{\theta}^{\rm ml}$ may be found by solving the score equations corresponding to (5) directly or via an EM algorithm (Dempster, Laird, & Rubin, 1977). The limiting distribution of $\widehat{\theta}^{\rm ml}$ depends on (1)-(3) so that asymptotically

$$\sqrt{N}(\widehat{\theta}^{\rm ml} - \theta) \sim N(0, \mathbb{A}(\Omega)^{-1}),$$

where $\Omega = (\alpha', \beta', \gamma', \delta')'$, $\mathbb{A}(\Omega) = -E\big[\partial \mathcal{S}_i(\theta)/\partial\theta'\big] = E\big[\mathcal{S}_i(\theta)\mathcal{S}_i'(\theta)\big]$, and $\mathcal{S}_i(\theta)$ is a contribution to the score function from a single individual obtained from the observed-data likelihood in (5). Note that $\mathbb{A}(\Omega)$ is a function of the full parameter set Ω since the expectation is taken with respect to (R, Y, X, V).

A model for the nuisance distribution of $X|V$ is required for (5), and misspecification becomes a real concern when X is continuous because it is then not possible to specify a saturated model for $X|V$ (Reilly & Pepe, 1995; Robins, Rotnitzky, & Zhao, 1995). One way to overcome this difficulty is through semiparametric restricted maximum likelihood (SPML), which involves maximization of

$$L(\alpha, G) = \prod_{i=1}^{N} \big[P(Y_i|X_i, V_i; \alpha)G(X_i|V_i)\big]^{R_i} \big[P(Y_i|V_i; G, \alpha)\big]^{1-R_i} ,$$

over the set of all discrete distributions G supported by the observed values of X. The complete-data likelihood can again be maximized by an EM algorithm (Zhao, Lawless, & McLeish, 2009) or via profile likelihood (Breslow & Holubkov, 1997; Scott & Wild, 1997).

3.2 Analysis via the mean score method

Under the EM algorithm, the contribution to (4) for an individual with unknown X is replaced with the conditional (given the observed data) expectation of their score contribution, as in the estimating equations

$$\sum_{i=1}^{N} \big\{ R_i S(Y_i|X_i, V_i; \alpha) + (1 - R_i) E_{X|Y,V}[S(Y_i|X_i, V_i; \alpha)] \big\} = 0 .$$

The MS method of Reilly & Pepe (1995) involves estimating this expectation empirically in a single step (Lawless, Kalbfleisch, & Wild, 1999). With discrete phase-I data, this expectation $E_{X|Y,V}[\cdot]$ can be estimated using the empirical conditional mean within strata defined by (Y, V), and the MS estimating equations can be simplified to

$$\sum_{i=1}^{N} U_i(\alpha; \widehat{\delta}) = \sum_{i=1}^{N} \frac{R_i}{\pi(Y_i, V_i; \widehat{\delta})} S(Y_i|X_i, V_i; \alpha) = 0 \tag{6}$$

where $\pi(Y, V; \widehat{\delta})$ is the empirical estimate of the stratum-specific selection probabilities based on phase-II data (Reilly & Pepe, 1995). The solution is denoted $\widehat{\alpha}^{\rm ms}$ and asymptotically,

$$\sqrt{N}(\widehat{\alpha}^{\rm ms} - \alpha) \sim N(0, \bar{\mathbb{A}}(\Psi)^{-1} + \bar{\mathbb{A}}(\Psi)^{-1}\mathbb{B}(\Omega)\, \bar{\mathbb{A}}(\Psi)^{-1}) ,$$

where $\bar{\mathbb{A}}(\Psi) = -E_{RYXV}\big[\partial U_i(\alpha; \delta)/\partial\alpha'\big] = -E_{YXV}\big[\partial S(Y_i|X_i, V_i; \alpha)/\partial\alpha'\big]$, $\mathbb{B}(\Omega) = \sum_{Y,V} P(Y, V; \Psi)$ $[\pi(Y, V; \delta)^{-1} - 1] \cdot \mathrm{var}_{X|Y,V}[S(Y|X, V; \alpha)]$, and $P(y, v; \Psi)$ is the joint probability that $(Y = y, V = v)$.

3.3 Analysis via inverse probability weighted estimating equations

Since the covariate of interest is missing at random with known phase-II selection probabilities $\pi(Y_i, V_i; \delta)$ that can be bounded away from zero, a consistent estimator can be obtained simply by suitably weighting the contributions to the complete-case estimating function as in the inverse probability weighted (IPW) estimating equation,

$$\sum_{i=1}^{N} \bar{U}_i(\alpha, \delta) = \sum_{i=1}^{N} \frac{R_i}{\pi(Y_i, V_i; \delta)} S(Y_i|X_i, V_i; \alpha) = 0 \tag{7}$$

(Robins, Rotnitzky, & Zhao, 1994). Note that it is not necessary to model the nuisance covariate model (2) to construct (7), so the analysis via IPW estimating equations is potentially more robust than analysis via parametric ML. The price for this robustness is that the IPW estimator is less efficient since none of the partial information available from the incomplete observations is exploited. Furthermore, when some phase-II selection probabilities are close to zero, the IPW estimator can perform poorly as estimates are greatly influenced by observations taken from these less frequently sampled strata (Tsiatis, 2006).

By comparing (6) and (7), it can be seen that the MS estimating equations are IPW estimating equations, but with sampling weights estimated; in general, estimation of δ results in an estimator with greater asymptotic efficiency (Robins, Rotnitzky, & Zhao, 1994; Lawless, Kalbfleisch, & Wild, 1999). By a series of conditioning arguments one can show that

$$E[\bar{U}_i(\alpha, \delta)\bar{U}_i'(\alpha, \delta)] = E_{YV}\left[\frac{E_{R|YXV}[R_i]}{\pi(Y_i, V_i; \delta)^2} E_{X|YV}[S(Y_i|X_i, V_i; \alpha)S'(Y_i|X_i, V_i; \alpha)]\right].$$

So, if $\widehat{\alpha}^{\text{ipw}}$ is the IPW estimator of α that uses known stratum-specific selection probabilities defined by δ, asymptotically

$$\sqrt{N}(\widehat{\alpha}^{\text{ipw}} - \alpha) \sim N(0, \bar{\mathbb{A}}(\Psi)^{-1}\, \bar{\mathbb{B}}(\Omega)\, \bar{\mathbb{A}}(\Psi)^{-1}),$$

where $\bar{\mathbb{B}}(\Omega) = \sum_{Y,V} \pi(Y, V; \delta)^{-1} P(Y, V; \Psi) E_{X|Y,V}\big[S(Y|X, V; \alpha)S'(Y|X, V; \alpha)\big]$.

3.4 Analysis via augmented inverse probability weighted estimating equations

Robins, Rotnitzky, & Zhao (1994) proposed augmented inverse probability weighted estimating equations (AIPW) of the form $\sum_{i=1}^{N} \bar{\bar{U}}_i(\alpha, \delta) = 0$, where

$$\bar{\bar{U}}_i(\alpha, \delta) = \frac{R_i}{\pi(Y_i, V_i; \delta)} S(Y_i|X_i, V_i; \alpha) - \frac{R_i - \pi(Y_i, V_i; \delta)}{\pi(Y_i, V_i; \delta)} \cdot \phi(Y_i, V_i), \tag{8}$$

using an arbitrary fixed function $\phi(Y_i, V_i)$ satisfying $E[\phi(Y_i, V_i)'\phi(Y_i, V_i)] < \infty$. The resulting AIPW estimator, $\widehat{\alpha}^{\text{aipw}}$, has the property that $\sqrt{N}(\widehat{\alpha}^{\text{aipw}} - \alpha)$ is asymptotically normal with mean 0 and variance

$$\bar{\mathbb{A}}(\Psi)^{-1}\, \bar{\bar{\mathbb{B}}}(\Omega)\, \bar{\mathbb{A}}(\Psi)^{-1}, \tag{9}$$

where $\bar{\bar{\mathbb{B}}}(\Omega) = E\big[\bar{\bar{U}}_i(\alpha, \delta)\bar{\bar{U}}_i'(\alpha, \delta)\big]$. In the absence of further auxiliary covariates, the optimal choice for the function $\phi(\cdot)$ in (8) is $\phi_S^{\text{opt}} = E[S(Y|X, V; \alpha)|Y, V]$ (Robins, Rotnitzky, & Zhao, 1994; Tsiatis, 2006). The double robustness property means that this AIPW estimator will remain consistent if either the selection model or the model for estimating ϕ_S^{opt} is correctly specified (Bang & Robins, 2005; Tsiatis, 2006). When selection probabilities are known, AIPW estimators may be particularly

appealing since, unlike the ML estimator, they will necessarily be consistent; the specification of $\phi(\cdot)$ will determine the efficiency of the estimators arising from (8).

In practice, Robins, Rotnitzky, & Zhao (1994) recommend estimating ϕ_S^{opt} with an empirical estimate of $E_{X|Y,V}[S(Y|X,V;\widehat{\alpha}^{\text{ipw}})]$. There has been some recent discussion on the utility of iteratively updating the estimate of ϕ_S^{opt} as α is estimated (Lumley, Shaw, & Dai, 2011; Scott & Wild, 2011). In our simulations (not shown), we found that such an iterative procedure results in perceivable small-sample efficiency gains for the AIPW estimator. Furthermore we note here that iteration is necessary to ensure the double robustness property of this AIPW estimating function since the expectation of the estimating function otherwise reduces to

$$E_{R,Y,V}\big[\frac{R_i-\pi(X_i,V_i;\delta)}{\pi(X_i,V_i;\delta)}(E_{Y|X,V}[S(Y|X,V;\alpha)]-E_{Y|X,V}[S(Y|X,V;\alpha^{\text{ipw}})])\big],$$

which is zero only if the selection model $\pi(X_i,V_i;\delta)$ is correctly specified.

Note that while ϕ_S^{opt} allows for the most efficient possible estimation of α amongst estimating equations of the form (8) (i.e., amongst estimating equations based on the optimal full-data estimating function S_i), we remark that it is possible in principle to achieve greater efficiency by deriving the optimal incomplete-data estimating function S^{eff} and its corresponding optimal augmentation term $\phi_{S^{\text{eff}}}^{\text{eff}}$ (Robins, Rotnitzky, & Zhao, 1994); see Yu & Nan (2006) for an accessible discussion of this point. This process, however, can require computationally-intensive iterative techniques (Robins, Rotnitzky, & Zhao, 1994; Tsiatis, 2006) and is rarely used in practice (Carpenter, Kenward, & Vansteelandt, 2006).

Due to these limitations and practical challenges, we henceforth restrict attention to the commonly used, efficient, but potentially sub-optimal, augmented estimating equations of the form (8) that utilize the optimal full-data estimating function S_i and ϕ_S^{opt}. This estimating function is called the efficient augmented estimator by Kulich & Lin (2004).

More details on the derivation of asymptotic theory discussed in this section can be found in McIsaac (2012).

4 Asymptotically Optimal Phase-II Sampling Designs

We consider six sampling designs which exploit phase-I data in different ways: simple random sampling, balanced sampling, optimal ML sampling, optimal MS sampling, optimal IPW sampling, and optimal AIPW sampling. In each optimal design, the selection models are derived to minimize the asymptotic variance of the estimator for the parameter associated with the biomarker of interest; that is, we wish to find the selection probabilities $\pi(Y,V;\delta)$ which will result in the most precise estimates of a particular component of α. These optimal designs require information on the phase-I stratum sizes and require specification of (unknown) parameter values. In contrast, simple random sampling does not exploit the phase-I information, and balanced sampling only requires knowledge of the sizes of the phase-I strata.

We reflect the budgetary constraints that limit the number of individuals that can be sampled in the second phase by specifying some $0 < P_R \leq 1$ so that

$$P(R=1;\delta) = N^{-1}\sum_{Y,V}\pi(Y,V;\delta)\cdot N_{YV} = P_R\ , \tag{10}$$

where it is assumed that N_{yv}, the number of individuals in the phase-I sample with $(Y=y,V=v)$, is known at the design stage.

Optimal designs for the MS method are discussed by Reilly & Pepe (1995) and Whittemore & Halpern (1997), amongst others. However, the optimal designs discussed here differ in that our

budgetary constraint (10) is based on the observed phase-II stratum sizes, whereas Reilly & Pepe (1995) and Whittemore & Halpern (1997) base their constraint on the expected stratum sizes, as in

$$P(R = 1; \delta) = \sum_{Y,V} \pi(Y, V; \delta) \cdot P(Y, V; \Psi) = P_R, \tag{11}$$

where Ψ is assumed to be known a priori or estimated from pilot data. We focus on the budgetary constraint in (10) for two reasons: (i) it has practical appeal when phase-I data are known, and (ii) it facilitates an exploration of the sensitivity of designs to changes in the parameters specified at the design stage. To illustrate this second point, note that our budgetary constraint in (10) does not depend on Ψ, but (11) is a function of the parameters used at the design stage. Therefore, misspecification of the unknown Ψ at the design stage would affect not only the optimal design for a given budgetary constraint, but also cause misspecification of the constraint itself! In fact, it can be seen that the optimal designs derived by Whittemore & Halpern (1997) (their Table IV) do not sample the expected number of individuals from their observed phase-I data (their Table I) due to differences between the expected and observed phase-I stratum sizes. The distinction between the use of (10) and (11) would become even more important if one were to consider basic stratified sampling (i.e. fixing the actual phase-II sample size rather than the expected sample size) or phase-I data that do not arise from a simple random sample (i.e. if N_{yv}/N was not a good estimate of $P(y, v; \Psi)$).

4.1 Simple random sampling

Under simple random sampling, the phase-II selection probabilities are the same for all individuals, so the R_i are i.i.d. Bernoulli random variables with mean P_R (i.e. $\pi(y, v) = P_R$ for all y, v). This design is easy to implement and renders covariate data *missing completely at random* (Little & Rubin, 2002). This naive sampling scheme will be used as a baseline to assess the efficiency gains of more refined designs.

4.2 Balanced sampling

Breslow & Cain (1988) and Breslow & Chatterjee (1999) advocate use of phase-II sampling probabilities which are inversely proportional to the size of the strata (i.e. $\pi(y, v) = c \cdot N_{yv}^{-1}$ where $c = P_R N / \sum_{Y,V} 1$), so that an equal number of completely observed individuals are expected in each stratum. While not necessarily efficient, this design is thought to offer a "reasonable compromise between the competing demands of efficiency and the need to check model assumptions" (Breslow & Chatterjee, 1999).

4.3 Optimal sampling for analysis via ML

The asymptotic variance of the ML estimator (given in Section 3.1) is a function of the phase-II selection probabilities, so the optimal design for estimation of the parameter of interest via ML can be obtained for any specified set of parameters Ψ by identifying the phase-II selection probabilities $\pi(Y, V; \delta)$ satisfying the budgetary constraints in (10) which minimize the entry in $\mathbb{A}(\Omega)^{-1}$ reflecting the asymptotic variance of the estimator of interest, say $\left\{\mathbb{A}(\Omega)^{-1}\right\}_{[k,k]}$. By introducing a new parameter λ and using Lagrange multipliers, the optimal selection probabilities can be found as stationary points of

$$\left\{\mathbb{A}(\Omega)^{-1}\right\}_{[k,k]} + \lambda[N^{-1} \sum_{Y,V} \pi(Y, V; \delta) N_{YV} - P_R] \ .$$

4.4 Optimal sampling for analysis via the MS method

The optimal stratum-specific selection probabilities $\pi(Y,V;\delta)$ satisfying (10) and resulting in the most efficient MS estimator for the parameter of interest are a stationary point of

$$\left\{\bar{\mathbb{A}}(\Psi)^{-1}+\bar{\mathbb{A}}(\Psi)^{-1}\mathbb{B}(\Omega)\,\bar{\mathbb{A}}(\Psi)^{-1}\right\}_{[k,k]}+\lambda[N^{-1}\sum_{Y,V}\pi(Y,V;\delta)N_{YV}-P_R]\,.$$

The fact that $\bar{\mathbb{A}}(\Psi)$ is functionally independent of δ means that a closed-form solution exists and, following some straightforward algebra, the optimal $\pi(y,v;\delta)$ for MS estimation can be written

$$\frac{P_R\big[P(y,v;\Psi)/(N_{yv}/N)\big]^{1/2}\left\{\,\bar{\mathbb{A}}(\Psi)^{-1}\mathrm{var}_{X|y,v}[S]\bar{\mathbb{A}}(\Psi)^{-1}\right\}^{1/2}_{[k,k]}}{\sum_{Y,V}\left[P(Y,V;\Psi)\cdot N_{YV}/N\right]^{1/2}\left\{\bar{\mathbb{A}}(\Psi)^{-1}\mathrm{var}_{X|Y,V}[S]\bar{\mathbb{A}}(\Psi)^{-1}\right\}^{1/2}_{[k,k]}}\,. \tag{12}$$

Here it is easy to see the appeal of (11) since this alternative budgetary constraint will result in N_{yv}/N being replaced with $P(y,v;\Psi)$, which will allow for simplification of the optimal design (see Reilly & Pepe, 1995). As noted earlier, however, only budgetary constraint (10) will properly fix the expected sample size and cost given observed N_{yv} and prespecified Ψ.

The form of the optimal design in (12) can easily be extended for any specified linear function, h, of elements of the asymptotic variance matrix by replacing $\{\mathbb{V}\}_{[k,k]}$ with $h(\mathbb{V})$; in particular, analogs of A-optimality and C-optimality (Emery & Nenarokomov, 1999) can be achieved by taking $h(\mathbb{V})=\mathrm{trace}(H\mathbb{V}H')$ with $H=I$ or $H=\mathrm{diag}\{\alpha\}^{-1}$, respectively. We do not pursue this here but there may be settings where interest lies in precise estimation of means, in which case this general approach could be useful.

4.5 Optimal sampling for analysis via IPW estimating equations

Similar arguments can be used to show that the asymptotically-optimal selection probabilities for efficient IPW estimation of our estimator of interest are stationary points of

$$\left[\bar{\mathbb{A}}(\Psi)^{-1}\,\bar{\mathbb{B}}(\Omega)\,\bar{\mathbb{A}}(\Psi)^{-1}\right]_{[k,k]}+\lambda[N^{-1}\sum_{Y,V}\pi(Y,V;\delta)N_{YV}-P_R],$$

and the optimal $\pi(y,v;\delta)$ for IPW estimation is

$$\frac{P_R\big[P(y,v;\Psi)/(N_{yv}/N)\big]^{1/2}\left\{\,\bar{\mathbb{A}}(\Psi)^{-1}E_{X|y,v}\big[SS'\big]\bar{\mathbb{A}}(\Psi)^{-1}\right\}^{1/2}_{[k,k]}}{\sum_{Y,V}\left[P(Y,V;\Psi)N_{YV}/N\right]^{1/2}\left\{\bar{\mathbb{A}}(\Psi)^{-1}E_{X|Y,V}\big[SS'\big]\bar{\mathbb{A}}(\Psi)^{-1}\right\}^{1/2}_{[k,k]}}. \tag{13}$$

In the consideration of optimal designs for binary data that follows, tedious but straightforward algebra shows that the asymptotically-optimal designs for MS and IPW estimation of our parameter of interest (α_2) are identical. This phenomenon does not hold more generally; in particular, solutions to the analogs of (12) and (13) differ if optimality is not defined simply in terms of the $[2,2]$ entry. Even in our simulations involving a continuous X, however, the asymptotic variances calculated for $\widehat{\alpha}_2^{\mathrm{ms}}$ and $\widehat{\alpha}_2^{\mathrm{ipw}}$ were always equal to at least seven decimal places, so somewhat surprisingly, we do not see here asymptotic efficiency gains for estimation of α_2 resulting from estimation of δ. Therefore, for our present purposes, the optimal designs for analysis via the MS method and for analysis via IPW estimating equations are identical.

4.6 Optimal sampling for analysis via AIPW estimating equations

As with ML, constrained numerical minimization procedures could be used to find the optimal choice of phase-II selection probabilities for analysis via AIPW estimating equations; for any specified Ψ, the optimal $\pi(Y, V; \delta)$ minimize the $[k, k]$ entry of (9) subject to the budgetary constraint in (10). However, the AIPW estimator considered here is asymptotically equivalent to the MS estimator in the presence of discrete phase-I data (Robins, Rotnitzky, & Zhao, 1994) and therefore the asymptotically-optimal stratum-specific sampling probabilities for AIPW estimation of the parameter of interest can also be found explicitly through (12); the optimal design for analysis via the MS method is, therefore, also optimal for analyses via the AIPW (and IPW) estimating equations considered in this paper.

5 Optimal Designs in the Psoriatic Arthritis Study

Here, we explore the utility of these asymptotically optimal phase-II sampling probabilities in a biomarker study for joint damage in psoriatic arthritis. We consider a phase-I sample of 504 patients who attended the University of Toronto Psoriatic Arthritis Clinic between 2003 and 2008. Information was recorded at a baseline visit on the patients' ESR levels and the extent of damage to the patients' joints. These patients also provided blood samples at the baseline visit which may be assayed to obtain the level of the biomarker MMP-3. At a follow-up visit roughly 2 years after the baseline assessment, it was determined whether there had been an increase in the number of damaged joints representing disease progression. Thus the data consist of a binary response Y indicating disease progression over the 2-year follow-up period, an inexpensive binary covariate V which indicates elevated ESR at baseline, and an expensive covariate X related to baseline MMP-3 levels. The response model of interest is

$$\mu(X, V; \alpha) = E[Y|X, V; \alpha] = \text{expit}(\alpha_1 + \alpha_2 X + \alpha_3 V),$$

where $\text{expit}(z) = \exp(z)/(1 + \exp(z))$, and we are particularly interested in efficient estimation of α_2.

The observed phase-I stratum sizes were $(N_{00}, N_{01}, N_{10}, N_{11}) = (196, 143, 61, 104)$, where N_{yv} denotes the number of patients with $(Y_i, V_i) = (y, v)$. MMP-3 levels are measured on a continuous scale, but are often dichotomized. Therefore, we consider optimal designs for both the situation where X is binary and where X is continuous.

Complete information was gathered for an additional pilot study of 53 PsA patients. In what follows, we take parameter estimates from the pilot study as true values and demonstrate the potential efficiency gains of optimal phase-two designs in this setting. We further examine the robustness of the optimal designs to misspecification of the parameter values used at the design stage, and explore how the potential efficiency gains of optimal designs differ when design parameters are estimated using pilot studies of different sizes.

5.1 Empirical properties of phase-II designs

We first investigate whether the empirical variation of the estimates of the parameter of interest are aligned with what is anticipated from the asymptotic results. Simulation studies were conducted for settings with binary and continuous X variables. In both cases, optimal designs were derived for each method of analysis with budgetary constraints reflected through the specification of $P(R{=}1){=}0.25$.

For the simulation study with binary X, 1000 complete datasets of size N were generated according to response model (1) and covariate distributions $P(X = 1|V; \beta) = \text{expit}(\beta_0 + \beta_v V)$, and

$$P(V = 1; \gamma) = \text{expit}(\gamma_0). \tag{14}$$

The parameter values $\Psi_0 = (\alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_v, \gamma_0)' = (-1.95, 1.00, 0.90, 1.05, -0.41, -0.04)'$ were obtained from analysis of the PsA pilot data where MMP-3 levels were dichotomized based on whether they exceed two standard deviations above the mean of controls (as specified by researchers at the PsA clinic).

For the simulation study with continuous X, 1000 complete datasets of size N were generated according to response model (1) and covariate distributions given by (14) and

$$g(X|V;\beta) = \frac{1}{\Gamma(\beta_0)(\beta_1 + \beta_v V)^{\beta_0}} X^{\beta_0 - 1} e^{-\frac{X}{\beta_1 + \beta_v V}}, \tag{15}$$

so that V remained binary, but $X|V$ followed a gamma distribution with shape β_0 and scale $\beta_1 + \beta_v V$. The parameters used in generating these data were $\Psi_{c0} = (\alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_v, \gamma_0)' = (-2.18,\ 0.03,\ .84,\ 1.40,\ 10,\ 5,\ -.04)$', where values of β were chosen to reflect the distribution of actual MMP-3 values given ESR status seen in the PsA pilot data.

Phase-I data were considered for each simulated dataset and phase-II samples were selected according to each of the four unique designs: simple random sampling (SRS); balanced sampling (Bal); the asymptotically-optimal design for estimation of α_2 via ML (Opt_{ml}) and the asymptotically-optimal design for estimation of α_2 via the MS, IPW, or AIPW estimating equations (Opt_{ms}). Each of the 1000 simulated incomplete datasets were analyzed via ML, MS, IPW, and AIPW estimating equations; when X was continuous, SPML was used in place of ML. The empirical standard errors (ESE) for estimators of α are presented in Table 1.

To avoid undesirable degenerate designs with near-zero selection probabilities in some strata (as discussed in Breslow & Cain, 1988), we constrain the stratum-specific selection probabilities to be at least 0.05; in our simulations, this restriction only affected the Opt_{ml} design. This constraint additionally ensured that samples could be analyzed with the IPW estimating equations, where near-zero selection probabilities are especially problematic. Of course, selection probabilities were also constrained to be at most 1; violations to these constraints can be avoided by proceeding along the boundaries, as in Reilly & Pepe (1995). The stratum-specific selection probabilities π = $[\pi(0,0),\ \pi(0,1),\ \pi(1,0),\ \pi(1,1)]$ employed by the SRS design were $\pi^{\text{srs}} = [0.25, 0.25, 0.25, 0.25]$. The balanced and optimal designs were based (at least in part) on phase-I data, so these designs depended on each simulated dataset. The average selection probabilities over the 1000 simulated datasets with binary X for the Bal, Opt_{ml}, and Opt_{ms} designs were $\bar{\pi}^{\text{bal}} = [0.16, 0.22, 0.52, 0.32]$, $\bar{\pi}^{\text{ml}} = [0.05, 0.33, 0.16, 0.58]$, and $\bar{\pi}^{\text{ms}} = [0.10, 0.25, 0.45, 0.43]$, respectively. The average selection probabilities over the 1000 simulated datasets with continuous X for the Bal, Opt_{ml}, and Opt_{ms} designs were likewise quite different at $\bar{\pi}^{\text{bal}} = [0.15, 0.19, 0.81, 0.38]$, $\bar{\pi}^{\text{ml}} = [0.07, 0.31, 0.46, 0.52]$, and $\bar{\pi}^{\text{ms}} = [0.12, 0.29, 0.46, 0.41]$, respectively.

The asymptotic variances of MS, IPW, and AIPW estimators of α_2 are equal, but for α_1 and α_3 the asymptotic variance of the IPW estimator is much greater than that for MS and AIPW estimators. With $N = 500$ (see Table 1), the MS estimating equations often resulted in much smaller empirical standard errors than the AIPW estimating equations; the MS estimator appears to have better small sample properties than the asymptotically equivalent AIPW estimator. We also explored the small-sample properties of an AIPW estimator where the approximation of ϕ_S^{opt} was updated iteratively with the estimator of α. This iterative AIPW estimator did have better small sample properties than the non-iterative AIPW estimator; the iterated AIPW estimator is reported in Table 1 for the simulations with continuous X because several AIPW estimates found without iteration were so poor that reporting an empirical standard error would be almost meaningless. A similar improvement in the small sample properties of AIPW estimators was achieved by using $\widehat{\alpha}^{\text{ms}}$ instead of $\widehat{\alpha}^{\text{ipw}}$ to estimate ϕ_S^{opt}; in this case, the MS and AIPW estimates were generally equivalent to at least 7 decimal places. Interestingly, the finite-sample efficiency of the IPW estimators of α_2 was actually generally greater here than the efficiency of either the MS or AIPW estimators, despite the fact that the latter methods of analysis are generally more asymptotically efficient.

Table 1: Empirical standard errors resulting from analyzing 1000 simulated datasets consisting of $N = 500$ individuals while employing four different phase-II sampling designs with an expected phase-II sample size of $P(R = 1) = 0.25$.

	Method of Analysis											
	ML / SPML			MS			IPW			AIPW[b]		
Design	α_1	α_2	α_3	α_1	α_2	α_3	α_1	α_2	α_3	α_1	α_2	α_3
						Binary Covariate X^a						
SRS	0.469	0.519	0.219	0.478	0.527	0.220	0.556	0.520	0.413	0.482	0.533	0.222
Bal	0.418	0.473	0.217	0.435	0.490	0.221	0.486	0.485	0.390	0.442	0.496	0.232
Opt_{ml}	0.391	0.426	0.227	0.486	0.577	0.241	0.676	0.556	0.565	0.633	0.608	0.414
Opt_{ms}	0.372	0.419	0.220	0.376	0.424	0.223	0.459	0.423	0.390	0.388	0.428	0.240
						Continuous Covariate X^a						
SRS	0.319	0.016	0.267	0.326	0.016	0.268	0.462	0.016	0.494	0.326	0.016	0.268
Bal	0.264	0.013	0.250	0.283	0.014	0.257	0.344	0.014	0.387	0.283	0.014	0.257
Opt_{ml}	0.261	0.013	0.256	0.270	0.015	0.285	0.425	0.015	0.461	0.319	0.015	0.298
Opt_{ms}	0.261	0.012	0.251	0.264	0.013	0.257	0.377	0.013	0.401	0.264	0.013	0.257

[a] For the results reported in the top half of the table, X was binary; for the results reported in the bottom half of the table, $X|V$ followed a gamma distribution. With a continuous X, we consider SPML instead of fully parametric ML.
[b] For continuous X, 10 iterations were used in the estimation of ϕ_S^{opt} in order to address some particularly poor non-iterative AIPW estimates.

As expected, the asymptotically-optimal designs resulted in the smallest empirical standard errors in the estimates of α_2 for the respective method of analysis in simulations where N was very large (not shown). However, the Opt_{ms} design actually demonstrated the greatest efficiency for estimation of α_2 with small sample sizes ($N = 500$) for both a binary and a continuous X. This apparent super-efficiency of the Opt_{ms} design suggests that it may be more appealing than Opt_{ml} in small samples, regardless of whether a likelihood or weighted estimating function based analysis is to be carried out.

In all cases, the Opt_{ms} design resulted in more efficient estimators of α_2 than either SRS or the balanced design; this design also generally resulted in more efficient estimators for α_1, but the balanced design was often more efficient for estimation of α_3. If efficient estimation of α_3 were also of primary importance, as note in Section 4.4, the definition of optimality could be modified and the optimal designs could be revised accordingly. While the Opt_{ml} design resulted in inefficient estimators from analyses based on weighted estimating equations, the Opt_{ms} design was more efficient than either the SRS or balanced designs for estimation of the parameter of interest even for ML estimators.

5.2 Robustness of optimal designs to misspecification

In the previous simulations, optimal designs were derived using the true parameter values, which are of course unknown in practice. In this section we explore the sensitivity of optimal phase-II sampling designs to misspecification of models and parameter values at the design stage. This section, therefore, represents an analysis of the sensitivity of designs to estimates of parameter values which are elicited from existing literature or expert knowledge.

Table 2 contains the asymptotic relative efficiencies (relative to simple random sampling) for estimation of α_2 with binary X when the supposedly-optimal designs have been derived based on misspecified parameter values. The asymptotic relative efficiency (ARE) of the estimator $\widehat{\alpha}_2$ under the Opt_{ml} design, for example, would be calculated as $\text{asvar}(\widehat{\alpha}_2; \delta^{\text{ml}})/\text{asvar}(\widehat{\alpha}_2; \delta^{\text{srs}})$ where $\text{asvar}(\widehat{\alpha}_2; \delta^{\text{ml}})$ is the asymptotic variance of $\widehat{\alpha}_2$ under the Opt_{ml} design – smaller AREs therefore correspond to more efficient designs. Here, the asymptotic variance of the ML and MS estimators under SRS are equivalent to 10 decimal places. The first column contains the results when the design is based on the 'true' parameter values (obtained from analysis of the PsA pilot data). The other columns display the relative efficiency for estimating this true α_2 when designs are derived to be optimal for incorrectly-specified parameter values. To provide a framework for the study of misspecification, we used incorrect parameter sets obtained as estimates from employing unwarranted independence assumptions when analysing the PsA pilot data.

Even when the parameter set was grossly misspecified (when it was mistakenly assumed that $X \perp V$ and $Y \perp (X, V)$), the 'optimal' designs were still asymptotically more efficient than SRS for their respective methods of analysis (see Table 2). In fact, the Opt_{ms} designs were always more efficient than SRS for both considered methods of analysis regardless of the misspecification. These Opt_{ms} designs were also more efficient than the balanced design for analysis via the MS method, while the Opt_{ml} design was again often quite inefficient for analysis via MS. The Opt_{ms} design was also often more efficient than the balanced design for analysis via ML, and when the balanced design was asymptotically more efficient, the efficiency gains were small. In addition, the Opt_{ms} was asymptotically more efficient than the Opt_{ml} for analysis via ML when the misspecification was large.

In Table 3 we display the empirical standard errors from a simulation study using continuous X in which the optimal designs are derived using an incorrectly modelled covariate distribution. Note that the incorrect specification of the model for $X|V$ will not affect the consistency of our estimators since this model is only used for specifying the $\pi(Y, V; \delta)$.

As before, the optimal designs were derived under the assumption that $X|V$ followed a gamma distribution. However, datasets of $N = 500$ individuals were simulated based on a log-normal model for $X|V$. The parameters used in generating the simulated data ($\Psi_{c1} = (\alpha_1, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_v, \gamma_0)' = (-2.18, 0.03, .84, 0.73, 2.77, 2.37, -.04)'$) were chosen so that the mean and variances of $X|V$ would

Table 2: Asymptotic efficiencies for estimation of α_2 relative to SRS when optimal designs are based on incorrect parameter values; different sets of parameter values correspond to estimates resulting from employing unwarranted independence assumptions when analysing the pilot data.

	Assumptions in Analysing Pilot Data				
	$-$	$X \perp V$	$Y \perp (X, V)$	$X \perp V$, $Y \perp X \mid V$	$X \perp V$, $Y \perp (X, V)$
			ML Analysis		
Bal	0.805	0.805	0.805	0.805	0.805
Opt_{ml}	0.728	0.730	0.781	0.779	0.938
Opt_{ms}	0.768	0.773	0.807	0.791	0.812
			MS Analysis		
Bal	0.829	0.829	0.829	0.829	0.829
Opt_{ml}	1.100	0.996	0.789	0.781	2.550
Opt_{ms}	0.771	0.773	0.814	0.793	0.826

be approximately equal to the mean and variances of the gamma distributed $X|V$ that was previously considered. For the simulated data, phase-II samples were selected using SRS, balanced, and the misspecified Opt_{ms} and Opt_{ml} designs; these four potential two-phase samples were each analyzed using SPML, MS, IPW, and AIPW estimating equations. This simulation represents what would arise if the optimal designs were derived for the PsA study under the mistaken assumption that MMP-3 followed a gamma distribution given ESR status, when in reality it followed a log-normal distribution. The optimal designs resulted in very efficient estimators of α_2 despite the fact that they were derived based on an incorrect specification of the covariate distribution. In fact, the misspecified Opt_{ml} design was the most efficient design for the SPML estimators and performed almost as well as the balanced design for the other methods of analysis. Importantly, the misspecified Opt_{ms} design was still more efficient than the SRS and balanced designs for all methods of analysis. As before, the optimal designs did not necessarily increase the efficiency for estimation of parameters other than α_2.

5.3 Robustness of optimal designs to use of pilot data

As mentioned above, the derivation of true optimal sampling designs requires a priori knowledge of parameter values which will generally be unknown. We have shown in the previous section that the Opt_{ms} design seems to be fairly robust to misspecification at the design stage. In practice, when a priori knowledge of the necessary parameters is not available, it is possible to derive optimal designs by estimating parameter values from a small validation sample (Reilly & Pepe, 1995). However, if optimal designs are sensitive to the small changes in design parameters that will inevitably result from the use of pilot data, then these optimal designs will be of little practical use. Here we explore the efficiency of optimal designs when they are based on parameter estimates from pilot studies of different sizes.

We considered external pilot studies of size m, where $m \in \{50, 200, 500, 1000\}$. For each pilot study, we simulated data (Y, X, V) for m individuals using parameter values estimated from the PsA pilot data taken as the true values and we added two observations (corresponding to a large and a small X value) to each of the 4 strata defined by the phase-I data (Y, V) in order to achieve greater stability in our estimates (Pepe, Reilly, & Fleming, 1994).

Table 3: Empirical standard errors resulting from analyzing 1000 simulated datasets with a phase-I sample size of $N = 500$ when $X|V$ followed a log-normal distribution.[a]

	Method of Analysis											
	SPML			MS			IPW			AIPW[b]		
Design	α_1	α_2	α_3	α_1	α_2	α_3	α_1	α_2	α_3	α_1	α_2	α_3
SRS	0.329	0.016	0.261	0.332	0.017	0.262	0.450	0.016	0.463	0.332	0.017	0.262
Bal	0.279	0.015	0.247	0.294	0.016	0.252	0.348	0.016	0.384	0.294	0.016	0.252
Opt_{ml}	0.275	0.013	0.255	0.288	0.016	0.278	0.503	0.016	0.528	0.332	0.016	0.292
Opt_{ms}	0.278	0.014	0.255	0.283	0.015	0.258	0.377	0.015	0.427	0.292	0.015	0.260

[a] The simulated $X|V$ followed a log normal distribution, but optimal designs were derived assuming $X|V$ followed a gamma distribution.
[b] Ten iterations were used in the estimation of ϕ_S^{opt} in order to address some particularly poor non-iterative AIPW estimates.

These simulated data were then used to find $\widehat{\Psi}$, a ML estimate. Optimal designs were derived using $\widehat{\Psi}$ and the asymptotic variances of estimators of the true α_2 that would result from employing these designs were recorded. This process was repeated 1000 times for each size of the pilot study. The results of these simulations are presented in Figure 1 for the scenario with a binary X and in Figure 2 for a continuous X.

The Opt_{ml} design was very inefficient for analysis via the MS method regardless of the size of the pilot study used to estimate the parameters used in deriving the optimal design; the Opt_{ml} design was also often inefficient for analysis via ML when the size of the pilot study was small. The Opt_{ms} design, however, was generally more efficient than both SRS and the balanced design for analyses via both ML and the MS method regardless of the size of the pilot studies. Furthermore, the efficiency resulting from the Opt_{ml} design was more variable than that from the Opt_{ms} design. Thus, the Opt_{ms} design appears to be more robust to the small changes that occur when pilot data are used to estimate design parameters, however, the Opt_{ml} design was still generally the most efficient design for ML estimators.

6 Conclusions and Recommendations for the PsA Study

We have derived closed-form asymptotically-optimal two-phase sampling designs to guide the selection of individuals for measurement of expensive covariates when analyses are carried out using MS, IPW or the most common AIPW estimators. Such optimal designs can be found numerically when analyses are to be carried out via ML and other AIPW. We have presented these derivations in the context of budgetary constraints that differ from what has been traditionally used in the MS literature and shown that while this results in a slightly more complex design, it allows for the expected study size to be fixed for any given set of phase-I data. We have further shown that the easy-to-implement optimal design for MS analysis is quite robust to misspecification of the design parameters and results in greater efficiency than other general sampling designs when using any of the considered methods of analysis. While this design is not asymptotically optimal for ML analysis, the simplicity and demonstrated robustness of this design make it appealing for use in practice where necessary design parameters are unknown a priori.

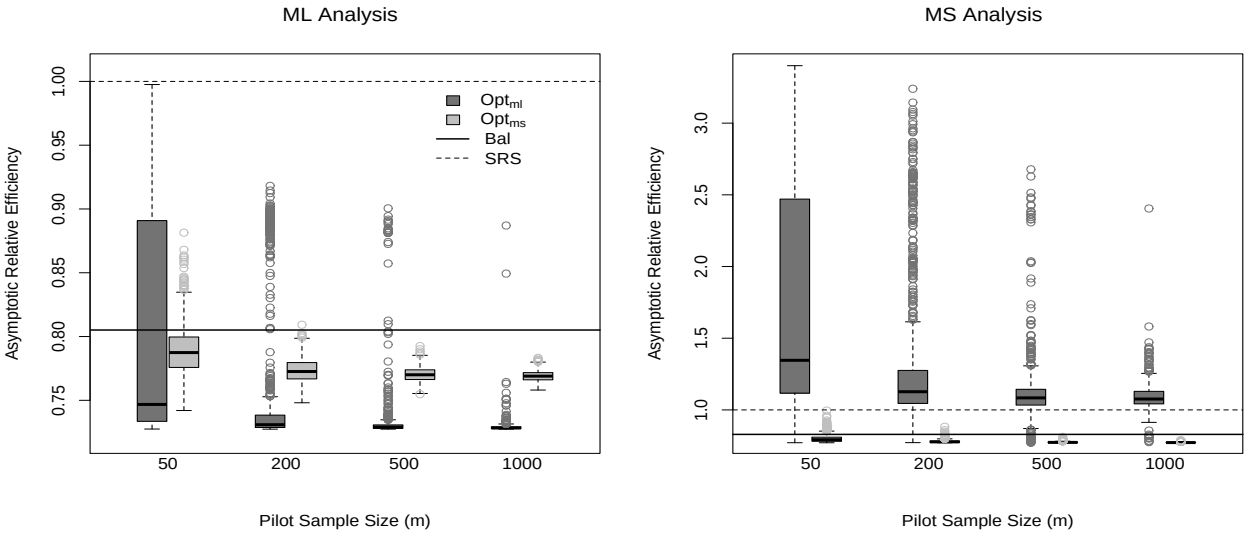

Figure 1: Asymptotic efficiencies of designs relative to SRS as a function of the size of the pilot studies used to estimate the design parameters with binary covariates and response. The balanced and SRS designs do not utilize the pilot data and are not affected by its size.

Our recommendations for the PsA clinic are that patients be selected from the phase-I sample according to the Opt_{ms} design based on parameter values estimated using the 53 patients in the pilot study. If budgetary constraints dictate that only 25% of the 504 available serum samples can be analysed for measurement of MMP-3, we recommend selecting according to the model $\pi^{\text{ms}} = [0.10, 0.25, 0.44, 0.42]$ if MMP-3 is to be dichotomized in the analysis, and according to the selection model $\pi^{\text{ms}} = [0.12, 0.30, 0.35, 0.35]$ if the effect of MMP-3 is to be modelled as a continuous covariate. The efficiency gains from this selection model could be quite substantial regardless of whether analysis is to be carried out using likelihood or weighted estimating equation approaches. Furthermore, these designs should work quite well even if the parameter estimates taken from the pilot data are poor.

We focussed here on the implementation of optimal designs in the presence of discrete phase-I data. If phase-I data are continuous, then discretization of these data will allow for the simple form of this design to be retained, though the efficiency and robustness of such a design need further exploration. Similar suggestions for discretization have been made for the purpose of analysis (Lawless, Kalbfleisch, & Wild, 1999); by discretizing only at the design stage, we avoid the potential resulting loss of efficiency in analysis noted by Chatterjee, Chen, & Breslow (2003).

ACKNOWLEDGEMENTS

This research was supported by an Alexander Graham Bell Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Michael McIsaac and a Discovery Grant from NSERC (RGPIN 155849) and a grant from the Canadian Institutes for Health Research (FRN 13887) to Richard Cook. 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.

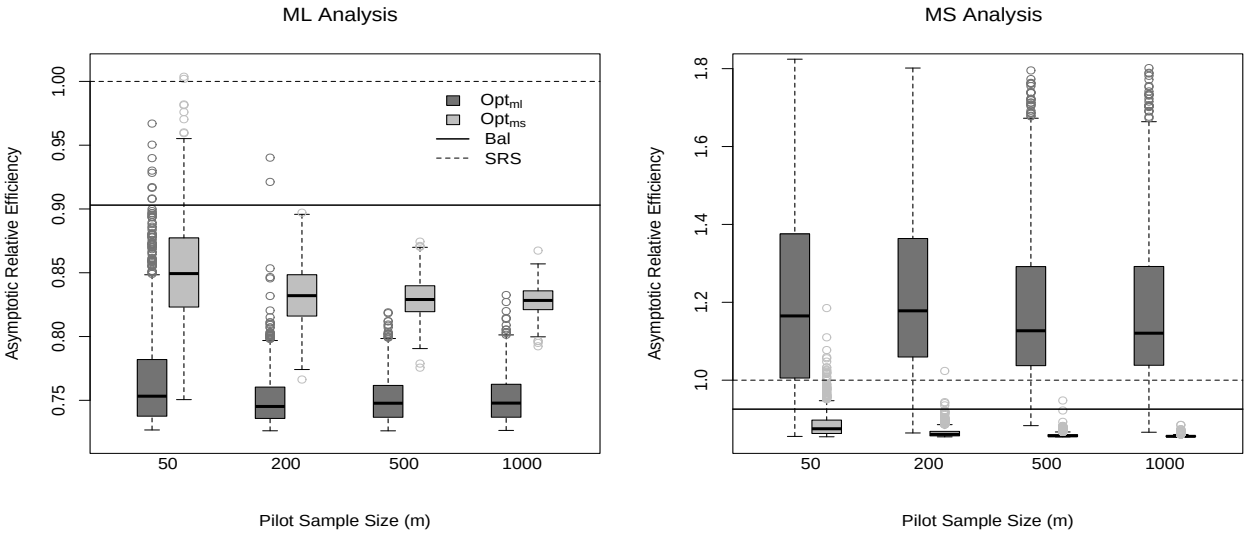

Figure 2: Asymptotic efficiencies of designs relative to SRS as a function of the size of the pilot studies used to estimate the design parameters with a continuous covariate X. The balanced and SRS designs do not utilize the pilot data and are not affected by its size.

REFERENCES

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973.

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

Breslow, N. E. & Chatterjee, N. (1999). design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Applied Statistics*, 48(4), 457–468.

Breslow, N. E. & Holubkov, R. (1997). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine*, 16(1), 103–116.

Carpenter, J. R., Kenward, M. G., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3), 571–584.

Chandran, V., Cook, R. J., Edwin, J., Shen, H., Pellett, F. J., Shanmugarajah, S., Rosen, C. F., & Gladman, D. D. (2010). Soluble biomarkers differentiate patients with psoriatic arthritis from those with psoriasis without arthritis. *Rheumatology*, 49(7), 1399–1405.

Chatterjee, N., Chen, Y. & Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association*, 98(461), 158–168

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Emery, A. F., & Nenarokomov, A. V. (1999). Optimal experiment design. *Measurement Science and Technology*, 9(6), 864–876.

Gladman, D. D. & Chandran, V. (2011). Observational cohort studies: Lessons learnt from the University of Toronto Psoriatic Arthritis Program. *Rheumatology*, 50, 25–31.

Kulich, M. & Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99(467), 832–844.

Lawless, J. F., Kalbfleisch, J. D. & Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 61(2), 413–438.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., John Wiley & Sons, New York.

Lumley, T., Shaw, P. A., & Dai, J. Y. (2011). Connections between survey calibration estimators and semiparametric models for incomplete data. *International Statistical Review*, 79(2), 200–220.

McIsaac, M. A. (2012). *Statistical Methods for Incomplete Covariates and Two-Phase Designs*. (Doctoral dissertation). Retrieved from UWSpace Electronic Theses and Dissertations. http://hdl.handle.net/10012/7259.

Okada, Y., Shinmei, M., Tanaka, O., Naka, K., Kimura, A., Nakanishi, I., Bayliss, M. T., Iwata, K., & Nagase, H. (1992). Localization of matrix metalloproteinase 3 (stromelysin) in osteoarthritic cartilage and synovium. *Laboratory investigation; a journal of technical methods and pathology*, 66(6), 680–690.

Pepe, M. S., Reilly, M., & Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*, 42(1), 137–160.

Reilly, M. (1996). Optimal sampling strategies for two phase studies. *American Journal of Epidemiology*, 143, 92–100.

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

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846–866.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995). analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429), 106–121.

Scott, A. J., & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1), 57–71.

Scott, A. J., & Wild, C. J. (2011). Discussions. *International Statistical Review*, 79(2), 228–230

Tosteson, T. D., & Ware, J. H. (1990). Designing a logistic regression study using surrogate measures for exposure and outcome. *Biometrika*, 77(1), 11–21.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer Science + Business Media.

Yu, M., & Nan, B. (2006). A revisit of semiparametric regression models with missing data. *Statistica Sinica*, 16(4), 1193–1212.

Whittemore, A. S., & Halpern, J. (1997). Multi-stage sampling in genetic epidemiology. *Statistics in Medicine*, 16, 153–167.

Zhao, Y., Lawless, J. F., & McLeish, D. L. (2009). likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal*, 51(1), 123–136.