Scandinavian Journal of Statistics, Vol. 39: 131–152, 2012
doi: 10.1111/j.1467-9469.2011.00745.x

A General Statistical Framework for Multistage Designs

MARIA GRÜNEWALD and OLA HÖSSJER

Department of Mathematics, Stockholm University

ABSTRACT. The efficiency of observational studies may be increased by applying multistage sampling designs. It is, however, not always transparent how to construct such a design to obtain increased efficiency. We here present a general statistical framework for describing and constructing multistage designs. We also provide tools for efficiency and cost-efficiency comparisons, to facilitate the choice of sampling scheme. The comparisons are based on Fisher information matrices and the results are presented in graphs, where either efficiency or cost-adjusted efficiency is plotted against a normalized measure of cost. The former curve resides in the unit square and is analogous to the receiver operating characteristic curve used for testing.

Key words: cost-efficiency, efficient design, Fisher information, hierarchical multistage model, multistage sampling

1. Introduction

Likelihood-based methods exhibit good efficiency under rather weak regularity conditions when the underlying model is correctly specified, see, for instance, Lehmann & Casella (1988). However, if the cost of collecting the full sample is large, one may collect a subsample at lower total cost, which, by careful choice of the sampling mechanism, only looses little efficiency compared with the full maximum likelihood (ML) estimator or likelihood ratio (LR) test. To choose an effective design, it is necessary to be able to calculate, and to compare, the effectiveness of different sampling strategies. The aim of this article is to provide tools for such comparisons within a multistage design framework that is general enough to be applicable for a broad range of statistical models and sampling schemes.

The term two-stage design was introduced by White (1982), and has since then been explored for different settings in various areas of research. Stage 1 data are first collected for all sampled individuals and Stage 2 data subsequently for a subset of them. It is motivated by differential costs and informativity of collecting data on different variables and individuals. Maydrech & Kupper (1978) allow for different costs of exposed/non-exposed or cases/controls when calculating required sample sizes for cohort and case–control studies. Reilly (1996) investigates optimal allocation of available resources for two-stage data, where either precision is maximized for a fixed budget or cost is minimized for a fixed precision. Thomas *et al.* (2004) provide cost-efficiency calculations for a genetic application where association between single-nucleotide polymorphism markers and disease status are tested, and the two-stage design is motivated by the cost of genotyping. Further examples of two-stage designs are provided in section 9.

Designs with $k \geq 2$ stages is the natural generalization of two-stage designs, with Stages 1 and k corresponding to minimal and full information. For a bottom-up design (BUD), individuals sequentially enter higher stages, and often data are increasingly more costly to collect. For instance, if the objective is to determine risk factors of a disease, a possible three-stage BUD is based on collecting registry data (affection status, sex, age, ...) at Stage 1, questionnaire variables (smoking, exercise, diet, at family history of disease, ...) at Stage 2 and more

expensive biological variables that require laboratory work (genetic maps, expression levels of DNA, protein data, ...) at Stage 3. Alternatively, for a top-down design (TDD), data are initially available up to Stage k for all individuals, but is then coarsened down to lower stages (Heitjan & Rubin, 1991). Here, the cost must be interpreted in a different way, since data are already collected. For instance, the cost of computing an estimate or ethical restrictions may force the data analyst to remove data at higher stages, she may choose to disregard uninformative data or she may act as a fusion centre, receiving information from several external sources that is compressed to various degrees.

Since a general statistical theory of multistage designs is lacking, it is the purpose of this article to contribute to filling this gap. Mathematically, the top-down perspective is more general. Therefore, we introduce a hierarchical multistage model where the full data set for an individual is coarsened down to Stage $J \in \{1,\ldots,k\}$ and retained at this level. The sampling scheme π is defined as the distribution of J, and the aim is to choose it in a cost-effective way. In contrast, in most applications, a bottom-up perspective is more appropriate, with data collected sequentially from Stage 1 up to Stage J. If uncollected data (for BUDs) or data lost after coarsening (for TDDs) is missing at random (Rubin, 1976; Little & Rubin, 2002), we obtain a Missing at Random Design (MARD). A BUD is always, and a TDD is sometimes, a MARD.

To systematically describe the efficiency-cost tradeoff, we introduce plots of efficiency and cost-adjusted efficiency as functions of average cost. We use the full sampling scheme ($J \equiv k$) as reference, and thus report efficiency as well as average cost in relative terms. Our framework can be viewed as a generalization of Grünewald & Hössjer (2010a), where two-stage retrospective designs are treated.

Our efficiency calculations are based on ML estimation. The likelihood theory for two-stage designs is well developed, with algorithms, Fisher information, asymptotic normality, variance and efficiency derived for estimators within quite a large class of models, see, for instance, Scott & Wild (1997), Breslow & Holubkov (1997), Breslow *et al.* (2003) and references therein. Here, we generalize some of these findings by deriving the Fisher information and efficiency of a general design and simplify these formulas for MARDs, a result we believe is of independent interest.

This article is organized as follows: in section 2, the multistage sampling model is described. The cost and efficiency of samples are defined in section 3 and the choice of cost function is discussed in section 4. Section 5 defines stage-dependent cost functions and MARDs, whereas strategies for comparing, visualizing and comparing designs are presented in section 6. In section 7, we outline how Monte Carlo methods can be used to approximate efficiency and cost. Related to multi-stage designs is the ascertainment problem, treated in section 8, where data are not recorded on units without full data. In section 9, the general theory is illustrated with a number of examples. The calculations are run in the software R (R Development Core Team, 2008), and the code used to produce efficiency and cost efficiency plots is available at http://www2.math.su.se/∼ola/, or at request from the authors. Finally, the main conclusions of the article are discussed in section 10, and proofs are collected in the appendix S1.

2. A multistage model

2.1. Data coarsening or level of sampling

Let Z denote full data of an individual, a random variable defined on a sample space $\mathcal{Z}$. Then, introduce a sequence of reduced sampling spaces $\mathcal{Z} = \mathcal{Z}_k, \mathcal{Z}_{k-1}, \ldots, \mathcal{Z}_1$, where reduction of complexity (or coarsening) from Stage $j+1$ to Stage j is achieved by means of the

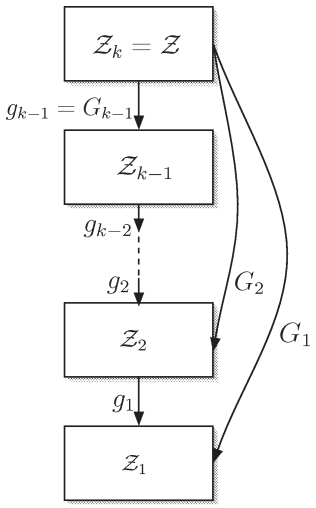

Fig. 1. A multistage model where sampling spaces are reduced sequentially. Stage k represents the most complex sampling space and Stage 1 represents the most sparse one.

non-invertible transformation $g_j : \mathcal{Z}_{j+1} \to \mathcal{Z}_j$ and the coarsening from Stage k down to Stage j is described by $G_j = g_j \circ g_{j+1} \circ \ldots g_{k-1}$, as shown in Fig. 1.

Let $Z_j = G_j(Z)$ denote the random variable obtained if full data $Z = Z_k$ are coarsened down to Stage j (or if data are sampled up to Stage j). In general, the observed level of coarsening $J \in \{1, \ldots, k\}$ may be random, giving a coarsened random variable $\tilde{Z} = Z_J$ defined on the combined sample space $\tilde{\mathcal{Z}} = \mathcal{Z}_1 \cup \ldots \cup \mathcal{Z}_k$. The joint distribution of Z and J is determined by

$$\pi_j(z) = P(J = j \mid Z = z),$$

the probability of using information up to Stage j for $z \in \mathcal{Z}$, so that $J \mid Z = z \sim \text{Mult}(1, \pi(z))$ is multinomial, with $\pi(z) = (\pi_1(z), \ldots, \pi_k(z))$. Also, let

$$\Pi_j(z) = P(J \geq j \mid Z = z) = \sum_{l=j}^{k} \pi_l(z)$$

be the probability that information on z is obtained at least up to Stage j.

For a TDD, a more relevant variable is

$$\lambda_j^{\text{down}}(z) = P(J \leq j \mid J \leq j+1, z) = [1 - \Pi_{j+1}(z)]/[1 - \Pi_{j+2}(z)],$$

the conditional probability of coarsening data down to Stage j given that it has already been coarsened down to Stage $j+1$, $j = k-1, \ldots, 1$. A TDD example is given in section 9. For a BUD, we rather use

$$\lambda_j^{\text{up}}(z) = P(J \geq j \mid J \geq j-1, \quad Z = z) = \Pi_j(z)/\Pi_{j-1}(z),$$

that is, the conditional probability of collecting data from Stage j, given that data have already been collected from Stage $j-1$, $j = 2, \ldots, k$.

Example 1 (Binary response-selective sampling). Let $z = (x, y)$, where x is a set of covariates and y a binary response. A healthy control individual has $y = 0$ and an affected case $y = 1$. Suppose affection status is available for all individuals, whereas covariates are collected for all cases but only for a small fraction η of controls. We formalize this as a two-stage BUD with

$$\begin{aligned} z_1 &= y, \\ z_2 &= (x, y), \end{aligned} \tag{1}$$

$\lambda_2^{\mathrm{up}}(x,1)=1$, $\lambda_2^{\mathrm{up}}(x,0)=\eta$. The idea behind this design is that little efficiency is lost for the reduced sample ($\eta<1$) when estimating the effect parameter in a logistic regression model compared with that of the much more costly full sample ($\eta=1$).

2.2. Likelihood-based inference

Assume Z has a density $f_k(z;\theta)$ w.r.t. some underlying measure μ_k on $\mathcal{Z}_k$, where $\theta=(\theta_1,\ldots,\theta_p)\in\Theta$ is the p-dimensional parameter vector which we wish to make inference on. To this end, there is a collection of i.i.d. full data random variables $Z^1,\ldots,Z^n$ defined on a $\mathcal{Z}$ with the same common density $f_k(z;\theta)$ as Z.

Example 2 (Regression model). In example 1, assume $\theta=(\gamma,\xi)$ and

$$f_2(z;\theta)=P(x;\gamma)P(y\mid x;\xi), \tag{2}$$

where γ contains nuisance parameters involved in the covariate distribution, and ξ the regression parameters. For the logistic regression model, the response is binary, with

$$P(y\mid x;\xi)=F(\alpha+\beta x^{\mathrm{T}})^{\{y=1\}}\left[1-F(\alpha+\beta x^{\mathrm{T}})\right]^{\{y=0\}}, \tag{3}$$

where $\xi=(\alpha,\beta)$ consists of one intercept parameter α, a number of slope parameters β and $F(x)=\exp(x)/(1+\exp(x))$.

Let J^i be the stage up to which we have information on individual i, either due to coarsening down to this stage or sampling information up to this stage. Then, $(Z^1,J^1),\ldots,(Z^n,J^n)$ is an i.i.d. sequence of random variables, not fully observed (or which we do not use all information from). Instead, only the reduced i.i.d. sample $\tilde{Z}^1,\ldots,\tilde{Z}^n$, where $\tilde{Z}^i=Z^i_{J^i}$, is used for inference.

If estimation of θ is of concern, we may employ the ML estimator

$$\hat{\theta}_{\mathrm{ML}}(\pi)=\arg\max_{\theta\in\Theta} L(\theta,\pi), \tag{4}$$

where

$$L(\theta,\pi)=\prod_{i=1}^{n} f(\tilde{z}^i;\theta,\pi) \tag{5}$$

is the likelihood function, $\tilde{z}^i$ the observed value of $\tilde{Z}^i$, $\pi=\{\pi(z);z\in\mathcal{Z}\}$ the (possibly infinite-dimensional) sampling (coarsening) parameter and $f(\cdot;\theta,\pi)$ the density of $\tilde{Z}$ on $\tilde{\mathcal{Z}}$. The density of $\tilde{Z}$ is

$$f(\tilde{z};\theta,\pi)=f_j(z_j)E[\pi_j(Z)\mid G_j(Z)=z_j],\quad \text{if } \tilde{z}=z_j\in\mathcal{Z}_j, \tag{6}$$

relative to some reference measure μ_j on $\mathcal{Z}_j$, typically the counting measure for discrete $\mathcal{X}_j$ or the Lebesgue measure for Euclidean $\mathcal{X}_j$. For proof, see appendix S1.

Example 3 (Binary response-selective sampling, continued). Continuing examples 1 and 2, we find that

$$f(\tilde{z})=\begin{cases}(1-\pi_2(y))\int P(x;\gamma)P(y\mid x;\xi)\,\mathrm{d}x, & \text{if } \tilde{z}=y,\\ \pi_2(y)P(x;\gamma)P(y\mid x;\xi), & \text{if } \tilde{z}=(x,y).\end{cases}$$

If we wish to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$, given some null parameter set $\Theta_0 \subset \Theta$, we may use the log-LR test statistic

$$T_{\mathrm{LR}}(\pi) = 2\Big[\max_{\theta\in\Theta} \log L(\theta, \pi) - \max_{\theta\in\Theta_0} \log L(\theta, \pi)\Big], \tag{7}$$

with H_0 rejected if T_{LR} exceeds a given threshold.

The full sample corresponds to $\pi_k(\cdot) \equiv 1$. We denote this sampling scheme by π_{full}, so that $\theta \to L(\theta, \pi_{\mathrm{full}})$ is the ordinary likelihood function. At the other extreme, we let π_{min} denote the design $\pi_1(\cdot) \equiv 1$, yielding a data set with minimal possible amount of information.

3. Cost and efficiency

Let $C_j(z)$ be the individual cost of using data $z \in \mathcal{Z}$ up to Stage j. Assume

$$0 \le C_1(z) \le \cdots \le C_k(z), \quad \forall z \in \mathcal{Z}, \tag{8}$$

so that the cost increases when more information on z is used. For a BUD, $C_j(z) - C_{j-1}(z)$ is the cost of sampling data from z at Stage j, whereas for a TDD, it is the cost of retaining data from z at Stage j not present at Stage $j-1$. The total average cost (TAC) of the cost-reduced sample is

$$\mathrm{TAC}(\theta, \pi) = nE(C_J(Z)) = n\sum_{j=1}^{k} \int_{\mathcal{Z}} \pi_j(z) C_j(z) f_k(z; \theta)\, \mathrm{d}\mu_k(z), \tag{9}$$

and the relative average cost $\mathrm{RAC}(\theta, \pi) = \mathrm{TAC}(\theta, \pi)/\mathrm{TAC}(\theta, \pi_{\mathrm{full}})$ compared with the full sample. Let

$$\psi(\tilde{z}; \theta, \pi) = \frac{\partial \log f(\tilde{z}; \theta, \pi)}{\partial \theta} \tag{10}$$

be the score function, which is a $1 \times p$ vector-valued function defined on $\tilde{\mathcal{Z}}$. The Fisher information of $\{\tilde{Z}^i\}_{i=1}^n$ is a $p \times p$ matrix

$$I(\theta, \pi) = nE[\psi(\tilde{Z}; \theta, \pi)^{\mathrm{T}} \psi(\tilde{Z}; \theta, \pi)], \tag{11}$$

where ψ^{T} is the transpose of ψ. Let $h(I)$ be a scalar function of I satisfying

$$\begin{aligned} &h(tI) = th(I) \quad \text{for any } t > 0, \\ &h(I_1) \le h(I_2) \quad \text{if } I_1 \le I_2, \end{aligned} \tag{12}$$

where $I_1 \le I_2$ means that $I_2 - I_1$ is positive semidefinite. We define the relative efficiency of the cost-reduced sample compared with the full one as

$$e(\theta, \pi) = h[I(\theta, \pi)]/h[I(\theta, \pi_{\mathrm{full}})]. \tag{13}$$

The first part of (12) ensures that e has the usual interpretation in terms of relative sample sizes: asymptotically, when n is large, a sample of size $n/e(\theta, \pi)$ is needed for design π to attain the same accuracy as a sample of size n using the full design π_{full}.

The cost-adjusted efficiency,

$$\mathrm{CE}(\theta, \pi) = e(\theta, \pi)/\mathrm{RAC}(\theta, \pi),$$

quantifies the relative efficiency of design π at parameter θ compared with a simple random sampling (SRS) with sample size $\mathrm{RAC}(\theta, \pi)n$, which exhibits the same TAC. It thus summarizes with a single number whether π is cost efficient ($\mathrm{CE} > 1$) or not ($\mathrm{CE} < 1$). It is frequently referred to as asymptotic relative cost efficiency and has been used by several authors (cf., e.g. Reilly, 1996; McNamee, 2003; Thomas *et al.*, 2004, and references therein).

4. Choice of efficiency function

If the estimation of θ is of interest, $h(I)$ in (12) is typically a function of the asymptotic covariance matrix $V = I^{-1} = (V_{rs})_{r,s=1}^{p}$. If θ_r is the parameter of main interest, $h = V_{rr}^{-1}$ is a natural choice. In example 8, we define h for a three-stage design to depend on the asymptotic variance of three effect parameters, which are the parameters of main interest in this model. For other examples, such as $\det(V)^{-1/p}$ and $\mathrm{tr}(V)^{-1}$, see, for example, Silvey (1980) and Melas (2006).

For testing, other functions can be used. Assume a simple null hypothesis $\Theta_0 = \{\theta_0\}$, and a true parameter value $\theta_0 + a$. Then asymptotically, in the limit of large samples (large I) and local alternatives (small a), $T_{\mathrm{LR}}(\pi)$ in (7) has a non-central χ^2 distribution with p degrees of freedom and non-centrality parameter $h(I) = a^{\mathrm{T}} I a$, where $I = I(\theta_0, \pi)$ (see, for instance, Serfling, 1980). Hence, the power of the LR test is asymptotically a monotone function of $h(I)$. More general choices of h for testing are discussed in Grünewald & Hössjer (2010b).

If $\theta = (\xi, \gamma)$ can be split into structural parameters ξ and nuisance parameters γ, with parameter space $\Theta = \Xi \times \Gamma$, we write

$$I(\theta, \pi) = \begin{pmatrix} I_{\xi\xi}(\theta, \pi) & I_{\xi\gamma}(\theta, \pi) \\ I_{\gamma\xi}(\theta, \pi) & I_{\gamma\gamma}(\theta, \pi) \end{pmatrix}. \tag{14}$$

The estimation-based functions h are defined as before. For testing, consider a composite null hypothesis $\Theta_0 = \{\xi_0\} \times \Gamma$. Then, an appropriate function, when $\xi_0 + a$ is the true structural parameter, is $h(I) = a I_{\mathrm{profile}} a^{\mathrm{T}}$, where $I_{\mathrm{profile}}(\theta, \pi) = I_{\xi\xi} - I_{\xi\gamma} I_{\gamma\gamma}^{-1} I_{\gamma\xi}$. is the profile likelihood Fisher information. More generally, $h(I) = \tilde{h}(I_{\mathrm{profile}})$ can be used, where $\tilde{h}$ is a function used for testing when the nuisance parameters are known.

5. MARDs and stage-dependent cost functions

We define the class $\mathcal{P}$ of MARDs as $\pi_j(\cdot)$ being constant on sets $G_j^{-1}(z_j) = \{z \in \mathcal{Z}; G_j(z) = z_j\}$. We will use the somewhat sloppy notation

$$\pi_j(z) = \pi_j(z_j), \tag{15}$$

to denote this, not only for π_j, but also for other functions; and (15) means that the probability of including information from z up to but not exceeding Stage j should not depend on information about z at stages above j. The name MARD is more transparent if we notice that (15) is equivalent to

$$\Pi_j(z) = \Pi_j(z_{j-1}), \tag{16}$$

and hence $1 - \Pi_j(z)$, the probability of 'missing' data above Stage $j-1$ should only depend on parts of z available up to Stage $j-1$, which corresponds exactly to the MAR condition of Rubin (1976). We always require a BUD to be MAR, and hence it is easily seen that $\lambda_j^{\mathrm{up}}(z) = \lambda_j^{\mathrm{up}}(z_{j-1})$, that is, the probability of collecting more data only depends on the data already present.

With condition (15), the density function of $\tilde{Z}$ simplifies to

$$f(\tilde{z}; \theta, \pi) = \pi_j(z_j) f_j(z_j; \theta), \quad \text{if } \tilde{z} = z_j \in \mathcal{Z}_j, \tag{17}$$

where $f_j(z_j; \theta)$ is the density of Z_j, cf. appendix S1. Inserting (17) into (10) and (11), we show in the appendix S1 that the Fisher information matrix can be expressed as a sum of k terms in two ways, either

$$I(\theta,\pi)=\sum_{j=1}^{k} I_j(\theta,\pi), \tag{18}$$

with I_j the part of all information obtained from individuals sampled up to (but not above) Stage j, or

$$I(\theta,\pi)=I(\theta,\pi_{\min})+\sum_{j=2}^{k} I_{j|j-1}(\theta,\pi), \tag{19}$$

where $I_{j|j-1}$ is the total information from Stage j data (i.e. not present at Stage $j-1$).

For cost functions, a simplification,

$$C_j(z)=C_j(z_j), \tag{20}$$

similar to (15) is possible, which may be violated for TDDs (cf. example 14), but always holds for BUDs, since the cost of gathering information at Stage j does not depend on information from stages above j not yet present. We refer to (20) as a *stage-dependent* cost function, for which (9) becomes

$$\begin{aligned}\mathrm{TAC}(\theta,\pi)&=n\sum_{j=1}^{k}\int_{\mathcal{Z}_j}\pi_j(z_j)C_j(z_j)f_j(z_j;\theta)\,\mathrm{d}\mu_j(z_j)\\&=n\sum_{j=1}^{k}\int_{\mathcal{Z}_j}\Pi_j(z_{j-1})[C_j(z_j)-C_{j-1}(z_{j-1})]f_j(z_j;\theta)\,\mathrm{d}\mu_j(z_j),\end{aligned} \tag{21}$$

where, for a BUD, the terms of the first sum regard the average cost of sampling individuals up to but not above Stage j, whereas the terms of the second sum regard the average cost of sampling at Stage j.

6. Cost-efficiency plots and optimal designs

6.1. Cost-efficiency

The tradeoff between cost and efficiency at a given parameter θ can be illustrated by varying π and plotting $e(\theta,\pi)$ as a function of $\mathrm{RAC}(\theta,\pi)$ (see Grünewald & Hössjer, 2010a). Some simple but useful properties of such cost-efficiency plots are summarized as follows.

Proposition 1. *Consider a fixed $\theta\in\Theta$ and function h satisfying* (12). *Let π,π' be two designs with $\pi\le\pi'$, that is, $\Pi_j(z)\le\Pi'_j(z)$ for all $z\in\mathcal{Z}$ and $j=1,\ldots,k$. Then,*

$$\begin{aligned}&0\le\mathrm{RAC}(\theta,\pi_{\min})\le\mathrm{RAC}(\theta,\pi)\le\mathrm{RAC}(\theta,\pi')\le\mathrm{RAC}(\theta,\pi_{\mathrm{full}})=1,\\&0\le e(\theta,\pi_{\min})\le e(\theta,\pi)\le e(\theta,\pi')\le e(\theta,\pi_{\mathrm{full}})=1,\end{aligned} \tag{22}$$

with equalities $0=\mathrm{RAC}(\theta,\pi_{\min})$ and $0=e(\theta,\pi_{\min})$ on the left-hand sides of (22) *if Stage* 1 *corresponds to no cost ($C_1(\cdot)\equiv 0$) and no information ($I(\theta,\pi_{min})=0$), respectively.*

It follows from proposition 1 that each design π corresponds to a point in the unit square $[0,1]\times[0,1]$ of the (RAC,e)-plane, with $(1,1)$ for the full design and $(0,0)$ for the minimal design if Stage 1 has zero cost and no information.

6.2. Optimal designs

Consider a finite-dimensional subclass

$$\mathcal{Q}=\{\pi\in\mathcal{P};\pi(\cdot)=\pi(\cdot;\eta)\text{ for some }\eta\} \tag{23}$$

of all MARDs, parameterized by $\eta=(\eta_1,\ldots,\eta_r)$. Keep $\theta\in\Theta$ fixed and define a $\mathcal{Q}$-optimal cost-efficiency curve

$$R\to e_{\max}(\theta,R)=\sup_{\pi\in\mathcal{Q}_R} e(\theta,\pi), \tag{24}$$

where $\mathcal{Q}_R=\{\pi\in\mathcal{Q};\mathrm{RAC}(\theta,\pi)\le R\}$. Any design attaining the maximum in (24) is referred to as $\mathcal{Q}$-optimal, since it maximizes $e(\theta,\cdot)$ over $\mathcal{Q}$ subject to a cost constraint $\mathrm{RAC}(\theta,\pi)\le R$. It will typically be locally optimal, that is, depend on the unknown θ. In practice, we may use training data to compute a preliminary estimate of θ, which is used as plug-in for the optimal design.

Proposition 2. *Assume h satisfies* (12) *and that $\mathcal{Q}$ is convex with $\pi_{full}\in\mathcal{Q}$. Then, given any $\theta\in\Theta$, the optimal efficiency curve* (24) *satisfies*

$$e_{\max}(\theta,R)=\sup_{\pi\in\mathcal{Q};\mathrm{RAC}(\theta,\pi)=R} e(\theta,\pi), \tag{25}$$

for any $R\in(R_{\min},1]$, where $R_{\min}=\min_{\pi\in\mathcal{Q}}\mathrm{RAC}(\theta,\pi)$. Hence, the maximal cost efficiency satisfies

$$\mathrm{CE}_{\max}(\theta,R):=\sup_{\{\pi\in\mathcal{Q};\mathrm{RAC}(\theta,\pi)=R\}}\mathrm{CE}(\theta,\pi)=e_{\max}(\theta,R)/R.$$

Proposition 2 applies with $\mathcal{Q}=\mathcal{P}$, since $\mathcal{P}$ is convex by definition. It is easily seen that π_{full} is optimal, with $(1,1)$ the right-hand end point of the optimal-efficiency curve. The minimal design $\pi_{\min}$ is optimal as well if we have strict inequalities in (8) and then $(R_{\min},e_{\min})$ is the left-hand end point of the optimal-efficiency curve, where $e_{\min}=\min_{\pi\in\mathcal{Q}}e(\theta,\pi)$. Grünewald & Hössjer (2010b) establish some other properties of the $e_{\max}$ and $\mathrm{CE}_{\max}$ curves.

6.3. ROC curve analogy

The performance of a statistical test, which rejects H_0 when a test statistic T exceeds a given threshold t, is usually assessed by reporting the significance level $\alpha(t)=P(T\ge t\,|\,H_0)$ and power $\beta(\theta,t)=P(T\ge t\,|\,\theta)$ for $\theta\in\Theta\setminus\Theta_0$. For a given θ, we may vary the threshold and obtain a receiver operator characteristic (ROC) curve by plotting β against α within the unit square. This could be compared with a cost-efficiency plot, with (RAC,e) replacing (α,β) and instead of a threshold the single parameter η of a one-dimensional $\mathcal{Q}$ is varied. Interestingly, the cost-adjusted efficiency corresponds to

$$\frac{\beta}{\alpha}=\frac{1-c}{c}\left[\frac{1}{\mathrm{FDR}}-1\right],$$

where $\mathrm{FDR}=(1-c)\alpha/[(1-c)\alpha+c\beta]$ is the false discovery rate, that is, the expected rate of false positives within a large population in which a fraction $1-c$ is drawn from the null distribution and a fraction c from a distribution parameterized by θ. Hence, for fixed c, maximizing CE is analogous to minimizing FDR in multiple testing.

7. Computation

For complex models, it may be needed to approximate efficiency (18) and cost (21) by, for instance, Monte Carlo. To this end, generate an i.i.d. sample $\{Z^i\}_{i=1}^N$ from $f_k(\cdot;\theta)$ and estimate

$$\begin{aligned}
\widehat{\mathrm{TAC}}(\theta,\pi) &= nN^{-1}\sum_{i=1}^{N}\sum_{j=1}^{k}\pi_j(Z_j^i)C_j(Z_j^i),\\
\widehat{\mathrm{RAC}}(\theta,\pi) &= \widehat{\mathrm{TAC}}(\theta,\pi)/\widehat{\mathrm{TAC}}(\theta,\pi_{\mathrm{full}}),\\
\hat{I}(\theta,\pi) &= nN^{-1}\sum_{i=1}^{N}\sum_{j=1}^{k}\pi_j(Z_j^i)\psi_j(Z_j^i;\theta)^{\mathrm{T}}\psi_j(Z_j^i;\theta),\\
\hat{e}(\theta,\pi) &= h[\hat{I}(\theta,\pi)]/h[\hat{I}(\theta,\pi_{\mathrm{full}})],
\end{aligned} \tag{26}$$

where $Z_j^i = G_j(Z^i)$. Since $\psi_j(Z_j^i;\theta)$ is defined by means of an integral when $j<k$ (cf. appendix S1, eq. (S.1)), we may need to approximate it by a Monte Carlo estimate $\hat{\psi}_j(Z_j^i;\theta)$. When $\mathcal{Z}_j$ is finite, we put

$$\hat{\psi}_j(z_j;\theta) = \frac{1}{N_{z_j}}\sum_{i;Z_j^i=z_j}\psi_j(Z^i;\theta),$$

provided $N_{z_j} = |\{i; Z_j^i = z_j\}|$ is positive. When $\mathcal{Z}_j$ is continuous, more refined methods can be used, for example, based on non-parametric regression.

If $I(\theta,\pi)$ and $e(\theta,\pi)$ are to be computed for several θ, one may use importance sampling (Hammersley & Handscomb, 1964). Instead of generating a new sample for each θ, it suffices to use *one* sample $\{Z^i\}_{i=1}^N$ from $f_k(\cdot;\theta')$, with

$$\begin{aligned}
\widehat{\mathrm{TAC}}(\theta,\pi) &= nN^{-1}\sum_{i=1}^{N}\sum_{j=1}^{k}\pi_j(Z_j^i)C_j(Z_j^i)w(Z^i;\theta),\\
\hat{I}(\theta,\pi) &= nN^{-1}\sum_{i=1}^{N}\sum_{j=1}^{k}\pi_j(Z_j^i)\psi_j(Z_j^i;\theta)^{\mathrm{T}}\psi_j(Z_j^i;\theta)w(Z^i;\theta),
\end{aligned} \tag{27}$$

and weight function $w(z;\theta) = f_k(z;\theta)/f_k(z;\theta')$. Each term $\psi_j(Z_j^i;\theta)$ with $j<k$ may be replaced by an estimate $\hat{\psi}_j(Z_j^i;\theta)$. For discrete $\mathcal{Z}_j$,

$$\hat{\psi}_j(z_j;\theta) = \sum_{i;Z_j^i=z_j}\psi_j(Z_j^i;\theta)w(Z^i;\theta)\Big/\sum_{i;Z_j^i=z_j}w(Z^i;\theta).$$

The accuracy of importance sampling is sometimes poor if the candidate parameter θ' is far away from θ (Hesterberg, 1995). Approximate $\mathcal{Q}$-optimal designs can be found using (26) or (27) and maximizing $\hat{e}(\theta,\cdot)$ over $\hat{\mathcal{Q}}_R = \{\pi\in\mathcal{Q}; \widehat{\mathrm{RAC}}(\theta,\pi)\le R\}$.

8. Ascertainment

Ascertainment is typically regarded as a one-stage model with an underlying complete data set $Z^1,\ldots,Z^n$. Each Z^i is either completely observed or not observed at all, depending on the ascertainment events $A^i = \{Z^i\text{observed}\}$. The sampling scheme $\pi_{\mathrm{asc}}(z) = P(A^i \mid Z^i = z)$ is often not known exactly. For regression models $Z^i = (X^i, Y^i)$, prospective and retrospective designs mean that observations are ascertained based on their covariates X^i and response variables Y^i, respectively.

It is well known that in general, π_{asc} has to be included into analysis to avoid inconsistent estimators (Fisher, 1934; Rao, 1965), either by relying of previous knowledge of π_{asc}, estimating θ and π_{asc} jointly or conditioning away the impact of π_{asc}. To this end, depending on the type of ascertainment, one may use a likelihood conditioned on ascertainment $\{A^i\}$, a retrospective likelihood conditioned on $\{Y^i\}$, a prospective likelihood conditioned on $\{X^i\}$,

an ascertainment-corrected prospective likelihood conditioned $\{X^i, A^i\}$, a semiparametric profile likelihood (Zhou *et al.*, 2007), or a full likelihood (Grünewald *et al.*, 2010). The case–control study is the most well-known retrospective example of ascertainment, for which the ascertainment scheme can be ignored in statistical analysis in two important cases; effect parameter estimation for logistic regression (Prentice & Pyke, 1979) or Cox regression (Prentice & Breslow, 1978) models.

To highlight the efficiency loss incurred by not including unascertained data, case–control sampling can be thought of as originating from example 1, where data are discarded for individuals without covariate information. In the same way, Grünewald & Hössjer (2010a) view ascertainment as a two-stage problem, where data are retained only for individuals sampled up to Stage 2. We now generalize this to multistage designs, where individuals sampled up to Stage k are ascertained, that is, $A^i = \{J^i = k\}$ and $\pi_{\text{asc}}(z) = \pi_k(z)$. Thus, we think of $\pi = (\pi_1, \ldots, \pi_k)$ as a MARD, although the mathematical development below does not require this.

8.1. Unconditional ascertainment

If it is known which observations that are not ascertained, we get a likelihood

$$L_{\text{asc}}(\theta, \pi) = \prod_{i=1}^{n} f_{\text{asc}}(\tilde{z}^i; \theta, \pi), \tag{28}$$

slightly different from (5), where n is the total population size, including both ascertained and not ascertained individuals,

$$f_{\text{asc}}(\tilde{z}; \theta, \pi) = \begin{cases} \pi_k(z) f_k(z; \theta); & \text{if } \tilde{z} \in \mathcal{Z}_k, \\ 1 - P(A \mid \theta, \pi); & \text{if } \tilde{z} \in \mathcal{Z}_1 \cup \ldots \mathcal{Z}_{k-1}, \end{cases}$$

$A = \{J = k\}$ is the ascertainment event and $P(A \mid \theta, \pi) = P(J = k \mid \theta)$ the probability of ascertainment, respectively. The TAC becomes

$$\text{TAC}_{\text{asc}}(\theta, \pi) = nP(A \mid \theta, \pi) E[C_J(Z) \mid J = k] \tag{29}$$

and the Fisher information (11) changes to

$$I_{\text{asc}}(\theta, \pi) = I_k(\theta, \pi) + nP'(A \mid \theta, \pi)^{\text{T}} P'(A \mid \theta, \pi) / [1 - P(A \mid \theta, \pi)], \tag{30}$$

where I_k is the information from sampled individuals and $P'(A \mid \theta, \pi) = \partial P(A \mid \theta, \pi) / \partial \theta$ the information from unsampled ones. Efficiency is calculated as $e_{\text{asc}}(\theta, \pi) = h[I_{\text{asc}}(\theta, \pi)] / h[I(\theta, \pi_{\text{full}})]$. Since $I_{\text{asc}}(\theta, \pi) \leq I(\theta, \pi)$, it follows from (12) that $e_{\text{asc}}(\theta, \pi) \leq e(\theta, \pi)$.

8.2. Conditional ascertainment

More commonly, the unascertained observations and hence also n, are unknown. We then condition on ascertainment status and use the likelihood

$$f_{\text{condasc}}(\tilde{z}; \theta, \pi) = \begin{cases} f_{\tilde{Z} \mid J=k}(\tilde{z}; \theta, \pi) = \pi_k(z) f_k(z; \theta) / P(A \mid \theta, \pi), & \text{if } \tilde{z} \in \mathcal{Z}_k, \\ 1, & \text{if } \tilde{z} \in \mathcal{Z}_1 \cup \cdots \cup \mathcal{Z}_{k-1}, \end{cases}$$

for one individual, giving a total likelihood

$$\begin{aligned} L_{\text{condasc}}(\theta,\pi) &= \prod_{i=1}^{n} f_{\text{condasc}}(\tilde{z}^i;\theta,\pi) \\ &= \prod_{i;J^i=k} f_{\text{condasc}}(\tilde{z}^i;\theta,\pi) \\ &= \prod_{i;J^i=k} \pi_k(z^i) f_k(z^i;\theta)/P(A\mid\theta,\pi) \\ &\propto P(A\mid\theta,\pi)^{-n^{\text{asc}}} \prod_{i;J^i=k} f_k(z^i;\theta), \end{aligned}$$

where $n^{\text{asc}} = |\{i; 1 \le i \le n, J^i = k\}|$ and $E(n^{\text{asc}}) = nP(A\mid\theta,\pi)$. The resulting total average cost $\text{TAC}_{\text{condasc}}(\theta,\pi)$ is the same as for unconditional ascertainment, cf. (29), but the Fisher information (11) changes to

$$\begin{aligned} I_{\text{condasc}}(\theta,\pi) &= nP(A\mid\theta,\pi)\text{cov}[\psi(\tilde{Z};\theta,\pi)] \\ &= I_k(\theta,\pi) - nP'(A\mid\theta,\pi)^{\mathrm{T}} P'(A\mid\theta,\pi)/P(A\mid\theta,\pi), \end{aligned}$$

where $\text{cov}[\psi(\tilde{Z})]$ is the $p \times p$ covariance matrix of $\psi(\tilde{Z})$. Efficiency is calculated as $e_{\text{condasc}}(\theta,\pi) = h[I_{\text{condasc}}(\theta,\pi)]/h[I(\theta,\pi_{\text{full}})]$. Since $I_{\text{condasc}}(\theta,\pi) \le I_{\text{asc}}(\theta,\pi)$, it follows from (12) that $e_{\text{condasc}}(\theta,\pi) \le e_{\text{asc}}(\theta,\pi)$.

9. Examples

Example 4 (Two-stage designs and missing data). Samples with missing data can be viewed as a two-stage design, where the Stage 2 variables are missing for some individuals and observed for others. If there is only a cost involved in collecting the Stage 2 variables, we put

$$C_1(\cdot) \equiv 0, \quad C_2(\cdot) \equiv 1, \tag{31}$$

which implies

$$\text{RAC}(\theta,\pi) = P(J=2) = \int_{\mathcal{Z}} \pi_2(z) f_2(z;\theta)\,\mathrm{d}z. \tag{32}$$

Since $k = 2$, the design is characterized by $\pi_2(\cdot)$. For a MARD,

$$\pi_2(z) = \pi_2(z_1), \tag{33}$$

that is, the probability of including Stage 2 data only depends on Stage 1 data.

When no variables are collected in Stage 1, $\mathcal{Z}_1 = \emptyset$, and data are either completely missing or completely observed. This is essentially equivalent to unconditional ascertainment, with Stages $1,\ldots,k-1$ condensed into one new Stage 1. Hence, the likelihood and the Fisher information can be obtained from (28) and (30). Further, (33) simplifies to

$$\pi_2(\cdot) \equiv \eta, \tag{34}$$

for some $0 \le \eta \le 1$, giving a one-dimensional class of MARDs. The resulting design is essentially a SRS with sample size $n\eta$ (Bin(n,η) to be exact). Condition (34) is referred to as data missing completely at random (MCAR), cf. Little & Rubin (2002), and a cost-efficiency plot $(\text{RAC}, e) = (\eta,\eta)$ is then located along the diagonal within the unit square.

Example 5 (Design of 'x-random' experiments). Assume full data $z = (x,y)$, where x is a set of covariates and y the response, and a two-stage BUD

$$\begin{aligned} z_1 &= x, \\ z_2 &= (x, y). \end{aligned} \tag{35}$$

The density of full data is given by (2), and a MARD satisfies

$$\pi_2(z) = \pi_2(x) = \lambda_2^{\mathrm{up}}(x).$$

Suppose cost function (31) is used, so that $\mathrm{RAC} = \int \pi_2(x) P(x;\gamma)\,\mathrm{d}x$ equals the probability that y is collected for a randomly chosen x, cf. (32). Finding an optimal design $\lambda_2^{\mathrm{up}}(\cdot)$ amounts to deciding for which fraction R of covariates $\{x^i\}_{i=1}^n$ we should collect responses y^i to maximize efficiency when estimating or testing ξ. This is very similar to optimal design, cf. Silvey (1980) and Melas (2006), although we focus on random x. Since x is ancillary for estimating ξ, the likelihood function factorizes as

$$L(\theta, \pi) \propto \prod_{i=1}^{n} P(x^i;\gamma) \cdot \prod_{i;J^i=2} P(y^i \mid x^i;\xi), \tag{36}$$

with proportionality constant not depending on θ. This implies that only the last term of (36), the prospective likelihood, is important for estimating ξ and $I_{\gamma\xi}(\theta,\pi) = I_{\xi\gamma}(\theta,\pi) = 0$ in (14). Hence, any function $h[I(\theta,\pi)]$ that involves estimation or testing of ξ will be a function of $I_{\xi\xi}(\theta,\pi)$ alone. Formula (36) gives rise to a decomposition

$$I(\theta,\pi_{\min}) = \begin{bmatrix} 0 & 0 \\ 0 & I_{\gamma\gamma}(\theta,\pi_{\min}) \end{bmatrix}, \quad I_{2\mid 1}(\theta,\pi) = \begin{bmatrix} I_{\xi\xi}(\theta,\pi) & 0 \\ 0 & 0 \end{bmatrix},$$

of the Fisher information matrix, cf. (19). It follows that (i) the design π has no effect on estimation of γ and (ii) all information about the effect parameters ξ is contained in the prospective likelihood.

Example 6 (Binary response-selective sampling, contd). An important difference of the response-selective design (1) compared with the prospective design (35) is that the x-variables are no longer ancillary and therefore, the first stage sample is informative for inference on ξ. The likelihood of response-selective sampling no longer factorizes as in (36) and any MARD in $\mathcal{P}$ must satisfy $\lambda_2^{\mathrm{up}}(z) = \lambda_2^{\mathrm{up}}(y)$. It is parameterized by $\eta = (\eta_0, \eta_1)$, where $\eta_y = \lambda_2^{\mathrm{up}}(y)$. The subclass $\mathcal{Q}$ with $\eta_1 = 1$ was treated in example 1.

In Fig. 2, efficiency for a logistic regression model with $X \sim \mathrm{Bin}[1, F(\gamma)]$, and $Y \mid X \sim \mathrm{Bin}[1, F(\alpha+\beta)]$ is illustrated, with F the logistic distribution function. To investigate sensitivity to the parameter specification, we simultaneously plot efficiencies for $\beta = 0, 2$ and 4. Note that while the scenario with $\beta = 4$ has the highest efficiency gain from the sampling scheme in the estimation of γ and α, the scenario with $\beta = 0$ has the largest efficiency gain in the estimation of β.

In Fig. 3, RAC for the same logistic regression model as before, with $\beta = 2$, is illustrated. Three different costs are here included, $(C_1, C_2) = (0,1)$, $(1/3, 1)$ and $(1/2, 1)$, representing scenarios where the cost per individual of collecting y is none, half and the same as the cost of collecting x. As x-axis η_0 is chosen. It translates easily into sampling probabilities when planning a study and is a linear transformation of $\mathrm{RAC} = C_1/C_2 + [\eta_0 P(Y=0;\theta) + P(Y=1;\theta)] \times (C_2 - C_1)/C_2$. The graphs show a larger cost-efficiency gain of response-selective sampling for all parameters the smaller C_1 is. Oversampling of cases is most beneficial for α, with a maximum CE over 2.5, followed by β, whereas for γ there is little gain in oversampling cases.

Example 7 (Continuous response-selective sampling, with a genetic application). For continuous y, $\mathcal{P}$ is infinite-dimensional, so we consider a subclass $\mathcal{Q}$

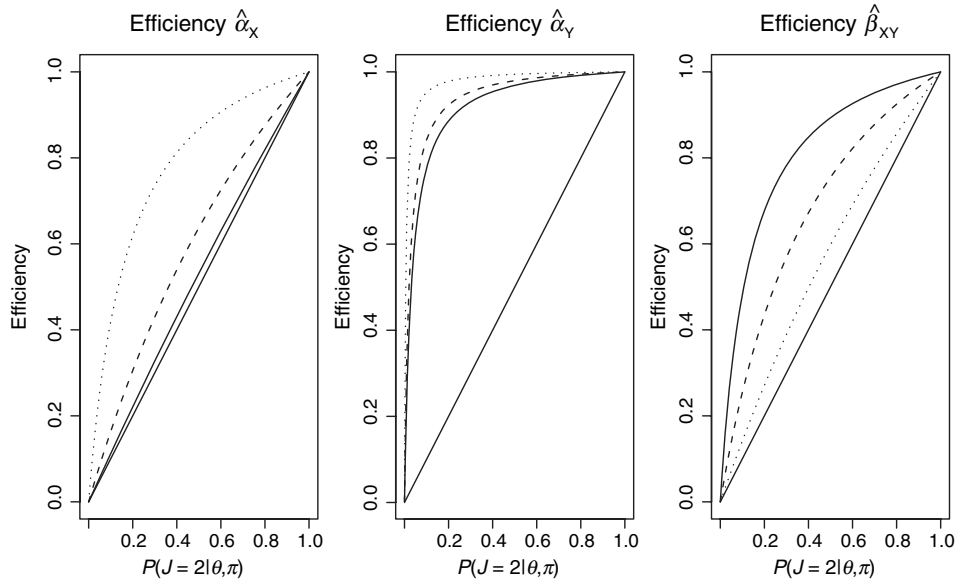

Fig. 2. Efficiency for a logistic regression model. $\gamma = -1$ and $\alpha = -2$. Solid, dashed and dotted lines represent $\beta = 0$, 2 and 4, respectively.

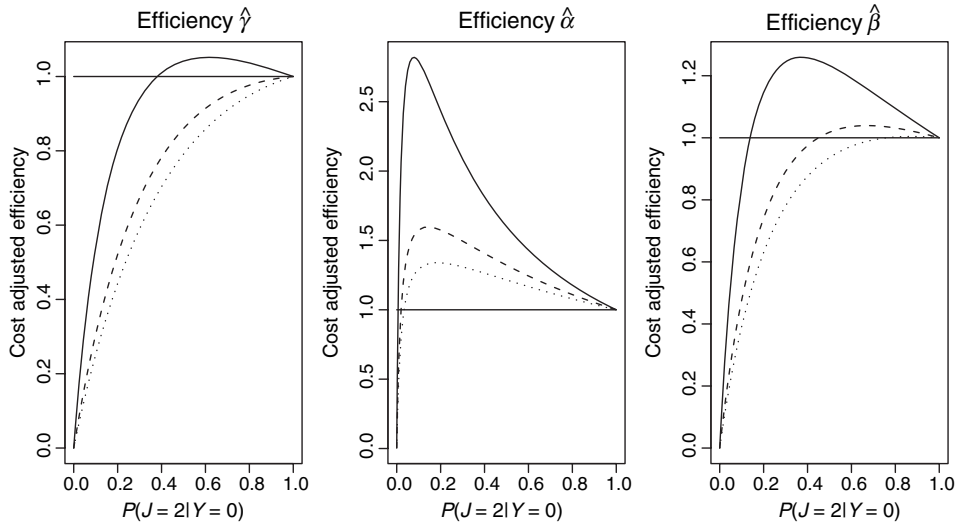

Fig. 3. Cost-adjusted efficiency for a logistic regression model. Solid, dashed and dotted lines represent costs $C_1 = 0, 1/3$ and $1/2$, respectively, for $C_2 = 1$, $\gamma = -1, \alpha = -2, \beta = 2$.

$$\lambda_2^{\mathrm{up}}(y) = \sum_{m=1}^{r} \eta_m 1_{\{y \in \mathcal{S}_m\}}, \tag{37}$$

corresponding to a decomposition $\{\mathcal{S}_m\}_{m=1}^r$ of the response region into r mutually disjoint strata on which the Stage 2 inclusion probability is constant. It is parameterized by $\eta = (\eta_1, \ldots, \eta_r)$, where $0 \le \eta_1, \ldots, \eta_r \le 1$.

Lyon *et al.* (2007) investigate the association between the C/C genotype of genetic marker rs7566605 and body mass index (BMI) in a number of cohorts. One of these samples (Maywood)

was enriched for obese individuals (BMI ≥ 30), whereas four other cohorts (FHS 1, Iceland, KORA S3 and Scandinavia) were not. We investigate if over-sampling of obese individuals would have been cost efficient in these four samples, if analysing the data as two-stage samples with continuous response variables $Y = \text{BMI}$ and binary covariate $X = 1_{\{\text{genotype}=\text{CC}\}}$. Assume $X \sim \text{Bin}[1, F(\gamma)]$, and a linear regression model

$$\log(Y) \mid X = x \sim N(\alpha + \beta x, \sigma^2),$$

for the logarithm of BMI. This yields regression parameters $\xi = (\alpha, \beta, \sigma)$. We use BMI=30 as cut-off value, giving $r = 2$ regions in (37), with $\mathcal{S}_1 = (0, 30)$, $\mathcal{S}_2 = [30, \infty)$ and $\eta = (\eta_1, \eta_2)$, with $\eta_2 = 1$ kept fixed and $0 < \eta_1 < 1$ varying. Parameter values are calculated as shown in Table 1, with no adjustment for age and sex, which was done in Lyon *et al.* (2007). In Fig. 4, the cost-adjusted efficiency for the effect parameter β is illustrated for $(C_1, C_2) = (0, 1)$, $(1/11, 1)$ and $(1/3, 1)$, when the cost of collecting BMI is none, one tenth and half the cost of genotyping, respectively. It is evident that over-sampling of obese subjects for genotyping ($\eta_1 < 1$) is only beneficial if the cost of measuring BMI is substantially lower than the genotyping cost. Over-sampling obese subjects was efficient only in three cohorts, while in the Iceland cohort the most efficient sampling scheme was close to $\eta_1 = 1$. In all four cohorts, the optimal proportions of subjects with BMI ≥ 30 in the samples were approximately 50 per cent (for high cost of genotyping). As seen from Table 1, the mean BMI in the original Iceland cohort was higher than that for the other three cohorts, and very close to the cut-off, so there was already a high proportion of obese individuals in the Iceland sample. However, the proportion of the Icelandic population with a BMI ≥ 30 is 12.4 (Steingrimsdottir *et al.*, 2003), suggesting that the Iceland cohort already has an over-representation of subjects with high BMI compared with the Icelandic population. Some of the participants in the Iceland cohort were 'relatives of probands', so there might be an unintended correlation between BMI and η in this cohort.

Example 8 (Sequential inclusion of covariates). We retain (2), but consider a k-stage BUD, with $z_j = (x_1, \ldots, x_{j-1}, y)$ containing an increasing number of covariates from $x = (x_1, \ldots, x_{k-1})$ as j increases. Then,

$$f(z_j; \theta, \pi) = \pi_j(z_j) P(x_1, \ldots, x_{j-1}; \gamma) \int P(x \mid x_1, \ldots, x_{j-1}; \gamma) P(y \mid x; \xi) \, \mathrm{d}x, \tag{38}$$

for $z_j \in \mathcal{Z}_j$ and $\lambda_j^{\text{up}}(z) = \lambda_j^{\text{up}}(x_1, \ldots, x_{j-1}, y)$. As an illustration, we consider a three-stage design. In the first step of the design, y is collected for the whole sample. In the second step, a

Table 1. *Specification of parameter values in example 7 that were used in Fig. 4*

	FHS	Iceland	KORA	Scandinavia
$\hat{\gamma}$	−2.04	−2.02	−2.12	−2.11
$\hat{\alpha}$	3.22	3.36	3.29	3.28
$\hat{\beta}$	0.0136	0.0237	0.0016	0.0111
$\hat{\sigma}$	0.169	0.232	0.165	0.137
Mean BMI	25.08	28.74	26.91	26.55

They were derived from Tables 1–2 in Lyon *et al.* (2007) as follows:
$F(\hat{\gamma}) = \exp(\hat{\gamma})/[1 + \exp(\hat{\gamma})] = n_{C/C}/(n_{C/C} + n_{C/G} + n_{G/G})$,
$\hat{\alpha} = \log[(\text{mean BMI}_{C/G} \times n_{C/G} + \text{mean BMI}_{G/G} \times n_{G/G})/(n_{C/G} + n_{G/G})]$,
$\hat{\beta} = \log(\text{mean BMI}_{C/C}) - \hat{\alpha}$,
$\hat{\sigma}^2 = (\text{SD BMI})^2 \times (1/\text{BMI mean})^2$.

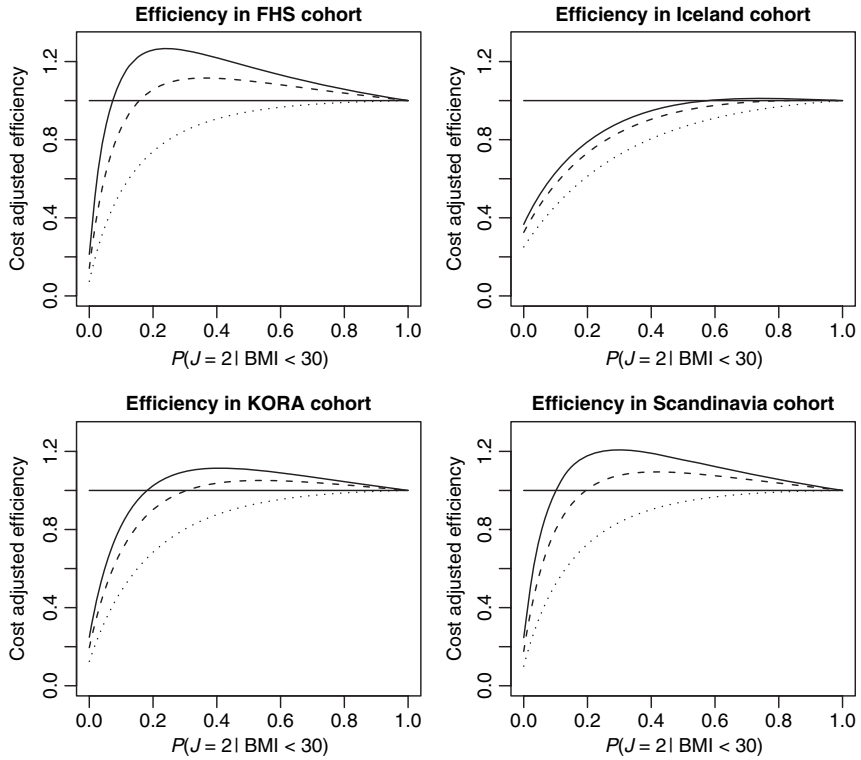

Fig. 4. Efficiency of over-sampling obese individuals (BMI ≥ 30) in different cohorts in Lyon *et al.* (2007) when estimating the effect (β) of the C/C genotype of genetic marker rs7566605 on log(BMI) in a linear regression model. Solid, dashed and dotted lines represent $C_1 = 0, 1/11$ and $1/3$, respectively, for $C_2 = 1$.

proportion of x_1 is selected, with selection probabilities determined by the value of y. A cut-off value t_Y is used, letting

$$\lambda_2^{\text{up}}(y) = \begin{cases} a; & y < t_Y, \\ 1; & y \geq t_Y, \end{cases}$$

and varying the value of a. In the third step, x_2 is collected, with selection probabilities determined by the values of y and x_1 simultaneously. For individuals with x_1, collected selection probabilities are

$$\lambda_3^{\text{up}}(y, x_1) = \begin{cases} b; & y < t_Y, x_1 < t_{x_1} \\ 1; & \text{otherwise}, \end{cases}$$

whereas for individuals with x_1 missing, x_2 is not eligible for selection. The design can thus be summarized by two parameters: $\eta = (a, b)$. We specify dependencies between variables in Fig. 5, where it is seen that X_2 acts as a confounder for the relation between X_1 and Y. In more detail, we have

$$X_2 \sim \text{Bin}\left[1, \frac{\exp(\alpha_{X_2})}{1 + \exp(\alpha_{X_2})}\right],$$

$$X_1 \mid \{X_2 = x_2\} \sim \text{Bin}\left[1, \frac{\exp(\alpha_{X_1} + \beta_{X_2 X_1} \times x_2)}{1 + \exp(\alpha_{X_1} + \beta_{X_2 X_1} \times x_2))}\right],$$

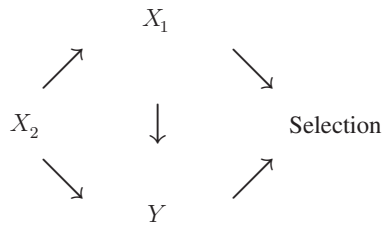

Fig. 5. A three-stage design.

$$Y \mid \{X_1 = x_1, X_2 = x_2\} \sim N(\alpha_Y + \beta_{X_1 Y} \times x_1 + \beta_{X_2 Y} \times x_2, \sigma_Y^2).$$

The likelihood contains seven parameters: $\gamma = (\alpha_{X_2}, \alpha_{X_1}, \beta_{X_2 X_1})$ from the covariate distribution and $\xi = (\alpha_Y, \beta_{X_1 Y}, \beta_{X_2 Y}, \sigma_Y)$ from the regression. To simplify, we focus on $\beta_{X_2 X_1}$, $\beta_{X_1 Y}$ and $\beta_{X_2 Y}$. The efficiency can be assessed for each parameter separately, or for all three simultaneously, using

$$h[I(\theta, \pi)] = \left[\sum_{r=3,5,6} \frac{V_{rr}(\theta, \pi)}{V_{rr}(\theta, \pi_{\text{full}})} \right]^{-1}, \tag{39}$$

a generalized form of the A-criterion in experimental design, which in our case depends on θ and puts equal interest into each of the three effect parameters. To visualize the results, three-dimensional plots are used in Fig. 6. With equal cost of sampling x_1 and x_2, there is no cost-efficiency gain in not sampling the full data set ($a = b = 1$). However, with no cost of sampling x_1, a considerable gain ($\text{CE} \geq 2$) is achieved by sampling optimally x_1 for a fraction $a \approx 0.2$ of individuals with non-extreme response variables. The optimum sampling fraction of x_2 is $b \approx 1$ for $\beta_{X_2 Y}$, $b \approx 0.2$ for $\beta_{X_2 X_1}$, $b \approx 0.3$ for $\beta_{X_1 Y}$ and $b \approx 0.5$ for all three parameters combined.

Example 9 (Models with surrogate data and latent variables). Suppose the covariates of a regression model are costly to sample but a cheaper surrogate variable S is available.

In the two-stage model,

$$z_1 = (s, y),$$
$$z_2 = (s, x, y),$$

the surrogate and response variable is collected at Stage 1 and covariate data at Stage 2 with a BUD inclusion probability $\lambda_2^{\text{up}}(s, y)$. The choice of likelihood depends on how Stage 1 data are collected. If it is sampled randomly from the population, we use the population distribution

$$f_2(z; \theta) = P(s; \gamma) P(x \mid s; \gamma) P(y \mid x; \xi)$$

of $Z = (S, X, Y)$, where $\theta = (\gamma, \xi)$ contains covariate and regression parameters, cf. Scott & Wild (1997). If Stage 1 data are sampled non-uniformly depending on Y, we use instead a retrospective likelihood with $f_2(z; \theta) = P(z \mid y; \theta)$, cf. Breslow & Holubkov (1997) and Scott & Wild (2007). For binary responses, this is a two-stage case–control design.

When the covariates are not known, but a proxy X' is available for sampling, the more general two-stage latent variable model

$$z_1 = (s, y),$$
$$z_2 = (s, x', y)$$

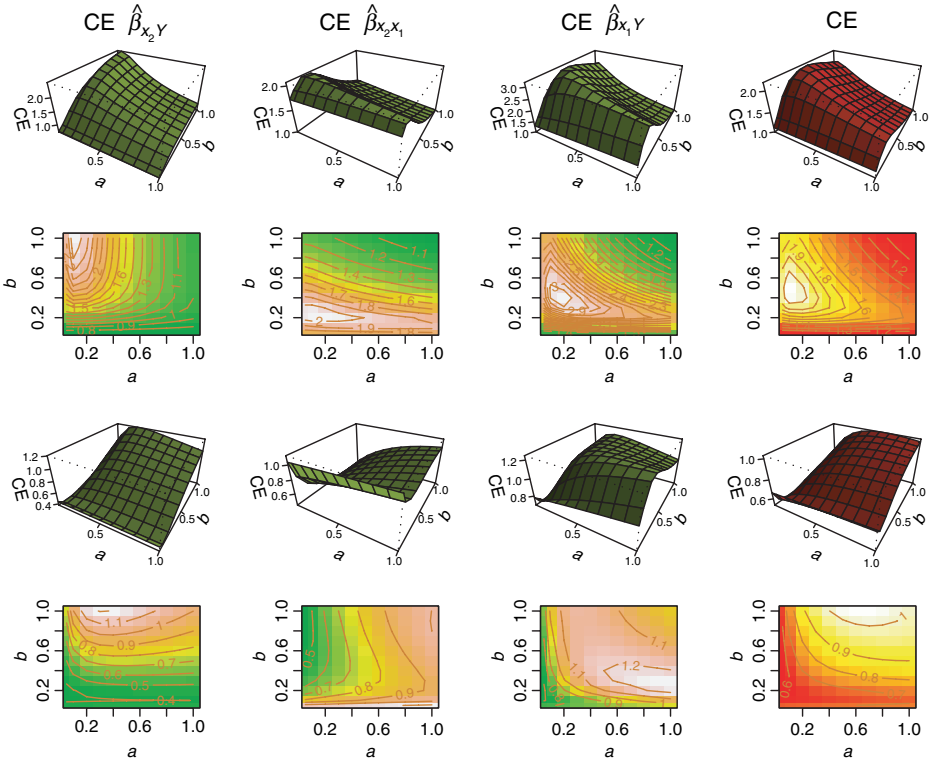

Fig. 6. CE for three parameters in a three-stage design, as described in example 8. Efficiency calculated for each parameter separately is presented in columns 1–3 (green), and efficiency calculated for all three parameters simultaneously is presented in column 4 (red). In the upper graphs, a surface is representing the cost efficiency. The same information is visualized in the graphs below, here instead projecting the height of the surface on a two-dimensional grid, letting a colour gradient represent the cost efficiency. Two cost functions are applied, $(C_1, C_2, C_3) = (0, 0, 1)$ (upper rows) and $(C_1, C_2, C_3) = (0, 1, 2)$ (lower rows). $t_Y = 3, t_{x_1} = 0.5$, $\alpha_{X_2} = 0, \alpha_{X_1} = -3, \beta_{X_2 X_1} = 2, \alpha_Y = 0, \beta_{X_1 Y} = 1, \beta_{X_2 Y} = 1, \sigma^2_Y = 1$. Monte Carlo approximation, based on (26), is used with $N = 10{,}000$.

can be used, cf. Thomas (2007). The BUD is still of the form $\lambda_2^{\mathrm{up}}(s, y)$, but the likelihood for randomly sampled data involves

$$f_2(z; \theta) = P(s; \gamma) \int P(x \mid s; \gamma) P(y \mid x; \xi) P(x' \mid x; \gamma) \, \mathrm{d}x,$$

which requires integration over the latent covariates. If Stage 1 data are sampled based on Y, a retrospective likelihood is used instead.

Example 10 (Stratum selection). Consider a regression model with $z = (x, y)$, for which the sample space is divided into $m_{\max}$ disjoint strata, $\mathcal{Z} = \cup_{m=1}^{m_{\max}} \mathcal{S}_m$. A two-stage BUD

$$z_1 = m,$$
$$z_2 = (x, y),$$

is defined by sampling stratum indicators m randomly from the population at Stage 1 and then full data for a proportion $\lambda_2^{\mathrm{up}}(m)$ of individuals within Stratum m. Breslow *et al.* (2003) consider ML estimation for this model when the covariate distribution parameter γ is infinite-dimensional.

Example 11 (Prevalence estimation with diagnostic tests). Let Y be a binary indicator of a given disease and suppose the prevalence $\xi = P(Y=1)$ is of main interest. For complex diseases, it may be costly to measure Y, and therefore a number of diagnostic tests $(\Phi_1, \ldots, \Phi_M)$ are available, with significance level $\alpha_m = P(\Phi_m = 1 \mid Y = 0)$ and power $\beta_m = P(\Phi_m = 1 \mid Y = 1)$. In the two-stage model

$$z_1 = (\phi_1, \ldots, \phi_M),$$
$$z_2 = (\phi_1, \ldots, \phi_M, y),$$

patients are sampled randomly into Stage 1, undergo diagnostic tests and so that divide the sample into 2^M strata. With a BUD, a proportion $\lambda_2^{\mathrm{up}}(\phi_1, \ldots, \phi_M)$ of Stratum $(\phi_1, \ldots, \phi_M)$ individuals proceed further to have disease status determined. If the diagnostic tests are independent,

$$f_2(z;\theta) = (1-\xi)^{\{y=0\}} \prod_{m=1}^{M} (1-\alpha_m)^{\{\phi_m=0\}} \alpha_m^{\{\phi_m=1\}} + \xi^{\{y=1\}} \prod_{m=1}^{M} (1-\beta_m)^{\{\phi_m=0\}} \beta_m^{\{\phi_m=1\}},$$

where $\theta = (\xi, \alpha_1, \beta_1, \ldots, \alpha_M, \beta_M)$. McNamee (2003) derives cost-efficient optimal designs for this model when $M = 1$. Salim & Welsh (2009) consider a related model when Y is replaced by a costly test Φ_{M+1} with higher precision than those of Stage 1 but yet not a perfect disease predictor.

Example 12 (Ascertainment versus two stage). The efficiency of conditional ascertainment $e_{\mathrm{condasc}}(\theta, \pi)$ is calculated for the logistic regression model (3), as presented in Fig. 7, together with the two-stage design efficiency, $e(\theta, \pi)$, for comparison. A two-stage design is preferable to an ascertainment design if $C_1 = 0$, since more information is available in the two-stage data set at no extra cost. In this example, the efficiency gain is, however, most prominent in estimating α, whereas no such effect is observed for the parameter of main interest, β.

More advanced examples of ascertainment occur in genetic applications with family data. For instance, Ginsburg *et al.* (2004) review the ascertainment problem in linkage analysis and another example is given here.

Example 13 (Ascertainment with three stages). Consider a population of families with three children, divided into $m_{\max}$ strata $\mathcal{Z} = \cup_{m=1}^{m_{\max}} \mathcal{S}_m$. We wish to estimate the prevalence of a certain rare disease whose occurrence varies between strata. Full data for one family is $z = \{m, (x_j, y_j)_{j=1}^3\}$, where m is the stratum indicator, x_j the covariate data and y_j the observed affection status of the jth child. A possible three-stage model is

$$z_1 = m,$$
$$z_2 = (m, y_1, y_2, y_3),$$
$$z_3 = (m, (x_j, y_j)_1^3).$$

Using logistic regression with stratum-specific intercept parameters,

$$f_3(z;\theta) = p_m \prod_{j=1}^{3} P(x_j \mid m; \gamma) F(\alpha_m + x_j \beta^{\mathrm{T}})^{\{y_j=1\}} [1 - F(\alpha_m + x_j \beta^{\mathrm{T}})]^{\{y_j=0\}},$$

where $\theta = (\gamma, p_1, \ldots, p_{m_{\max}}, \alpha_1, \ldots, \alpha_{m_{\max}}, \beta)$, γ contains covariate distribution parameters and $p_m = P(M = m)$ are the strata proportions. The parameter of main interest is the covariate-adjusted overall logodds prevalence $\alpha = \sum_m p_m \alpha_m$. A possible subclass $\mathcal{Q}$ of BUDs is obtained by collecting affection status with different proportions over strata and then covariates

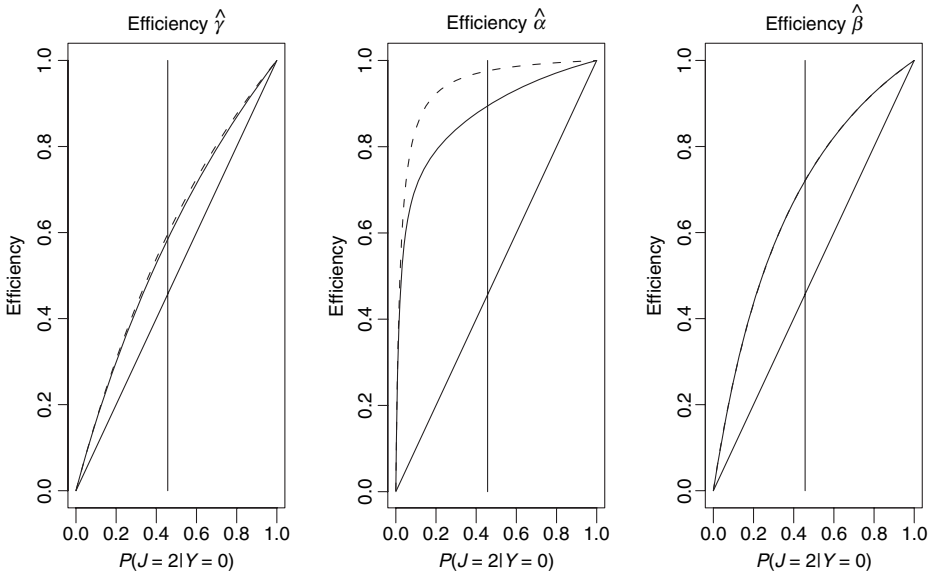

Fig. 7. Efficiency for estimation of parameters using two-stage (dashed line) likelihood and conditional ascertainment likelihood (solid line) in logistic regression, cf. examples 6 and 12, with $\gamma=-1, \alpha=-2, \beta=2, \eta_1=P(J=2 \mid Y=1)=1$ and $0<\eta_0=P(J=2 \mid Y=0)\leq 1$.

for families with at least one affected child, that is, $\lambda_2^{\text{up}}(z)=\eta_m$, $\lambda_3^{\text{up}}(z)=1_{\{y_1+y_2+y_3\geq 1\}}$ and $\eta=(\eta_1,\ldots,\eta_{m_{\max}})$.

Now, suppose we have a (conditionally) ascertained data set, so that only families with covariate data are retained. This is very similar to the model treated by Burton *et al.* (2001). It is shown in this article that lost stratum indicators in the ascertained sample may result in severely biased estimates of the overall prevalence when the whole population is treated as one stratum, thereby mis-specifying the ascertainment procedure.

Example 14 (Distributed detection). Consider a wireless network of n sensors distributed over a certain region. Their task is to locate a possible target. Sensor i registers data $Z^i=(X_1^i,\ldots,X_T^i)$ during T time points, where $X_t^i\sim N(0,\sigma^2)$ are independent. Detection of the target is performed by a fusion centre (FC), the task of which is to test a null hypothesis of no target ($\Theta_0=\{\sigma_0^2\}$) against the alternative of a present target ($\Theta\backslash\Theta_0=(\sigma_0^2,\infty)$). Sensor i compresses data in the form of average power $Y^i=\sum_{t=1}^{\mathrm{T}}(X_t^i)^2/T\sim\sigma^2\chi^2(T)/T$ and transmits it to the FC. To save energy and communication resources, this is only done when $Y^i\geq\lambda$ for some threshold λ. A sent message from Sensor i may be lost with probability p. We formulate this as a TDD with three stages

$$z_3=(x_1,\ldots,x_T),$$
$$z_2=y=\sum_t x_t^2/T,$$
$$z_1=\emptyset.$$

Since messages are always compressed, $\lambda_2^{\text{down}}(x_1,\ldots,x_T)\equiv 1$. A compressed message is sent and reaches the FC with probability $\lambda_1^{\text{down}}(y)=p1_{\{y\geq\lambda\}}$. Thus, either $J^i=2$ if the FC receives a signal from Sensor i or $J^i=1$ if it does not, in which case the signal is either not sent or sent and lost. Since the send condition depends on y, this is not a MARD. The design is para-

meterized by $\eta=(\lambda,p)$. Suppose the cost for each sensor of compressing data is a and the cost of sending y is $b(p)$, a decreasing function of p, since a higher sending power is required to achieve a small p. This gives cost functions $C_1(y)=C_2(y)=C_3(y)=a+b(p)1_{\{y\geq\lambda\}}$, of which C_1 is not stage-dependent, and $\mathrm{TAC}=n[a+b(p)P(Y\geq\lambda)]$. For more details on distributed detection, see Chamberland & Veeravalli (2007) and references therein.

10. Discussion

Even though multistage designs are often used in observational studies, it is usually not transparent in the design phase how the data selection will affect the efficiency of the study. This article provides a framework describing the design procedure, in the attempt to facilitate description and discussion of such. We also describe how efficiency and cost-adjusted efficiency can be calculated using Fisher information matrices adjusted for the selection procedure.

The relative performance of different selection schemes π in the examples varied with costs and assumed parameters θ, which illustrates that it is advisable to calculate efficiency for several parameter values, and choose a design that has acceptable efficiency for most plausible values of θ.

One obtains a broader picture by presenting efficiencies for a whole class $\mathcal{Q}$ of designs in graphs, rather than just calculating one optimal sampling scheme. In this setting, the area under the optimal efficiency curve (25) may be used to quantify an average cost efficiency of the designs in $\mathcal{Q}$ and compare it with $(1+R_{\min})/2$, the corresponding area for SRSs.

Alternatively, an iterative approach is often successfully used in experimental design, that is, to first do a small study and then fill in more data where the initial sample suggests is efficient (Montgomery, 1984). Similarly, pilot studies are sometimes used in observational studies to identify practical issues in data collection, but can also be used for assessment of crude parameter estimates for subsequent cost-efficiency calculations.

The choice of design may also affect the plausibility of the model assumptions, such as normally distributed errors in regression. For example, Allison *et al.* (1988) argue that selecting only extreme outcomes is not advisable in genetic studies, since extremes are likely to result from rare exposures with strong effects not in the model.

We have assumed in (4) that only θ is estimated, whereas π is known. This is natural if the sampling scheme is controlled by the investigator. If π is rather an unknown nuisance parameter, in general, it has to be estimated either from training data or from the data set at hand by maximizing $L(\theta,\pi)$ jointly w.r.t. θ and π. However, for MARDs, it is not necessary to estimate π, since the likelihood can be factorized as:

$$L(\theta,\pi)=A(\theta)B(\pi), \tag{40}$$

so that inference on θ can be based solely on $A(\theta)$ if the joint parameter space of θ and π is the product of the individual parameter spaces, cf. Rubin (1976), Heitjan & Rubin (1991) and Little & Rubin (2002).

The methodology described here is equally valid for other estimators than ML. Let $\theta(\pi)$ be a given estimator of θ using design π, with asymptotic covariance matrix $V(\hat{\theta};\theta,\pi)$. Then, the efficiency of estimating the rth component of θ, compared with the ML estimator $\hat{\theta}_{\mathrm{ML}}$ of the full data set, is

$$e(\hat{\theta};\theta,\pi)=V_{rr}(\hat{\theta}_{\mathrm{ML}};\theta,\pi_{\mathrm{full}})/V_{rr}(\hat{\theta};\theta,\pi). \tag{41}$$

This corresponds to our previous definition (13) when $\hat{\theta}=\theta_{\mathrm{ML}}$ and $h(I)=V_{rr}^{-1}$. When $\pi=\pi_{\mathrm{full}}$, (41) is the usual definition of efficiency (see, e.g. Lehmann & Casella, 1998). Similarly, $e(T;\theta,\pi)$ for a test statistic T, with design π, is defined by comparing its non-centrality parameter with that of T_{LR} for the full data set. There may be several reasons for choosing other esti-

mators and tests than ML and LR, such as (i) robustness against model mis-specification (e.g. generalized estimating equations; Liang & Zeger, 1986), (ii) simple and explicit algorithms (e.g. inverse probability weighted ML estimates; Breslow & Wellner, 2007), (iii) approximations of likelihood methods with explicit optimal cost-efficient designs (e.g. mean score method estimates; cf. Pepe *et al.*, 1994; Reilly & Pepe, 1995; Reilly, 1996). Also, we have focused on asymptotic expressions of the covariance matrix based on the Fisher information matrix. Another possibility is to use distributional properties of finite sample estimators of the covariance matrix, for example, based on a sandwich estimator (see, e.g. Kauermann & Carroll, 2001).

Acknowledgements

Maria Grünewald was supported by the Swedish Foundation for Strategic Research, grant A3 02:129. Ola Hössjer was supported by the Swedish Research Council, grants 621-2005-2810 and 621-2008-4946, and the Gustafsson Foundation for Research in Natural Sciences and Medicine. The authors thank one Associate Editor, Duncan Thomas, Sandrah Eckel and two other referees, whose comments considerably improved the quality of the article.

Supporting Information

Additional supporting material may be found in the online version of this article:

Appendix S1: Mathematical derivations and proofs.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Allison, D., Heo, M., Schork, N., Wong, S. & Elston, R. (1998). Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. *Human Hered.* **48**, 97–107.

Breslow, N. E. & Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *J. Roy. Statist. Soc.* **59**, 447–461.

Breslow, N. E., McNeney, B. & Welnner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *Ann. Statist.*, **31**, 1110–1139.

Breslow, N. E. & Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scand. J. Statist.* **34**, 86–102.

Burton, P. R., Palmer, L. J., Jacobs, K., Keen, K. J., Olson, J. M. & Elston, R. C. (2001). Ascertainment adjustment: Where does it take us? *Amer. J. Hum. Genet.* **67**, 1505–1514.

Chamberland, J.-F. & Veeravalli, V. V. (2007). Wireless sensors in distributed detection applications. *IEEE Signal Processing Magazine*, 16–25, May.

Fisher, R. (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Ann. Eugenics.* **6**, 13–25.

Ginsburg, E., Malkin, I. & Elston, R. C. (2004). Sampling correction in linkage analysis. *Genet. Epidemiol.* **27**, 87–96.

Grünewald, M. & Hössjer, O. (2010a). Efficient ascertainment schemes for maximum likelihood estimation. *J. Statist. Plann. Inference.* **140**, 2078–2088.

Grünewald, M. & Hössjer, O. (2010b). A general statistical framework for multistage designs. Rep. 2010:6, Mathematical Statistics, Stockholm University.

Grünewald, M., Humphreys, K. & Hössjer, O. (2010). A stochastic EM type algorithm for parameter estimation in models with continuous outcomes, under complex ascertainment. *Int. J. Biostatist.* **6**, Article 23.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo methods.* Methuen, London.

Heitjan, D. & Rubin, D. (1991). Ignorability and coarse data. *Ann. Statist.* **19**, 2244–2253.
Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics* **37**, 185–194.
Kauermann, G. & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.* **96**, 1387–1396.
Lehmann, E. & Casella, G. (1998). *Theory of point estimation*, 2nd edn. Springer, New York.
Liang, K.-Y. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.
Little, R. J. A. & Rubin, D. (2002). *Statistical analysis with missing data*. Wiley, New York.
Lyon, H. N., Emilsson, V., Hinney, A., Heid, I. M., Lasky-Su, J., Zhu, X., Thorleifsson, G., Gunnarsdottir, S., Walters, G. B., Thorsteinsdottir, U., *et al.* (2007). The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. *PLoS Genet*. **3**, e61.
Maydrech, E. & Kupper, L. (1978). Cost considerations and sample size requirements in cohort and case-control studies. *Amer. J. Epidemiol.* **107**, 201–205.
McNamee, R. (2003). Efficiency of two-phase designs for prevalence estimation. *Int. J. Epidemiol.* **32**, 1072–1078.
Melas, V. (2006). *Functional approach to optimal experimental design*, number 184 *in* 'Lecture notes in statistics'. Springer, USA (Chapter 1.4).
Montgomery, D. C. (1984). *Design and analysis of experiments*, 2nd edn., Chapter 1. John Wiley & Sons, New York.
Pepe, M., Reilly, M. & Fleming, T. (1994). Auxiliary outcome data and the mean score method. *J. Statist. Plann. Inference* **42**, 137–160.
Prentice, R. L. & Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–158.
Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case control studies. *Biometrika* **66**, 403–411.
R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for statistical computing, Vienna, Austria.
Rao, C. (1965). *Classical and contagious discrete distributions*. Pergamon Press and Statistical Publishing Society, Calcutta, 320–332.
Reilly, M. (1996). Optimal sampling strategies for two-stage studies. *Amer. J. Epidemiol.* **143**, 92–100.
Reilly, M. & Pepe, M. S. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299–314.
Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
Salim, A. & Welsh, A. H. (2009). Designing 2-phase prevalence studies in the absence of a 'gold standard' test. *Amer. J. Epidemiol* **170**, 369–378.
Scott, A. J. & Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71.
Scott, A. J. & Wild, C. J. (2007). On the Breslow-Holubkov estimator. *Lifetime Data Anal*. **13**, 545–563.
Serfling, R. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.
Silvey, S. (1980). *Optimal design*. Chapman and Hall, London.
Steingrímsdóttir, L., Thorgeirsdóttir, H. & Olafdóttir, S. (2003). The diet of Icelanders: Dietary survey of the icelandic nutrition council 2002. *Rannsóknir Manneldisráds Íslands V, 2002*.
Thomas, D. C. (2007). Multistage sampling for latent variable models. *Lifetime Data Anal*. **13**, 565–581.
Thomas, D. C., Xie, R. & Gebregziabher, M. (2004). Two-stage sampling designs for gene association studies. *Genet. Epidemiol.* **27**, 401–414.
White, J. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Amer. J. Epidemiol.* **115**, 119–128.
Zhou, H., Chen, J., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T. & Longnecker, M. P. (2007). Outcome dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology*, **18**, 461–468.

Received May 2010, in final form March 2011

Ola Hössjer, Department of Mathematics, Stockholm University, SE. 10691, Stockholm, Sweden.
E-mail: ola@math.su.se