Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 53 (2009) 1782–1793
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates

Juha Karvanen[a,*], Sangita Kulathinal[b], Dario Gasbarra[c]

[a] *Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Mannerheimintie 166, 00300 Helsinki, Finland*
[b] *Indic Society for Education and Development, Nashik, India*
[c] *University of Helsinki, P.O. Box 68 (Gustaf Hällströmin katu 2b), Helsinki FIN-00014, Finland*

Available online 16 February 2008

Abstract

In gene-disease association studies, the cost of genotyping makes it economical to use a two-stage design where only a subset of the cohort is genotyped. At the first-stage, the follow-up data along with some risk factors or non-genetic covariates are collected for the cohort and a subset of the cohort is then selected for genotyping at the second-stage. Intuitively the selection of the subset for the second-stage could be carried out efficiently if the data collected at the first-stage are utilized. The information contained in the conditional probability of the genotype given the first-stage data and the initial estimates of the parameters of interest is being maximized for efficient selection of the subset. The proposed selection method is illustrated using the logistic regression and Cox's proportional hazards model and algorithms that can find optimal or nearly optimal designs in discrete design space are presented. Simulation comparisons between D-optimal design, extreme selection and case-cohort design suggest that D-optimal design is the most efficient in terms of variance of estimated parameters, but extreme selection may be a good alternative for practical study design.

1. Introduction

Multi-stage designs are commonly used in epidemiology especially when the collection of data on some covariates is expensive. In a typical gene-disease association study, baseline and event follow-up data are collected for the study cohort at the first-stage and then a subset of the cohort is selected for genotyping at the second-stage of the study. The well-known examples include case-cohort (Prentice, 1986), case-control and nested case-control designs (Clayton and Hills, 1993). An alternative approach considered in this paper is to select the subgroup for genotyping so that the information on the parameters of interest is maximized in the conditional distribution of the genotype given the first-stage data. A smaller sized subgroup may then provide better information on the gene-disease association.

Let us assume that at the first-stage we have already measured the covariates $\mathbf{x}(i)$ and the response $\mathbf{y}(i)$ for the whole cohort $i \in C = \{1, 2, \ldots, N\}$. At the second-stage, the genetic covariate of interest $g(i)$ is either not measured

* Corresponding author. Tel.: +358 9 4744 8641; fax: +358 9 4744 8338.
E-mail addresses: juha.karvanen@ktl.fi (J. Karvanen), sangita.kulathinal@inseed.org (S. Kulathinal), dag@rni.helsinki.fi (D. Gasbarra).

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.010

yet for anybody or is measured only for a (small) subset of the cohort. Our goal is to choose $n \ll N$ individuals for genotyping at the second-stage in a way that is optimal according to some well defined optimality criterion. The classical criteria of optimality are based on the expected Fisher information matrix

$$I_{\mathbf{X},\mathbf{Y},G}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left(\frac{\partial^2 \log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{X},\mathbf{Y},G)}{\partial \boldsymbol{\theta}^2}\right), \tag{1}$$

where $\boldsymbol{\theta}$ is the vector of the parameters of interest, $\boldsymbol{\psi}$ is the vector of nuisance parameters and the observations $\mathbf{x}(i)$, $\mathbf{y}(i)$ and $g(i)$ are understood to be realizations of the random variables $\mathbf{X}$, $\mathbf{Y}$ and G, respectively. We choose to use the D-optimality, which maximizes the determinant of the information matrix. Using the property $p(\mathbf{X},\mathbf{Y},G \mid \boldsymbol{\theta},\boldsymbol{\psi}) = p(\mathbf{X},\mathbf{Y} \mid \boldsymbol{\theta},\boldsymbol{\psi})\, p(G \mid \mathbf{X},\mathbf{Y},\boldsymbol{\theta},\boldsymbol{\psi})$ in (1) we obtain

$$\begin{aligned} I_{\mathbf{X},\mathbf{Y},G}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}\left(-\frac{\partial^2 \log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(\mathbf{X},\mathbf{Y})}{\partial \boldsymbol{\theta}^2}\right) + E_{\boldsymbol{\theta}}\left(E_{\boldsymbol{\theta}}\left(-\frac{\partial^2 \log p_{\boldsymbol{\theta},\boldsymbol{\psi}}(G \mid \mathbf{X},\mathbf{Y})}{\partial \boldsymbol{\theta}^2}\,\middle|\,\mathbf{X},\mathbf{Y}\right)\right) \\ &= I_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta}) + E_{\boldsymbol{\theta}}\left(I_{G|\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta})\right), \end{aligned} \tag{2}$$

where the inner expectation in the second term is with respect to the conditional distribution of G given $(\mathbf{X},\mathbf{Y})$ and the outer expectation is with respect to the joint distribution of $(\mathbf{X},\mathbf{Y})$. The first term is fixed in the sense that the data on $(\mathbf{X},\mathbf{Y})$ are already collected. The overall information about $\boldsymbol{\theta}$ can still be maximized by collecting $g(i)$ for given $(\mathbf{x}(i),\mathbf{y}(i))$ such that $I_{G|\mathbf{x}(i),\mathbf{y}(i)}(\boldsymbol{\theta})$ is maximum.

The general theory of optimal design (Atkinson and Donev, 1992; Pukelsheim, 1993) is frequently applied to dose-response experiments, see e.g. Myers et al. (1996), but less often in other fields of biometry. Optimal sample size allocation between the stages in epidemiological two-stage studies is considered by McNamee (2002) and Reilly (1996). Recently, Wright and Bailer (2006) applied D-optimal design to start-stop experiments in environmental toxicology. Their approach is similar to ours but the models considered are different. In genetics, extreme selection (Lander and Botstein, 1989; Carey and Williamson, 1991; Darvasi and Soller, 1992; Allison et al., 1998; VanGestel et al., 2000; Tenesa et al., 2005; McElroya et al., 2006; Macgregor et al., 2006) is a well-known selection strategy where individuals with the highest and lowest phenotype values are selected for genotyping. Extreme selection may be motivated by the fact that in linear regression the optimal design is comprised of extreme covariate values (Elfving, 1952) but the design can be also applied as an ad-hoc design when the model is non-linear. In nested case-control studies, a somewhat similar concept is counter matching (Langholz and Borgan, 1995; Langholz, 2007) where the controls are selected in such a way that they differ maximally from the cases. It is to be noted that the above-mentioned genetics literature address an issue of selecting individuals already phenotyped for a quantitative trait where this quantitative trait is of the main interest. Further, the tails of the distribution of the main trait are then used for selection of individuals to enable more efficient mapping of loci. This is slightly different from analyzing a binary or survival trait as the main outcome for given non-genetic covariates and then selecting individuals based on the distributions of the covariates among cases and controls.

In this article, we develop methods for optimally selecting individuals for collecting data on the genetic covariate g. In Section 2, Fisher information matrices are derived for situations where the response is modeled by logistic regression or Cox's proportional hazards model. In Section 3, we present algorithms that can find optimal or nearly optimal designs in discrete design space. Statistical analysis of data collected according to the D-optimal design is considered in Section 4. Simulation studies comparing D-optimal design, extreme selection, case-cohort design and simple random sampling are presented in Section 5. Section 6 concludes the paper.

2. D-optimal designs for nonlinear models

2.1. General model

The likelihood of the models we consider can be written in general form as $L(\boldsymbol{\theta};\mathbf{x},\mathbf{y},g)$, where the likelihood is a function of the parameters of interest $\boldsymbol{\theta}$ but not the nuisance parameters $\boldsymbol{\psi}$. The interpretation of the likelihood depends on the stage that is considered: at the first-stage, we have observed the covariates $\mathbf{x}$ and the response $\mathbf{y}$ but the genetic covariate g is unknown; at the second-stage, g is available for some individuals but missing for the majority of the

cohort. Typically, the genetic covariate g may be a genotype variable with three possible values or a binary variable indicating presence or absence of certain allele. Our objective is to find the explicit form of information matrices $I_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta})$ and $E_{\boldsymbol{\theta}}\left(I_{G|\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta})\right)$ given in (2). We denote

$$L_g = p_{\boldsymbol{\theta}}(G = g, \mathbf{X}, \mathbf{Y}), \tag{3}$$

where the notation $p_{\boldsymbol{\theta}}$ includes only the parameters of interest and write the conditional probability of G given $(\mathbf{x}, \mathbf{y})$ as

$$p_{\boldsymbol{\theta}}(G = g \mid \mathbf{x}, \mathbf{y}) = \frac{L_g}{\sum_k L_k}, \tag{4}$$

where the summation goes over all possible values of g. Making the simplifying but realistic assumption that g has possible values 0 and 1 (i.e. dominant allele effect) we obtain for the marginal distribution of $(\mathbf{X}, \mathbf{Y})$

$$p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y}) = \sum_g p_{\boldsymbol{\theta}}(G = g, \mathbf{X}, \mathbf{Y}) = L_1 + L_0, \tag{5}$$

where L_0 and L_1 are defined as in (3). Using this result, we may write the elements of the information matrix $I_{\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta})$ as

$$\begin{aligned} &E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_i}\log(L_1 + L_0)\,\frac{\partial}{\partial\theta_j}\log(L_1 + L_0)\right) \\ &\quad = E_{\boldsymbol{\theta}}\left(\frac{1}{(L_1 + L_0)^2}\left(L_1\frac{\partial \log L_1}{\partial\theta_i} + L_0\frac{\partial \log L_0}{\partial\theta_i}\right)\left(L_1\frac{\partial \log L_1}{\partial\theta_j} + L_0\frac{\partial \log L_0}{\partial\theta_j}\right)\right), \end{aligned} \tag{6}$$

where θ_i and θ_j are some model parameters.

For the conditional log-likelihood we obtain the following form

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(G = g \mid \mathbf{x}, \mathbf{y}) &= g\log\left(\frac{L_1}{L_1 + L_0}\right) + (1 - g)\log\left(\frac{L_0}{L_1 + L_0}\right) \\ &= g\log L_1 + (1 - g)\log L_0 - \log(L_1 + L_0). \end{aligned} \tag{7}$$

The elements of the information matrix $E_{\boldsymbol{\theta}}\left(I_{G|\mathbf{X},\mathbf{Y}}(\boldsymbol{\theta})\right)$ are obtained from the products of the partial derivatives of (7) by taking the expectation with respect to the conditional distribution of G given $(\mathbf{x}, \mathbf{y})$

$$\begin{aligned} &E_{\boldsymbol{\theta}}\left(E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_i}\log p_{\boldsymbol{\theta}}(G|\mathbf{X}, \mathbf{Y})\;\frac{\partial}{\partial\theta_j}\log p_{\boldsymbol{\theta}}(G|\mathbf{X}, \mathbf{Y})\,\middle|\,\mathbf{X}, \mathbf{Y}\right)\right) \\ &\quad = E_{\boldsymbol{\theta}}\left(\sum_{g=0,1} p_{\boldsymbol{\theta}}(G = g|\mathbf{X}, \mathbf{Y})\frac{\partial}{\partial\theta_i}\log p_{\boldsymbol{\theta}}(G = g|\mathbf{X}, \mathbf{Y})\frac{\partial}{\partial\theta_j}\log p_{\boldsymbol{\theta}}(G = g|\mathbf{X}, \mathbf{Y})\right) \\ &\quad = E_{\boldsymbol{\theta}}\left(\frac{L_0}{L_0 + L_1}\frac{\partial}{\partial\theta_i}\log\left(\frac{L_0}{L_0 + L_1}\right)\frac{\partial}{\partial\theta_j}\log\left(\frac{L_0}{L_0 + L_1}\right)\right. \\ &\qquad \left. + \frac{L_1}{L_0 + L_1}\frac{\partial}{\partial\theta_i}\log\left(\frac{L_1}{L_0 + L_1}\right)\frac{\partial}{\partial\theta_j}\log\left(\frac{L_1}{L_0 + L_1}\right)\right) \\ &\quad = E_{\boldsymbol{\theta}}\left(\frac{L_1 L_0}{(L_1 + L_0)^2}\left(\frac{\partial \log L_0}{\partial\theta_i} - \frac{\partial \log L_1}{\partial\theta_i}\right)\left(\frac{\partial \log L_0}{\partial\theta_j} - \frac{\partial \log L_1}{\partial\theta_j}\right)\right). \end{aligned} \tag{8}$$

The two terms of Eq. (2), which are further derived in equations (6) and (8), require taking expectation over the joint distribution of $(\mathbf{X}, \mathbf{Y})$. This expectation may be difficult to calculate analytically but fortunately we may replace the expected information by observed information $J_{\mathbf{x},\mathbf{y},G}(\boldsymbol{\theta})$, whose elements can be calculated as a sum over the observations

$$\sum_{k=1}^{N}\left(\frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Y})}{\partial\theta_i}\,\frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Y})}{\partial\theta_j}\bigg|\,\mathbf{X}=\mathbf{x}(k),\mathbf{Y}=\mathbf{y}(k)\right)$$
$$+\sum_{k=1}^{n}\left(E_{\boldsymbol{\theta}}\left(\frac{\partial \log p_{\boldsymbol{\theta}}(G\mid\mathbf{X},\mathbf{Y})}{\partial\theta_i}\,\frac{\partial \log p_{\boldsymbol{\theta}}(G\mid\mathbf{X},\mathbf{Y})}{\partial\theta_j}\bigg|\,\mathbf{X}=\mathbf{x}(k),\mathbf{Y}=\mathbf{y}(k)\right)\right). \tag{9}$$

In the equation above, the first term is a sum over all first-stage observations $k = 1, 2, \ldots, N$ whereas the second term is a sum over the second-stage observations $k = 1, 2, \ldots, n$ only. In other words, the values of $J_{\mathbf{x},\mathbf{y},G}(\boldsymbol{\theta})$ can be calculated using equations (6) and (8) where the expectations are replaced by summations. The D-criterion to be maximized is defined as

$$D = \det(J_{\mathbf{x},\mathbf{y},G}(\boldsymbol{\theta})). \tag{10}$$

It is often practical to handle the standardized version of the D-criterion

$$D^* = \sqrt[m]{\det(J_{\mathbf{x},\mathbf{y},G}(\boldsymbol{\theta}))}, \tag{11}$$

where m is the number of parameters in $\boldsymbol{\theta}$.

We consider two models, logistic regression and proportional hazards models, which are of interest in genetic epidemiology for assessing gene-disease association. For simplicity, we consider only a single non-genetic covariate x. The possible dependence between g and x is modeled assuming that the conditional distribution of X given $G = g$ is normal with mean $\mu + \gamma g$ and variance σ^2 but the corresponding parameters μ, σ^2 and γ are considered as to be part of the nuisance parameters $\boldsymbol{\psi}$. The D-optimal design is derived only with respect to the parameters of interest $\boldsymbol{\theta}$.

2.2. Logistic regression

Let the response Y to be a dichotomous variable that follows the logistic regression model

$$p(Y = 1 \mid a, b, c, x, g) = F(ax + bg + c), \tag{12}$$

where

$$F(z) = \frac{\exp(z)}{1+\exp(z)}, \tag{13}$$

which satisfies

$$\frac{\mathrm{d}}{\mathrm{d}z}\log F(z) = 1 - F(z) \quad \text{and} \tag{14}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}\log(1 - F(z)) = -F(z). \tag{15}$$

The likelihood has form

$$L(a, b, c, \pi; x, y, g) = \pi^g(1-\pi)^{1-g}\phi((x-\mu-\gamma g)/\sigma)F(ax+bg+c)^y(1-F(ax+bg+c))^{1-y}, \tag{16}$$

where $P(G = 1) = \pi$ and $\phi()$ is the probability density function of the standard normal distribution.

The model parameters of interest are $\boldsymbol{\theta} = (a, b, c, \pi)$ and the nuisance parameters are $\boldsymbol{\psi} = (\mu, \sigma^2, \gamma)$. In order to apply the results (6) and (8) we calculate the partial derivatives (score functions) of

$$\log L_1 = \log\pi + \log\phi((x-\mu-\gamma)/\sigma) + y\log F(ax+b+c) + (1-y)\log(1-F(ax+b+c)) \quad \text{and} \tag{17}$$

$$\log L_0 = \log(1-\pi) + \log\phi((x-\mu)/\sigma) + y\log F(ax+c) + (1-y)\log(1-F(ax+c)). \tag{18}$$

Noting that

$$\frac{\mathrm{d}}{\mathrm{d}a}\left(y\log F(z(a)) + (1-y)\log(1-F(z(a)))\right) = (y - F(z))\frac{\mathrm{d}z(a)}{\mathrm{d}a}, \tag{19}$$

we obtain

$$\frac{\partial \log L_0}{\partial a} = (y - F(ax + c))x, \tag{20}$$

$$\frac{\partial \log L_0}{\partial b} = 0 \tag{21}$$

$$\frac{\partial \log L_0}{\partial c} = (y - F(ax + c)), \tag{22}$$

$$\frac{\partial \log L_0}{\partial \pi} = -\frac{1}{1-\pi}, \tag{23}$$

$$\frac{\partial \log L_1}{\partial a} = (y - F(ax + b + c))x, \tag{24}$$

$$\frac{\partial \log L_1}{\partial b} = \frac{\partial \log L_1}{\partial c} = (y - F(ax + b + c)) \quad \text{and} \tag{25}$$

$$\frac{\partial \log L_1}{\partial \pi} = \frac{1}{\pi}. \tag{26}$$

Note that the derivatives above depend on the parameters of interest $\boldsymbol{\theta}$ but not on the nuisance parameters $\boldsymbol{\psi}$. The D-optimal design, however, depends also on $\boldsymbol{\psi}$ through the terms L_0 and L_1.

2.3. Proportional hazards model

Let t be the time of the disease event and δ the status indicator that gives $\delta = 1$ for the event and $\delta = 0$ for censoring. The likelihood has form

$$L(a, b, \pi; x, t, \delta, g) = \pi^g (1-\pi)^{1-g} \phi((x - \mu - \gamma g)/\sigma)(1 - F(t; x, g, a, b))^{1-\delta} f(t; x, g, a, b)^{\delta}, \tag{27}$$

where $P(G = 1) = \pi$ and $\phi()$ is the probability density function of the standard normal distribution. Under Cox's proportional hazards model we have

$$F(t; x, g, a, b) = 1 - (1 - F_0(t))^{\exp(ax+bg)} \quad \text{and} \tag{28}$$

$$\begin{aligned} f(t; x, g, a, b) &= f_0(t) \exp(ax + bg)(1 - F_0(t))^{\exp(ax+bg)-1} \\ &= \lambda_0(t) \exp(ax + bg)(1 - F_0(t))^{\exp(ax+bg)}, \end{aligned} \tag{29}$$

where $F_0(t)$ is a cumulative distribution function (cdf) and $\lambda_0(t) = f_0(t)/(1 - F_0(t))$ is the corresponding hazard rate. The baseline hazard $\lambda_0(t)$ is considered as a nuisance parameter.

The model parameters of interest are $\boldsymbol{\theta} = (a, b, \pi)$ and the nuisance parameters are $\boldsymbol{\psi} = (\mu, \sigma^2, \gamma, \lambda_0(t))$. In order to apply the results (6) and (8) we calculate the partial derivatives

$$\frac{\partial \log L_0}{\partial a} = h(ax)x, \tag{30}$$

$$\frac{\partial \log L_0}{\partial b} = 0, \tag{31}$$

$$\frac{\partial \log L_0}{\partial \pi} = -\frac{1}{1-\pi}, \tag{32}$$

$$\frac{\partial \log L_1}{\partial a} = h(ax + b)x, \tag{33}$$

$$\frac{\partial \log L_1}{\partial b} = h(ax + b) \quad \text{and} \tag{34}$$

$$\frac{\partial \log L_1}{\partial \pi} = \frac{1}{\pi}, \tag{35}$$

where

$$h(z) = \delta + \exp(z)\log(1 - F_0(t)). \tag{36}$$

3. Finding optimal designs in discrete design space

In order to find the optimal subset of individuals for the second-stage, we need initial estimates of the model parameters and a computational method that maximizes the D-criterion at these estimates when the sample size of the second-stage is specified. The initial estimates can be obtained from literature or from a previous study. If these are not available a sophisticated guess may be used. It is also possible to use case-control or case-cohort design first and apply D-optimal design when the follow-up has been extended and more individuals are needed for genotyping.

In general, it is not computationally possible to explore all possible subsets of individuals and therefore heuristic search methods are needed. The search strategies applied in this paper are the greedy method and the iterative replacement method. The sample size of the second-stage is usually determined by the genotyping budget. In the greedy method (Dykstra, 1971), the individuals are selected sequentially one by one so that the D-criterion D_n for n individuals is maximized on the condition that $n-1$ individuals have been already selected. Let S be the set of the $n-1$ individuals already selected and let $J_{\mathbf{x}(j),\mathbf{y}(j),G}(\hat{\boldsymbol{\theta}})$ be the observed Fisher information for individual j calculated on the condition $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ represents the initial parameter estimates. The new individual $j \notin S$ is selected so that

$$D_n = \det\left(\sum_{i\in S} J_{\mathbf{x}(i),\mathbf{y}(i),G}(\hat{\boldsymbol{\theta}}) + J_{\mathbf{x}(j),\mathbf{y}(j),G}(\hat{\boldsymbol{\theta}})\right) \tag{37}$$

is maximized. Note that this is different from

$$\max_{j\notin S} \det(J_{\mathbf{x}(j),\mathbf{y}(j),G}(\hat{\boldsymbol{\theta}})) \tag{38}$$

because of the nonlinearity of the determinant. If there is more than one individual that maximizes (37), the choice between them may be done randomly. The individual j is added to set S and the selection continues sequentially so that on the next round D_{n+1} is maximized on the condition that n individuals have been already selected.

In the iterative replacement method, which is also known as modified Fedorov method (Cook and Nachtsheim, 1980), the search starts with an initial selection which is often obtained by the greedy method. Then the selected individuals are considered one by one and replaced by another individual if that increases the value of D. The same procedure is iterated until no more changes that increase D can be done. The greedy method and the iterative replacement method are presented in algorithmic form e.g. by Wright and Bailer (2006). More complicated search strategies have also been proposed in literature (Lejeune, 2003; Montepiedra et al., 1998).

4. Statistical analysis

The analysis of the data collected using D-optimal designs can be carried out using the likelihood-based approach for incomplete data (Rubin, 1976). Let $\mathcal{F}_0$ denote the data collected on (x, t, δ) at the first-stage for the entire study cohort. Now the selection of the individuals for collecting data on G is carried out using D-optimal design described earlier. Let $(i_1, \ldots, i_n)$ denote the order in which n individuals are selected sequentially for genotyping. According to the study design, individual i_k is selected conditional on all the information available at the time of selection and we denote this by $\mathcal{F}_{k-1} = (x, t, \delta, i_1, G(i_1), \ldots, i_{k-1}, G(i_{k-1})) = \mathcal{F}_0 \cup (i_1, G(i_1), \ldots, i_{k-1}, G(i_{k-1}))$. Further, the true value of $G(i_k)$ depends only on $\mathcal{F}_0$ and does not depend on i_k given $\mathcal{F}_{k-1}$. These characteristics of the design allows us to write the likelihood function in the product form as given below.

$$\begin{aligned} L(\psi,\theta) &= p_{\boldsymbol{\psi},\boldsymbol{\theta}}(G = g, \mathbf{X}, \mathbf{Y}) \\ &= p_{\boldsymbol{\psi},\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y}) \prod_{k=1}^{n} p(i_k \mid \mathcal{F}_{k-1}) p_{\boldsymbol{\psi},\boldsymbol{\theta}}(G(i_k) \mid \mathcal{F}_{k-1}, i_k) \\ &= p_{\boldsymbol{\psi},\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Y}) \prod_{k=1}^{n} p(i_k \mid \mathcal{F}_{k-1}) p_{\boldsymbol{\psi},\boldsymbol{\theta}}(G(i_k) \mid \mathcal{F}_0) \end{aligned}$$

$$\propto \prod_{j=1}^{N} p_{\psi,\theta}(x(j), t(j), \delta(j)) \prod_{k=1}^{n} p_{\psi,\theta}(G(i_k) \mid \mathcal{F}_0)$$

$$\propto \prod_{j=1}^{n} p_{\psi,\theta}(x(i_j), t(i_j), \delta(i_j)) p_{\psi,\theta}(G(i_j) \mid x(i_j), t(i_j), \delta_{i_j}) \prod_{j=n+1}^{N} p_{\psi,\theta}(x(i_j), t(i_j), \delta(i_j))$$

$$\propto \prod_{j=1}^{n} p_{\theta}(G(i_j)) p_{\psi}(x(i_j) \mid G(i_j)) p_{\theta}(t(i_j), \delta(i_j) \mid G(i_j), x(i_j))$$

$$\times \prod_{j=n+1}^{N} \sum_{g} p_{\theta}(G(i_j) = g) p_{\psi}(x(i_j) \mid g) p_{\theta}(t(i_j), \delta(i_j) \mid g, x(i_j)), \tag{39}$$

where $\mathbf{Y} = (t, \delta)$ and G is not observed for individuals $\{i_{n+1}, \ldots, i_N\}$.

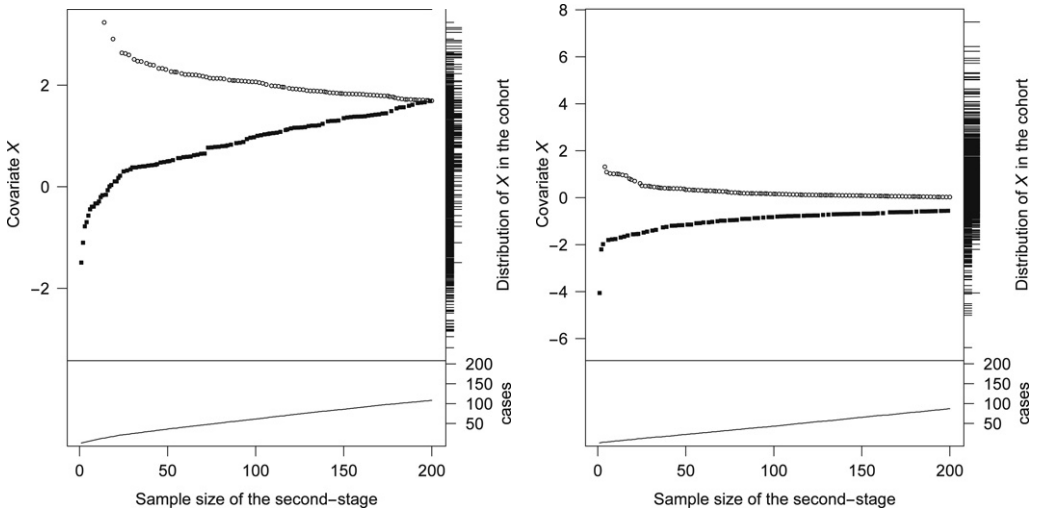
Because the missing genetic information can be handled with summation over the possible values of the genetic variables, it is often possible to maximize the above likelihood function by direct numerical maximization. In complicated situations, methods such as expectation-maximization (EM) algorithm (Scheike and Martinussen, 2004) or Bayesian data augmentation (Kulathinal and Arjas, 2006) can be applied to the above likelihood function to estimate the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. An alternative approach for analysis is post-stratification (Samuelsen et al., 2007) where the observations are weighted by the inverse of inclusion probabilities that are calculated by stratifying the selected individuals according to the non-genetic covariate x.

5. Examples

The simulation examples compare the D-optimal design with extreme selection, (generalized) case-cohort design and simple random sampling. D-optimal designs were found using the greedy method. The Iterative replacement method was also tried using the result of the greedy method as the initial design, but typically only one individual was changed. In extreme selection, the individuals for the second-stage are selected starting from the individuals with smallest and largest values of covariate x. The selection is made separately for cases and non-cases and the number of cases equal to the number of non-cases (presuming that there are enough cases). In the other words, there are four categories each containing one fourth of the second-stage sample: cases with high x, cases with low x, non-cases with high x and non-cases with low x. In the standard case-cohort design (Prentice, 1986) all cases and a random subset of the cohort are selected. The generalization applied here has the additional restriction that the number of cases cannot be greater than the number of non-cases. In simple random sampling, a random sample of the cohort is selected.

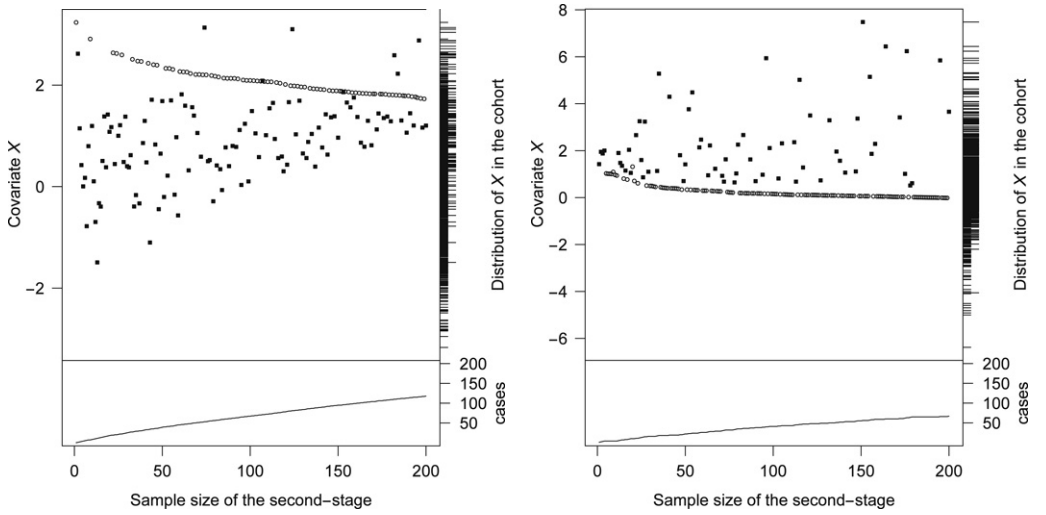
In the first simulation example, follow-up data for 2000 individuals are generated. The event times of a rare disease follow the Weibull regression model where the covariates are a normally distributed phenotype x (regression coefficient $a = 1$) and a genetic indicator variable g (regression coefficient $b = 0.5$, allele frequency $\pi = 0.4$). Phenotype x is generated from the distribution $N(\mu + \gamma g, \sigma^2)$, where $\mu = 0$, $\sigma^2 = 1$ and $\gamma = 0.3$. Covariate x is known for everyone but covariate g is unknown. During the follow-up of ten years, 138 disease events occurred. Our goal is to optimally select the individuals for the genotyping stage where the value of g is determined. We consider two types of response: the disease status (case/non-case) at the end of the study and right censored event times and model the data by the logistic model (12) or by the proportional hazards model (27), respectively. The parameters of the Weibull distribution are known for all designs because the focus is on the estimation of the parameters a, b and π. In the calculation of the D-optimal designs μ and σ^2 are estimated from the data and it is taken that $\gamma = 0$, which corresponds to assuming independence of x and g.

In the second simulation example we study a common disease and non-normally distributed phenotype. The example also demonstrates the effect of model misspecification because the models (16) and (27) assume that x follows a normal distribution. Note that the model misspecification affects all designs considered, including simple random sampling, because the parameters are estimated using the full likelihood approach. Event times follow the Weibull regression model where phenotype x is generated from a heavy-tailed normal-polynomial quantile mixture (Karvanen, 2006) with mean 0, L-scale 0.56, L-skewness 0.10 and L-kurtosis 0.35. The other simulation parameters except the Weibull parameters and covariate x are set as in the first simulation example. During the follow-up of ten years, 1398 disease events occurred.



(a) Rare disease, normally distributed X. (b) Common disease, non-normally distributed X.

Fig. 1. Sequential selection of observations to be included in the second-stage when the response is dichotomous. The greedy method is used. In the upper panels of (a) and (b), sample size on the x-axis indicates the order in which the observations should be included. Non-cases are marked by circles and cases are marked by squares. The y-axis on the left presents the covariate values of the selected observations. The tick-marks on the y-axis on the right present the distribution of the covariate values in the whole cohort. The longer tick-marks correspond to cases and the shorter tick-marks correspond to non-cases. The second-stage observations are selected from these 2000 observations. In the lower panels, the number of cases selected is shown as a function of the second-stage sample size.



(a) Rare disease, normally distributed X. (b) Common disease, non-normally distributed X.

Fig. 2. Sequential selection of observations to be included in the second-stage when the response is time-to-event. The greedy method is used. In the upper panels of (a) and (b), sample size on the x-axis indicates the order in which the observations should be included. Non-cases are marked by circles and cases are marked by squares. The y-axis on the left presents the covariate values of the selected observations. The tick-marks on the y-axis on the right present the distribution of the covariate values in the whole cohort. The longer tick-marks correspond to cases and the shorter tick-marks correspond to non-cases. The second-stage observations are selected from these 2000 observations. In the lower panels, the number of cases selected is shown as a function of the second-stage sample size.

The D-optimal designs for the simulation examples are illustrated in Figs. 1 and 2. It can be seen that the number of cases selected is smaller in the common disease scenario than in the rare disease scenario under both logistic and proportional hazards models. The figures also show the selection order in the greedy method. It can be seen that there are some patterns but it is not easy to give a universally applicable explanation for the observed selection order. A

Table 1
Comparison of different designs: rare disease, logistic model

Parameter	Design	$n = 100$		$n = 200$		$n = 500$	
		Estim.	SE	Estim.	SE	Estim.	SE
a	SRS	1.08	0.11	1.08	0.11	1.06	0.11
	CC	1.08	0.11	1.07	0.10	1.07	0.10
	Extreme	1.08	0.11	1.07	0.10	1.07	0.10
	D-optimal	1.08	0.10	1.07	0.10	1.07	0.10
b	SRS	0.63	0.85	0.65	0.61	0.64	0.38
	CC	0.52	0.40	0.54	0.29	0.59	0.22
	Extreme	0.58	0.37	0.60	0.27	0.59	0.20
	D-optimal	0.57	0.30	0.57	0.24	0.56	0.20
c	SRS	−3.54	0.46	−3.53	0.35	−3.53	0.24
	CC	−3.46	0.23	−3.48	0.19	−3.49	0.17
	Extreme	−3.52	0.23	−3.52	0.19	−3.49	0.17
	D-optimal	−3.49	0.20	−3.50	0.17	−3.49	0.17
$\pi = 0.4$	SRS	0.40	0.041	0.40	0.032	0.40	0.021
	CC	0.40	0.049	0.40	0.039	0.40	0.023
	Extreme	0.40	0.046	0.40	0.037	0.40	0.023
	D-optimal	0.40	0.046	0.40	0.035	0.40	0.021

The designs are simple random sampling (SRS), (generalized) case-cohort design (CC), extreme selection and D-optimal design. The reported point estimates and their standard errors are medians from 100 simulation runs.

reader interested in interpreting the selection order should compare these results with the theoretical results arising from Carathéodory's theorem applied to the support points of the canonical forms of nonlinear designs (Ford et al., 1992; Li and Majumdar, 2007).

In both rare disease and common disease scenarios, 100 datasets were generated and estimates of a, b, c and π were compared when the second-stage sample was selected according to different designs. The second-stage sample size had values $n = 100$, $n = 200$ and $n = 500$. The point estimates and their standard errors are reported in Tables 1 and 2 for the logistic model, and in Tables 3 and 4 for the proportional hazards model. Because the data were generated as time-to-event data, we cannot specify the true values of parameters a, b and c in the logistic model. However, we can see in the rare disease scenario that the point estimates of a and b in Table 1 are close to the point estimates of a and b from the proportional hazards model in Table 3. In the common disease scenario, covariate x was generated from normal-polynomial quantile mixture but modeled with normal distribution. This causes a bias that can be seen especially in the estimates of b in Tables 2 and 4. The bias seems to be a more serious problem in the logistic model. The results also suggests that the full likelihood inference based on extreme selection or D-optimal designs might be more sensitive to model misspecification than the inference based on simple random sampling. On the other hand, by looking at the standard errors of the estimates, the designs may be ranked as D-optimal design (best), extreme selection, case-cohort design, simple random sampling (worst). This order is what one might expect. The D-optimal design uses information on the observed phenotype and case status whereas the simple random sampling does not utilize the information available. It is worth of noting that extreme selection works very well taking into account the simplicity of the selection procedure.

In addition to the examples presented here, we also varied the simulation parameters to see what are their effects on the results. Different cohort sizes and different follow-up times provided essentially similar results. The effect of misspecification of the initial parameters need to be always considered when likelihood-based designs are used for non-linear models. In both examples, it was wrongly assumed in the calculation of D-optimal design that x and g are independent but this did not have a major effect on the results. According to our experiences, D-optimal designs are useful even if initial parameters are misspecified but the advantage compared to extreme selection and case-cohort design may be lost if the initial parameters are far from the correct values.

6. Discussion

In many practical situations, the second-stage sample size must be rather small compared to the first-stage sample size due to the cost consideration. Hence, the study has to be designed very carefully so that we can draw inferences about the questions of interest with the limited resources. The approach presented here offers an alternative to the

Table 2
Comparison of different designs: common disease, logistic model with misspecified distribution of covariate x

Parameter	Design	$n = 100$		$n = 200$		$n = 500$	
		Estim.	SE	Estim.	SE	Estim.	SE
a	SRS	1.80	0.13	1.73	0.11	1.70	0.11
	CC	1.77	0.12	1.73	0.11	1.70	0.11
	Extreme	1.75	0.11	1.72	0.11	1.70	0.11
	D-optimal	1.71	0.11	1.70	0.11	1.70	0.11
b	SRS	1.29	0.69	1.02	0.42	0.89	0.26
	CC	1.14	0.60	1.01	0.38	1.00	0.23
	Extreme	1.09	0.40	0.71	0.28	0.24	0.19
	D-optimal	0.69	0.27	0.37	0.20	0.44	0.16
c	SRS	0.52	0.19	0.61	0.14	0.64	0.097
	CC	0.54	0.18	0.61	0.14	0.60	0.095
	Extreme	0.61	0.12	0.69	0.11	0.84	0.089
	D-optimal	0.68	0.11	0.78	0.099	0.81	0.075
$\pi = 0.4$	SRS	0.40	0.039	0.40	0.031	0.40	0.021
	CC	0.40	0.041	0.40	0.032	0.39	0.022
	Extreme	0.37	0.042	0.39	0.032	0.39	0.022
	D-optimal	0.42	0.040	0.43	0.032	0.33	0.021

The designs are simple random sampling (SRS), (generalized) case-cohort design (CC), extreme selection and D-optimal design. The reported point estimates and their standard errors are medians from 100 simulation runs.

Table 3
Comparison of different designs: rare disease, proportional hazard model

Parameter	Design	$n = 100$		$n = 200$		$n = 500$	
		Estim.	SE	Estim.	SE	Estim.	SE
$a = 1$	SRS	0.99	0.095	0.99	0.093	0.98	0.087
	CC	1.00	0.088	0.99	0.082	0.99	0.077
	Extreme	0.99	0.087	0.98	0.082	0.99	0.076
	D-optimal	1.00	0.085	0.99	0.080	0.99	0.076
$b = 0.5$	SRS	0.50	0.24	0.52	0.23	0.53	0.20
	CC	0.50	0.22	0.54	0.19	0.52	0.16
	Extreme	0.50	0.21	0.52	0.18	0.52	0.15
	D-optimal	0.51	0.20	0.50	0.17	0.52	0.15
$\pi = 0.4$	SRS	0.40	0.041	0.40	0.032	0.40	0.021
	CC	0.40	0.045	0.40	0.037	0.40	0.023
	Extreme	0.41	0.045	0.40	0.035	0.40	0.022
	D-optimal	0.41	0.045	0.40	0.035	0.40	0.021

The designs are simple random sampling (SRS), (generalized) case-cohort design (CC), extreme selection and D-optimal design. The reported point estimates and their standard errors are medians from 100 simulation runs.

case-control and case-cohort designs when we already have some initial estimates of the parameters of interest and can therefore calculate the value of the D-criterion for candidate designs. The first-stage of multi-stage epidemiological studies is observational but the second-stage is almost experimental. We are of course restricted to the covariate values that have been observed at the first-stage but if the sample size of the first-stage is large and the sample size of the second-stage is comparatively small, we have a lot of freedom in the choice of the design.

The statistical model can be more complicated than the simplified model considered in the examples and instead of the D-criterion any other design criterion can be used. For instance, we could minimize the variance of the genetic effect. The functional form of the design criterion (that is the score functions required in (6) and (8)) for a particular model can be derived similarly to what was done with the D-criterion and logistic or Cox's regression model. The same optimization algorithms can be used for any design criterion.

The D-optimality approach proposed here is model-based because the Fisher information is model-based and this raises a question on the generality of the results. There are two types of misspecification that need to be taken into

Table 4
Comparison of different designs: Common disease, proportional hazard model with misspecified distribution of covariate x

Parameter	Design	$n = 100$		$n = 200$		$n = 500$	
		Estim.	SE	Estim.	SE	Estim.	SE
$a = 1$	SRS	0.99	0.026	0.99	0.026	1.00	0.026
	CC	0.99	0.026	0.99	0.026	0.99	0.026
	Extreme	1.00	0.026	1.01	0.025	1.02	0.025
	D-optimal	1.00	0.025	1.01	0.025	1.01	0.025
$b = 0.5$	SRS	0.50	0.078	0.50	0.070	0.49	0.060
	CC	0.50	0.075	0.50	0.068	0.50	0.058
	Extreme	0.48	0.070	0.45	0.061	0.38	0.053
	D-optimal	0.46	0.055	0.43	0.050	0.47	0.049
$\pi = 0.4$	SRS	0.41	0.038	0.41	0.030	0.40	0.021
	CC	0.40	0.037	0.40	0.029	0.39	0.021
	Extreme	0.41	0.039	0.42	0.031	0.41	0.021
	D-optimal	0.43	0.035	0.45	0.030	0.39	0.021

The designs are simple random sampling (SRS), (generalized) case-cohort design (CC), extreme selection and D-optimal design. The reported point estimates and their standard errors are medians from 100 simulation runs.

account: wrong distributional assumptions and misspecification of the initial estimates. The former problem is not specific to D-optimality or the models considered, but concerns the full likelihood analysis in general. The latter problem is relevant for design based on Fisher information and we recommend the sensitivity to the choice of the initial estimates be studied when D-optimal designs are applied. If the misspecification of initial estimates is a major concern, one might consider applying minimax (King and Wong, 2000; Sitter, 1992) or Bayesian approach (Zhou et al., 2003; Chaloner and Verdinelli, 1995).

On the basis of the simulation results, we recommend extreme selection as a practical study design. Extreme selection does not require initial estimates, is easy to implement and gives relatively good results compared to D-optimal design. It is probably possible improve the results of extreme selection further by specifying the ratio of cases and non-cases according to some suitable criterion. D-optimality and other criteria based on Fisher information provide the theoretical background for efficient study design and serve as benchmarks for the ad-hoc designs. One should be aware that if the data are analyzed using the full likelihood, also extreme selection may be sensitive to wrong distributional assumptions. This was seen in the second simulation example where the covariate x was generated from a non-normal distribution but modeled by a normal distribution and as result, especially the estimates of the genotype effect were clearly biased. Fortunately, the empirical distribution of x is observed and we have a possibility to check our distributional assumptions. The normality assumption is often justified, for instance, in the cardiovascular epidemiology, the risk factors or their log-transformations typically have a distribution close to the normal distribution.

D-optimal design and extreme selection may be applied also in situations where the number of genetic or non-genetic covariates is greater than one. For a vector $\mathbf{x}$ of non-genetic covariates we may consider the linear combination $z = \mathbf{ax}$, where $\mathbf{a}$ is a vector of initial parameter estimates, and proceed as in the case of a single non-genetic covariate. When there are several genetic covariates of interest, extreme selection can be applied without modifications and for D-optimal design we may compute the optimal design for a typical genetic covariate or alternatively define the selected subset as a union of the optimal designs computed separately for each genetic covariate.

Acknowledgements

The research of the first author was supported by the GenomEUtwin Project grant from the European Commission under the programme ‘Quality of Life and Management of the Living Resources’ of 5th Framework Programme and by the Academy of Finland via its grant number 53646. The present work was carried out when the second author was a visiting researcher at the Department of Mathematics and Statistics, University of Helsinki and her research was supported by the Academy of Finland via its grant number 114786. The research of the third author was supported by the Academy of Finland via its funding of the “Centre of Population Genetic Analyses”. Authors thank Dr. Mikko Sillanpää for useful discussions.

References

Allison, D.B., Schork, N.L., Wong, S.L., Elston, R.C., 1998. Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. Human Heredity 15, 261–267.

Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Design. Oxford University Press, Oxford.

Carey, G., Williamson, J., 1991. Linkage analysis of quantitative traits: Increased power by using selected samples. American Journal of Human Genetics 49, 786–796.

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statistical Science 10 (3), 273–304.

Clayton, D., Hills, M., 1993. Statistical Models in Epidemiology. Oxford University Press, New York.

Cook, R.D., Nachtsheim, C.J., 1980. A comparison of algorithms for constructing exact D-optimal design. Technometrics 22, 315–324.

Darvasi, A., Soller, M., 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. TAG Theoretical and Applied Genetics 85 (2–3), 353–359.

Dykstra, O., 1971. The augmentation of experimental data to maximize $|X'X|$. Technometrics 13, 682–688.

Elfving, G., 1952. Optimum allocation in linear regression theory. The Annals of Mathematical Statistics 23 (2), 255–262.

Ford, I., Torsney, B., Wu, C.F.J., 1992. The use of a canonical form in the construction of locally optimal designs for non-linear problems. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 54 (2), 569–583.

Karvanen, J., 2006. Estimation of quantile mixtures via L-moments and trimmed L-moments. Computational Statistics & Data Analysis 51 (2), 947–959.

King, J., Wong, W.-K., 2000. Minimax D-optimal designs for the logistic model. Biometrics 56, 1263–1267.

Kulathinal, S., Arjas, E., 2006. Bayesian inference from case-cohort data with multiple end-points. Scandinavian Journal of Statistics 33, 25–36.

Lander, E.S., Botstein, D., 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121, 185–199.

Langholz, B., 2007. Use of cohort information in the design and analysis of case-control studies. Scandinavian Journal of Statistics 34, 120–136.

Langholz, B., Borgan, O., 1995. Counter-matching: A stratified nested case-control sampling method. Biometrika 82, 69–79.

Lejeune, M.A., 2003. Heuristic optimization of experimental designs. European Journal of Operational Research 147 (3), 484–498.

Li, G., Majumdar, D., 2007. D-optimal designs for logistic models with three and four parameters. Journal of Statistical Planning and Inference. doi:10.1016/j.jspi.2007.07.010.

Macgregor, S., Craddock, N., Holmans, P.A., 2006. Use of phenotypic covariates in association analysis by sequential addition of cases. European Journal of Human Genetics 14, 529–534.

McElroya, J.P., Zhangb, W., Koehlerb, K.J., Lamonta, S.J., Dekkersa, J.C., 2006. Comparison of methods for analysis of selective genotyping survival data. Genetics Selection Evolution 38, 637–655.

McNamee, R., 2002. Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. Statistics in Medicine 21, 3609–3625.

Montepiedra, G., Myers, D., Yeh, A.B., 1998. Application of genetic algorithms to the construction of exact D-optimal designs. Journal of Applied Statistics 25 (6), 817–826.

Myers, W.R., Myers, R.H., Carter, W.H., White, K.L., 1996. Two-stage designs for the logistic regression model in single-agent bioassays. Journal of Biopharmaceutical Statistics 6 (3), 283–301.

Prentice, R.L., 1986. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73 (1), 1–11.

Pukelsheim, F., 1993. Optimal Design of Experiments. Wiley, New York.

Reilly, M., 1996. Optimal sampling strategies for two-stage studies. American Journal of Epidemiology 143 (1), 92–100.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63 (3), 581–592.

Samuelsen, S.O., Ånestad, H., Skrondal, A., 2007. Stratified case-cohort analysis of general cohort sampling designs. Scandinavian Journal of Statistics 34 (1), 103–119.

Scheike, T.H., Martinussen, T., 2004. Maximum likelihood estimation for Cox's regression model under case-cohort sampling. Scandinavian Journal of Statistics 31, 283–293.

Sitter, R.R., 1992. Robust designs for binary data. Biometrics 48, 1145–1155.

Tenesa, A., Visscher, P.M., Carothers, A.D., Knott, S.A., 2005. Mapping quantitative trait loci using linkage disequilibrium: Marker- versus trait-based methods. Behavior Genetics 35, 219–228.

Van Gestel, S., Houwing-Duistermaat, J.J., Adolfsson, R., van Duijn, C.M., Broeckhoven, C.V., 2000. Power of selective genotyping in genetic association analyses of quantitative traits. Behaviour Genetics 30 (2), 141–146.

Wright, S.E., Bailer, A.J., 2006. Optimal experimental design for a nonlinear response in enviromental toxicology. Biometrics 62, 886–892.

Zhou, X., Joseph, L., Wolfson, D.B., Bélishe, P., 2003. A Bayesian A-optimal and model robust design criterion. Biometrics 59, 1082–1088.