Statistica Sinica

COST CONSIDERATIONS FOR EFFICIENT GROUP TESTING STUDIES

Shih-Hao Huang[1,2], Mong-Na Lo Huang[3] and Kerby Shedden[2]

[1]*Academia Sinica,* [2]*University of Michigan and* [3]*National Sun Yat-sen University*

Abstract: A group testing study involves collecting samples from multiple individuals, pooling them, and testing them as a group. A realistic cost model for such a study should consider the costs both for collecting the samples, and for running the assays. Moreover, an efficient design should accommodate inaccuracies in any prespecified nominal test sensitivity and specificity values, and allow them to vary with group size. In this work, we derive locally optimal designs in this setting, and characterize their theoretical properties. We also provide a guaranteed algorithm for constructing the designs on discrete design spaces. Several simulated examples based on a real-world group testing study show that the proposed designs have high efficiency, and are not strongly sensitive to the working parameter specification that is used to obtain the locally optimal design.

Key words and phrases: Budget-constrained design, dilution effect, group testing.

1. Introduction

Group testing, first discussed by Dorfman (1943), plays an important role in prevalence estimation and case diagnosis, and may become

increasingly important in public health, environmental monitoring, and risk surveillance, as sensors, assays, and data-driven risk monitoring proliferate – see for example, Gastwirth (2000), Xie et al. (2001), Pilcher et al. (2005) and Liu et al. (2011). A successful group testing study should be based on an efficient and tractable design, in order to get the most information out of limited resources. One critical aspect of efficient design in this setting is the overall study cost (Turner, Stamey and Young, 2009), which arises from separate costs due to collecting samples and running assays. Another important issue is that specificity and especially sensitivity of the test may decline with increasing group sizes, which is called *dilution effects* in the literature (Zenios and Wein, 1998; McMahan, Tebbs and Bilder, 2013).

Many group testing studies for prevalence estimation utilize prespecified values for the sensitivity and specificity, and therefore their designs involve only one group size (Tu, Litvak and Pagano, 1995; Liu et al., 2012). However, Zhang et al. (2014) indicate that misspecified sensitivity and specificity may introduce bias in the prevalence estimate. Therefore, here we estimate the prevalence while treating the sensitivity and specificity as nuisance parameters inferred from the data.

Huang et al. (2017) theoretically characterize optimal designs for group testing with uncertain testing parameters. However, since they do not in-

corporate costs for assays and subjects, the optimal designs may place untenably large numbers of subjects into large groups. In group testing, large groups are important for sensitivity estimation, but it is arguably unlikely that scarce samples would be utilized this way in practice. Therefore, introducing differential costs for subjects and assays, and in particular placing a realistic cost on subject recruitment can lead to optimal designs in which the largest group sizes are moderated.

Here we develop a theory and an algorithm to obtain optimal designs for prevalence estimation in a realistic group testing setting. We allow different costs for assays and for subjects, and accommodate uncertain test accuracies which may vary with the group sizes. Our results indicate that the optimal design substantially depends on the relative costs of assays and subjects. Therefore, a simplified approach in which either assays or subjects are taken to be cost-free may not be appropriate in many cases.

2. Preliminaries

Let $\theta = (p_0, p_1, p_2)^{\mathrm{T}}$, where p_0 is the prevalence (the proportion of diseased people in the population), and p_1 and p_2 are the sensitivity and specificity (true positive rate and the true negative rate of the test, respectively). We first consider the case with unknown sensitivity and specificity that do not change with the group size. We assume that $p_0 \in (0, 1)$, $p_1, p_2 \in (0.5, 1]$,

and false positives and false negatives occur randomly with rates $1-p_2$ and $1-p_1$, respectively. Hence, the positive response probability (either true or false positive) of a trial with group size x is

$$\pi(x) = \pi(x|\theta) = p_1 - (p_1 + p_2 - 1)(1 - p_0)^x . \tag{2.1}$$

We consider designs subject to a known group size constraint $1 \leq x_L \leq x \leq x_U < \infty$, where the limits on the group sizes are driven from practical considerations such as the feasibility of the test. We note that when the upper bound x_U is large enough, it is often not a support point of the optimal design in our setting, and therefore it does not impact the design or analysis.

To introduce costs, we let the total budget be C_0, and we assume that the costs of performing an assay and enrolling a subject are, respectively, q_0 and q_1, which in practice are known, where $q_0, q_1 \geq 0$ and $q_0 + q_1 > 0$. Without loss of generality, we rescale the total budget and the costs for assay and subject with respect to the cost for individual test, $q_0 + q_1$. That is, the (rescaled) total budget is $C = C_0/(q_0 + q_1)$, and the (rescaled) costs for assay and subject are $1 - q$ and q, respectively, for $q = q_1/(q_0 + q_1) \in [0, 1]$. We then model the cost of a trial with group size x as

$$c(x) = 1 - q + qx.$$

Under a fixed budget, having $q = 0$ means that subjects incur no cost, thus is equivalent to the scenario with a fixed number of trials. Similarly, the scenario with $q = 1$, i.e., assays are cost-free, is equivalent to the scenario with a fixed number of subjects.

For a study consisting of n_i trials with group size x_i for $i = 1, \ldots, k$, we denote its *budget-constrained design* as $\xi = \{(x_i, w_i)\}_{i=1}^{k}$, where w_i is the proportion of budget expended at group size x_i, expressed as

$$w_i = n_i c(x_i)/C, \tag{2.2}$$

and the total budget $C = \sum_j n_j c(x_j)$. The log-likelihood function in θ is (omitting an unimportant additive constant)

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{k} \{y_i \log(\pi(x_i|\theta)) + (n_i - y_i) \log(1 - \pi(x_i|\theta))\} \\ &= C \left(\sum_{i=1}^{k} \frac{w_i}{c(x_i)} \left\{ \frac{y_i}{n_i} \log(\pi(x_i|\theta)) + \left(1 - \frac{y_i}{n_i}\right) \log(1 - \pi(x_i|\theta)) \right\} \right). \end{aligned} \tag{2.3}$$

The maximum likelihood estimate (MLE) of θ, $\hat{\theta}$, is obtained by maximizing (2.3), and the covariance matrix of $\hat{\theta}$ is asymptotically proportional to the inverse of the information matrix of ξ, which is

$$M(\xi) = \sum_{i=1}^{k} w_i \lambda(x_i) f(x_i) f(x_i)^{\mathrm{T}}, \tag{2.4}$$

where

$$\begin{aligned}\lambda(x) &= \{c(x)\pi(x)(1-\pi(x))\}^{-1}, \\ f(x) &= \left((p_1+p_2-1)x(1-p_0)^{x-1}, 1-(1-p_0)^x, -(1-p_0)^x\right)^{\mathrm{T}}.\end{aligned}$$

We note that in equations (2.3) and (2.4), $c(x)$ plays the role of an inverse weight in both the log-likelihood function and the information matrix.

Our main goal is to accurately estimate the prevalence, where other unknown parameters are treated as nuisance parameters. Therefore, we use the D_s-optimality criterion, which seeks a design minimizing the asymptotic generalized variance of a given subset of model parameters. In this study a D_s-optimal design maximizes

$$\Phi_s\{M(\xi)\} = -\log\left(M(\xi)^-\right)_{11} \tag{2.5}$$

among all designs under which p_0 is estimable, where for a matrix of M, M_{11} is its $(1,1)$ entry and M^- is a generalized inverse M. Note that the D_s-optimality above is equivalent to c-optimality with $c = (1,0,0)^{\mathrm{T}}$ (Atkinson, Donev and Tobias, 2007, Chap. 17.5), which minimizes the asymptotic variance of $c^{\mathrm{T}}\hat{\theta}$. The optimal group sizes of a D_s-optimal design may be non-integer-valued. For practical use, we further say that a design is D_s^I-optimal ('I' stands for 'integers') if it is D_s-optimal among all designs supported on the positive integers $[x_L, x_U] \cap \mathbb{N}$. According to (2.4) and (2.5), we can see

that the optimality of a design depends on unknown parameters $(p_0, p_1, p_2)^{\mathrm{T}}$ and the cost parameter q, but is invariant to the total budget C.

3. D_s-optimal budget-constrained designs

We first consider the design space as the interval $[x_L, x_U]$ to get an overview of the behavior of D_s-optimal budget-constrained designs. The main tools used in this section are the general equivalence theorem (Kiefer, 1974) and the following two lemmas. Note that the three results still hold when the design space $[x_L, x_U]$ is replaced by $[x_L, x_U] \cap \mathbb{N}$, which are used to obtain D_s^I-optimal designs in Section 3.1. For the D_s-criterion, we say that a design ξ with finitely many group sizes is *valid* if p_0 is estimable under ξ. The first result describes the collection Ξ of all valid designs, through the following lemma. The proofs of this lemma and other results are detailed in the on-line supplement.

Lemma 1. *For the D_s-criterion (2.5), Ξ consists of all designs having at least three support points in $[x_L, x_U]$.*

This lemma also shows that all valid designs under model (2.1) have nonsingular information matrices. Moreover, for three group sizes $x_1 < x_2 < x_3$, letting $F = (f(x_1), f(x_2), f(x_3))$, $(v_1, v_2, v_3)^{\mathrm{T}} = F^{-1} \cdot (1, 0, 0)^{\mathrm{T}}$, $u_i = \{\lambda(x_i)^{-1} v_i^2\}^{1/2}$ and $w_i^s = u_i / \sum_j u_j$ for $i = 1, 2, 3$. We have the following lemma to determine the optimal weights on the three group sizes.

Moreover, when a three-point design is described by its support points, its weights are obtained from this lemma without further mention.

Lemma 2. *The weights* $\{w_i^s\}_{i=1}^3$ *are the unique optimal weights for the group size* $x_1 < x_2 < x_3 \in [x_L, x_U]$*, with*

$$\Phi_s\{M(\{x_i, w_i^s\}_{i=1}^3)\} = -2\log\sum_{i=1}^{3} u_i. \tag{3.1}$$

For completeness, we introduce the general equivalence theorem as follows. For $x \in [x_L, x_U]$, let δ_x be the one-point design supported on x. For a design $\xi \in \Xi$ and $x \in [x_L, x_U]$, let $\phi_s(x, \xi)$ be the directional derivative of Φ_s at $M(\xi)$ in the direction $M(\delta_x)$,

$$\begin{aligned}\phi_s(x,\xi) &= \lim_{\alpha\to 0^+}\frac{1}{\alpha}\left(\Phi_s\{M(\xi)\} - \Phi_s\{M((1-\alpha)\xi + \alpha\delta_x)\}\right)\\ &= \lambda(x)f(x)^{\mathrm{T}}M^{-1}(\xi)f(x) - \lambda(x)f_s(x)^{\mathrm{T}}M_s^{-1}(\xi)f_s(x) - 1,\end{aligned} \tag{3.2}$$

where $f_s(x)$ is the 2×1 subvector of $f(x)$ deleting its first element, and $M_s(\xi)$ is the 2×2 submatrix of $M(\xi)$ after deleting its first row and first column. Then we have the following general equivalence theorem.

Theorem 1. *For a design* $\xi_s \in \Xi$*, the three assertions are equivalent:*

(a) $\Phi_s\{M(\xi_s)\} = \max_{\xi\in\Xi}\Phi_s\{M(\xi)\}$*;*

(b) $\max_{x\in[x_L,x_U]}\phi_s(x,\xi_s) = \min_{\xi\in\Xi}\max_{x\in[x_L,x_U]}\phi_s(x,\xi)$*;*

(c) for an arbitrary group size x_s *of* ξ_s*,* $\phi_s(x_s,\xi_s) = \max_{x\in[x_L,x_U]}\phi_s(x,\xi_s) = 0$.

Any linear combination of designs satisfying (a)-(c) also satisfies (a)-(c).

Based on Lemmas 1, 2, and Theorem 1, we characterize the D_s-optimal design by Theorem 2. It extends Theorem 3 in Huang et al. (2017) from the special case with cost parameter $q = 0$ to an arbitrary $q \in [0, 1]$.

Theorem 2. *The D_s-optimal design ξ_s for estimating only the prevalence is unique. It has three group sizes $x_L = x_1^s < x_2^s < x_3^s \leq x_U$ with weights given in Lemma 2.*

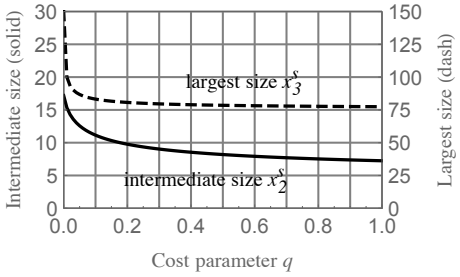
Theorem 2 shows that some properties of Theorem 3 in Huang et al. (2017) continue to hold for $q \in [0, 1]$: (i) the unique D_s-optimal design has exactly three group sizes; to run a design with four or more distinct sizes would lose efficiency for prevalence estimation; (ii) the information about the prevalence, sensitivity, and specificity mainly come from x_2^s, x_3^s, and x_1^s, respectively; (iii) having a smaller value of x_L strictly improves the accuracy of the estimation of p_0, so x_L should be set to one whenever possible.

On the other hand, as q increases, the inverse weight $c(x)$ tends to penalize larger group sizes. Therefore, when $q > 0$, x_U may not be a support point of ξ_s, and thus a two-dimensional optimization problem (x_2 and x_3 in equation (3.1)) needs to be solved for obtaining ξ_s. In contrast, Theorem 3 in Huang et al. (2017) showed that when $q = 0$, x_3^s must be x_U, and x_2^s can be obtained via a one-dimensional root finding algorithm.

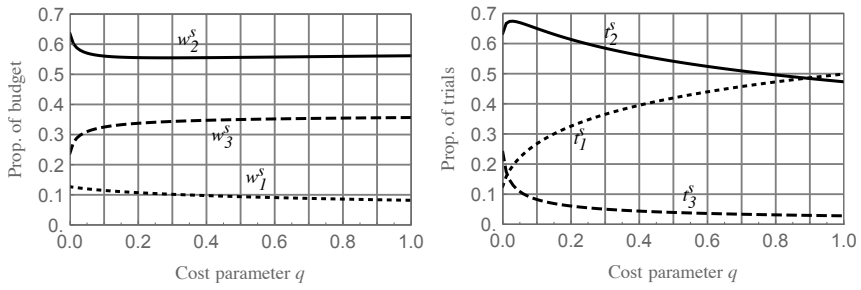
Example 1. Let $\theta = (p_0, p_1, p_2)^{\mathrm{T}} = (0.07, 0.93, 0.96)^{\mathrm{T}}$ (this parameter setting is based on a chlamydia study in McMahan, Tebbs and Bilder (2012)) and let $[x_L, x_U] = [1, 150]$. We obtain the D_s-optimal design for each $q \in [0, 1]$, as shown in Figure 1. First we focus on the group sizes of the D_s-optimal designs. Theorem 2 shows that the smallest group size x_1^s of ξ_s must be the lower boundary x_L. In Figure 1(a) we observe that the intermediate and largest group sizes x_2^s and x_3^s decrease as the cost parameter q increases. Moreover, when x_U is as large as 150, the largest group size of a D_s-optimal design is strictly less than x_U unless q approaches 0.

The optimal weights (proportions of budget) $\{w_1^s, w_2^s, w_3^s\}$ of D_s-optimal designs are shown in Figure 1(b). Under this parameter setting, the weight w_2^s at x_2^s always dominates the other two weights w_1^s and w_3^s. By observing Figures 1 (a) and (b) we note that what really matters is whether $q \approx 0$ or not, and the D_s-optimal designs are quite stable when $q \geq 0.4$, which is roughly supported on $\{1, 7, 77\}$ with weights $\{0.09, 0.55, 0.36\}$.

Figure 1(c) shows another perspective on trial allocation related to cost. Roughly speaking, we find that as the cost parameter q increases, only the proportion of trials t_1^s at x_1^s increases, but the other two proportions decrease. By comparing Figures 1(b) and (c), we conclude that, as q increases, to obtain a trial with large group size gets more expensive, and more budget



(a) Group sizes against q



(b) Proportions of budget against q (c) Proportions of trials against q

Figure 1: Properties of the D_s-optimal designs for Example 1.

at larger sizes is needed to get enough information about p_1 for efficiently estimating p_0; on the other hand, the proportion of trials at the smallest size still increases, which reflects the preference for less expensive trials. □

Remark 1. In the online supplement, Section S2, we also consider the D-optimal design under the setting of Example 1, where D-criterion treats p_0, p_1, and p_2 as equally important. The D-optimal design also have exactly three group sizes with the low boundary x_L, and an intermediate size close to that of the D_s-optimal design. However, a D_s-optimal design puts much

more weight (proportion of budget) on its intermediate size (≥ 0.55 vs. 0.33). □

3.1. D_s^I-optimal designs

In practice, the group sizes in a group testing design must be supported on the finite set $[x_L, x_U] \cap \mathbb{N}$ instead of the interval $[x_L, x_U]$. To obtain the optimal integer-valued group sizes, a natural approach would be simple rounding of the D_s-optimal design ξ_s; however, to attain optimality we develop an efficient numerical search procedure that yields D_s^I-optimal designs on $[x_L, x_U] \cap \mathbb{N}$.

Intuitively, a D_s^I-optimal design should be close to the corresponding D_s-optimal design ξ_s obtained based on Theorem 2. Therefore, the three support points of ξ_s after rounding form a good initial design. Then, by Theorem 1, we know that either the initial design is optimal, or it can be improved by adding a point which has a positive derivative (3.2). We then recalculate the weights, by numerically optimizing (2.5). After dropping points with zero weight, if any exist, we check the optimality of the new design. These steps are iterated until optimality is attained.

The algorithm stops when the resulting design satisfies Theorem 1(c), which guarantees optimality; otherwise, the design obtained in each iteration is strictly better than the previous ones. Since $[x_L, x_U] \cap \mathbb{N}$ is finite,

this algorithm must stop in finitely many steps. Also, due to the convexity of the design criterion, this stepwise ascent algorithm converges to a global optimum. The details of the search algorithm for obtaining a D_s^I-optimal design ξ_I are shown later in Section 4, together with the scenario having dilution effects, as Algorithm 1. In practice, we have observed that the algorithm tends to converge in very few steps, since the initial design is often already close to (and in many cases, exactly equal to) a D_s^I-optimal design.

We note that heuristic numerical search (e.g. Zhang et al. (2014)) may yield incorrect results. When $\theta = (0.05, 0.95, 0.995)^{\mathrm{T}}$, $q = 1$, and $[x_L, x_U] = [1, 150]$, our algorithm obtains a D_s^I-optimal design having group sizes $\{1, 8, 113\}$. Alternatively, a design supported on $\{1, 12, 150\}$ is reported by Zhang et al. (2014). By using Theorem 1, we find that our design is optimal, but the other is not.

3.2. Design implementation

In the approximate design framework, the optimal weights only involve the constraints $w_i > 0$ and $\sum w_i = 1$. For practical use with a total budget C, equation (2.2) shows that the number of trials at each point x_i is $n_i = Cw_i/c(x_i)$, which should be positive integers, introducing additional restrictions on the weights.

For implementing a D_s^I-optimal design ξ_I, we obtain the number of

Table 1: Allocations of C_1 on support points $\{1, 10, 81\}$ when $C = 10000$.

x_i	1	10	81	remaining	$\mathrm{Var}(\hat{p}_0)$
$c(x_i)$	1	2.8	17	budget	$(\times 1/C)$
additional	0	4	0	0.0	0.137634
trials Δ_i	2	3	0	0.8	0.137645
	5	2	0	0.6	0.137642
	8	1	0	0.4	0.137640
	11	0	0	0.2	0.137638

trials based on a variant of the efficient rounding procedure (Pukelsheim, 2006). For a budget-constrained design $\xi = \{(x_i, w_i)\}_{i=1}^k$ and a total budget C, let $\{n_i^0\}_{i=1}^k = \{\lfloor Cw_i/c(x_i) \rfloor\}_{i=1}^k$ and let $C_1 = C - \sum_i n_i^0 c(x_i)$, where $\lfloor x \rfloor$ is the largest integer that is not greater than x. Then, we allocate C_1 at each x_i to obtain a design having trial counts $\{n_i^0 + \Delta_i\}_{i=1}^k$ with minimum variance of the prevalence estimator, where $\Delta_i \in \mathbb{N} \cup \{0\}$ for each i and $\sum \Delta_i c(x_i) \leq C_1$. Note that $\sum \Delta_i c(x_i)$ may be strictly less than C_1, when the remaining cost is less than $c(\min(x_i))$. Some details are presented in the following example.

Example 2. Following Example 1, we let $\theta = (0.07, 0.93, 0.96)^{\mathrm{T}}$, let $q = 0.2$, and let $[x_L, x_U] \cap \mathbb{N} = \{1, 2, \ldots, 150\}$. A D_s^I-optimal design ξ_I is sup-

ported on {1, 10, 81} with weights {0.104, 0.555, 0.341} and costs per test {1, 2.8, 17}, respectively. The asymptotic variance of prevalence estimate from ξ_I is $0.137633/C$.

When the total budget is C =10,000, we have $\{n_i^0\}_{i=1}^3 = \{1042, 1981, 200\}$ and $C_1 = 11.2$. Table 1 shows all possible allocations of C_1, and the variance attains its minimum when the additional trials are at $\{0, 4, 0\}$. Thus, we set the numbers of trials of the implemented design $\xi_I(C)$ to be $\{1042, 1985, 200\}$, with total number of trials 3,227, and total number of subjects 37,092. We also note that, when C is large enough, such as this example, the loss of design efficiency tends to be negligible, no matter how we allocate C_1 in Table 1. □

4. D_s^I-optimal designs under dilution effects

In Section 3 we treated the sensitivity and specificity as constants with unknown values. As noted in the introduction, dilution effects, which reduce sensitivity or specificity for larger group sizes, are commonly seen, especially when the allowable range of group sizes $[x_L, x_U]$ is wide. In this section, we provide an algorithm to accommodate group testing with dilution effects.

The most natural form for dilution is decreasing sensitivity with increasing group size (Zenios and Wein, 1998). For completeness, we also consider the presence of diluted specificity. When there is a dilution effect

on the sensitivity or on the specificity, we work respectively with the model,

$$p_1(x) = p_1(x|\alpha) = \mathit{link}(\alpha_0 - \alpha_1 \log(x)) \quad \text{or} \tag{4.1}$$

$$p_2(x) = p_2(x|\beta) = \mathit{link}(\beta_0 - \beta_1 \log(x)), \tag{4.2}$$

where $\mathit{link} : \mathbb{R} \to [0, 1]$ is a link function for probability. For convenience of interpreting the dilution models, we adopt the logistic regression in the following context: $\mathit{link}(u) = \text{expit}(u) = \{1 + \exp(-u)\}^{-1}$ (also see equation (4) in Zhang et al. (2014)). Thus, for instance, the sensitivity model has the properties that $\text{expit}(\alpha_0)$ is the baseline sensitivity $p_1(1)$, and that the sensitivity has a nearly polynomial rate of decay, $\{1 + x^{\alpha_1} \exp(-\alpha_0)\}^{-1}$, as the group size x grows. In other scenarios, $\log(x)$ in equations (4.1) and (4.2) can be replaced by x, $\log^2(x)$, etc., and another link function can be adopted. Here we assume that $\alpha_0, \beta_0 > 0$ and $\alpha_1, \beta_1 \geq 0$ so that $p_1(1), p_2(1) > 0.5$ and p_1, p_2 monotonically decrease as x increases.

When only the sensitivity has a dilution effect, the corresponding information matrix becomes a variant of (2.4)

$$M_\alpha(\xi) = \sum_{i=1}^{k} w_i \lambda(x_i) f_\alpha(x_i) f_\alpha(x_i)^{\mathrm{T}}, \tag{4.3}$$

where $f_\alpha(x) = H_\alpha(x) f(x) \in \mathbb{R}^4$, and $H_\alpha(x)$ is a 4×3 block-diagonal matrix with diagonal blocks $(1, \partial p_1(x)/\partial\alpha, 1)$. Similarly, when only the specificity has a dilution effect, or when both the sensitivity and specificity have dilu-

tion effects, the corresponding information matrices are, respectively,

$$M_\beta(\xi) = \sum_{i=1}^{k} w_i \lambda(x_i) f_\beta(x_i) f_\beta(x_i)^{\mathrm{T}}, \quad \text{and} \tag{4.4}$$

$$M_{\alpha\beta}(\xi) = \sum_{i=1}^{k} w_i \lambda(x_i) f_{\alpha\beta}(x_i) f_{\alpha\beta}(x_i)^{\mathrm{T}}, \tag{4.5}$$

where $f_\beta(x) = H_\beta(x) f(x) \in \mathbb{R}^4$, $f_{\alpha\beta}(x) = H_{\alpha\beta}(x) f(x) \in \mathbb{R}^5$, $H_\beta(x) = \mathrm{diag}(1, 1, \partial p_2(x)/\partial\beta)$, and $H_{\alpha\beta}(x) = \mathrm{diag}(1, \partial p_1(x)/\partial\alpha, \partial p_2(x)/\partial\beta)$.

Extending the ideas in Section 3.1, our search algorithm is described as follows. By Theorem 2, the D_s-optimal design supported on $\{x_1^s, x_2^s, x_3^s\}$ can be used to efficiently estimate p_0 in the absence of dilution effects, and when dilution effects exist, more points should be added. Note that in Theorem 2, the information of p_1 and p_2 mainly comes from the larger and smaller group sizes, x_3^s and x_1^s, respectively. Therefore, to form an initial design, we add a size between x_2^s and x_3^s (or a size in (x_1^s, x_2^s)) if the sensitivity (or specificity) is diluted. The optimal weights for these group sizes can be obtained through Lemma 2 (when there is no dilution effect) or the lemma below (when dilution effects exist).

Lemma 3.

(a) For group testing with one dilution effect (either sensitivity or specificity), if the four distinct sizes $\{x_1, x_2, x_3, x_4\}$ *satisfy that* $F_* = (f_*(x_1), f_*(x_2), f_*(x_3), f_*(x_4))$ *is invertible, where* $f_* = f_\alpha$ *or* f_β*, respectively,*

then the D_s-optimal weights at these sizes are proportional to $(\lambda(x_i)^{-1}v_i^2)^{1/2}$ *for* $i = 1 \dots 4$, *where* $(v_1, v_2, v_3, v_4) = F_*^{-1} \cdot (1, 0, 0, 0)^{\mathrm{T}}$.

(b) For group testing with two dilution effects (both sensitivity and specificity), if the five distinct sizes $\{x_1, x_2, x_3, x_4, x_5\}$ *satisfy that* $F_{\alpha\beta} = (f_{\alpha\beta}(x_1), f_{\alpha\beta}(x_2), f_{\alpha\beta}(x_3), f_{\alpha\beta}(x_4), f_{\alpha\beta}(x_5))$ *is invertible, the D_s-optimal weights at these sizes are proportional to* $(\lambda(x_i)^{-1}v_i^2)^{1/2}$ *for* $i = 1 \dots 5$, *where* $(v_1, v_2, v_3, v_4, v_5) = F_{\alpha\beta}^{-1} \cdot (1, 0, 0, 0, 0)^{\mathrm{T}}$.

Based on the ideas above, and applying Theorem 1 (where the information matrix (2.4) should be replaced by M_α, M_β, or $M_{\alpha\beta}$ if dilution effects exist), we provide Algorithm 1 to obtain D_s^I-optimal designs. The use of Algorithm 1 is demonstrated in the example below, and we also comment on some of the patterns we observe.

Algorithm 1. *Let* $\Omega = [x_L, x_U] \cap \mathbb{N}$ *and let* $x_1^{(0)} < x_2^{(0)} < x_3^{(0)}$ *be the three support points of* ξ_s *for* $\theta = (p_0, p_1(1), p_2(1))^{\mathrm{T}}$ *in Theorem 2 after rounding. Set* $X_0 = \left\{x_1^{(0)}, x_\beta^{(0)} \text{(if } p_2 \text{ is diluted)}, x_2^{(0)}, x_\alpha^{(0)} \text{(if } p_1 \text{ is diluted)}, x_3^{(0)}\right\}$, *where* $x_\alpha^{(0)} = \lfloor (x_2^{(0)} + x_3^{(0)})/2 \rfloor$ *and* $x_\beta^{(0)} = \lfloor (x_1^{(0)} + x_2^{(0)})/2 \rfloor$. *Set W_0 be the optimal weights obtained from Lemma 2 or Lemma 3 at points in X_0. Set* $\xi_0 = \{X_0, W_0\}$. *For* $j = 0, 1, \dots,$ *do:*

Step 1. *Set* $x_j = \arg\max_{\Omega \setminus X_j} \phi_s(x, \xi_j)$. *If* $\phi_s(x_j, \xi_j) \le 0$, *output* ξ_j *and stop.*

Step 2. *Set* $X_{j+1} = X_j \cup \{x_j\}$*, and obtain*

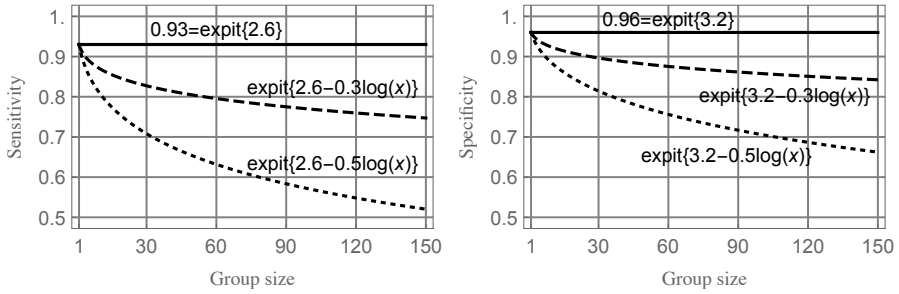
$$W_{j+1} = \arg\max_{W} \left\{ \Phi_s\{M(X_{j+1}, W)\};\ \min(W) \geq 0,\ \sum_{w \in W} w = 1 \right\}.$$

The weights W *are available in closed-form (Lemmas 2 and 3) if the design is minimally supported. Otherwise,* W *can be obtained by solving a convex optimization.*

Step 3. *Set* $\xi_{j+1} = \{X_{j+1}, W_{j+1}\}$ *after deleting those* (x, w) *with* $w = 0$*.*

Example 3. To better understand the structure of optimal designs in the presence of dilution effects, and how they relate to the optimal designs in the setting without dilution, we considered several numerical examples. Following Examples 1 and 2, we let $p_0 = 0.07$, $[x_L, x_U] \cap \mathbb{N} = \{1, 2, \ldots, 150\}$, $q = 0.2$. We further let the sensitivity and the specificity be respectively 0.93 and 0.96 at group size 1 ($\alpha_0 = 2.6$ and $\beta_0 = 3.2$), and consider α_1, β_1 respectively vary from 0 to 0.5. Figure 2 shows how the sensitivity and specificity decay as the group sizes.

Table 2 shows D_s^I-optimal designs for several setting with or without dilution effects. When there is no dilution effect, the design supported on $\{1, 10, 81\}$ is D_s^I-optimal under the model with information matrix (2.4). When the experimenters include dilution effects in the group testing model, the information matrix becomes (4.3), (4.4) or (4.5).



(a) Sensitivity functions (b) Specificity functions

Figure 2: The sensitivity and specificity functions for Example 3.

If sensitivity is diluted, the new support point falls between x_2 and x_3 but does not approach either of them, while if specificity is diluted, the new support point falls near the lower end of the range of group sizes. This is consistent with the fact that larger group sizes are more informative about sensitivity, and small group sizes are more informative about specificity. However, the new support points cannot approach the extremes of the range of allowable group sizes, because these points are already included in the design, and we need to observe results for sufficiently many distinct group sizes to be able to estimate the slope parameters α_1 and β_1.

We also considered how the population parameters for dilution effects influence the structure of the optimal designs. As the slope parameter α_1 becomes larger, x_2 and x_α tend to get smaller, while when the slope parameter β_1 becomes larger, x_β and x_2 tend to get larger. These changes

Table 2: D_s^I-optimal designs for several scenarios in Example 3. (Here $\alpha_0 = 2.6$ and $\beta_0 = 3.2$)

Model	α_1	β_1	x_1	x_β	x_2	x_α	x_3	w_1	w_β	w_2	w_α	w_3
(2.4)	–	–	1	–	10	–	81	.11	–	.55	–	.34
(4.3)	.0	–	1	–	11	53	150	.04	–	.27	.38	.31
	.3	–	1	–	7	44	150	.08	–	.30	.33	.29
	.5	–	1	–	6	38	150	.10	–	.31	.31	.28
(4.4)	–	.0	1	3	18	–	89	.07	.19	.44	–	.30
	–	.3	1	3	20	–	91	.11	.25	.38	–	.26
	–	.5	1	3	22	–	92	.15	.27	.34	–	.24
(4.5)	.0	.0	1	3	15	57	150	.05	.14	.29	.31	.21
	.0	.3	1	3	17	58	150	.09	.20	.29	.26	.16
	.0	.5	1	3	18	58	150	.13	.24	.27	.23	.13
	.3	.0	1	2	13	51	150	.08	.17	.25	.29	.21
	.3	.3	1	3	14	52	150	.09	.21	.27	.26	.17
	.3	.5	1	3	15	53	150	.12	.24	.25	.24	.15
	.5	.0	1	2	12	46	150	.08	.17	.24	.29	.22
	.5	.3	1	3	13	47	150	.08	.21	.26	.27	.18
	.5	.5	1	3	13	46	150	.12	.25	.25	.23	.15

may allow for improved estimation of the sensitivity or specificity curves, but note that since we are using the D_s-criterion focusing on prevalence, the changes are not large. □

In the example above, it seems that the upper bound x_U is always present in a D_s^I-optimal design when the sensitivity is diluted. However, x_U is not necessarily present, especially when x_U is sufficiently large. For instance, under the same parameter setting as the example above, with $\alpha_1 = \beta_1 = 0.5$, and moving x_U up to 1000, the D_s^I-optimal design is supported on $\{1, 3, 14, 51, 674\}$.

5. Design performance

In this section, we study the performance of the D_s^I-optimal design when the working parameter is moderately misspecified. We can see below that its performance is relatively stable when the working parameter is not too far from the true value. Following Examples 1-3, and focusing on the most common setting where only the sensitivity is diluted, we let $[x_L, x_U] \cap \mathbb{N} = \{1, 2, \ldots, 150\}$ and $q = 0.2$, and let the working parameter $\tilde{\theta}_0 = (\tilde{p}_0, \tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{p}_2)^{\mathrm{T}} = (0.07, 2.6, 0.3, 0.96)^{\mathrm{T}}$. From Table 2 (Model (4.3), $\alpha_1 = 0.3$), the D_s^I-optimal design $\tilde{\xi}$ under $\tilde{\theta}$ is supported on {1, 7, 44, 150}.

For studying how the misspecified working parameter affect the performance of $\tilde{\xi}$, we consider that the true value of $\theta = \{p_0, \alpha_0, \alpha_1, p_2\}^{\mathrm{T}}$ comes

Table 3: AEFF($\tilde{\xi}|\theta$) for selected $\theta \in \Theta$.

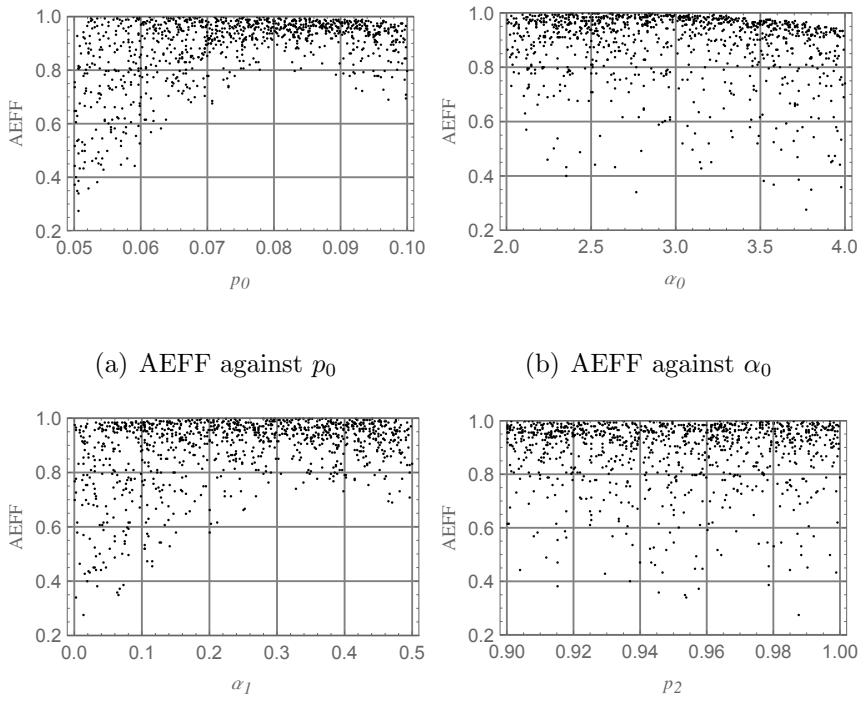
		$p_0 = 0.05$		$p_0 = 0.10$	
$p_1(x) = \text{expit}\{\alpha_0 - \alpha_1 \log(x)\}$		$p_2 = 0.9$	$p_2 = 1.0$	$p_2 = 0.9$	$p_2 = 1.0$
$\alpha_0 = 2$	$\alpha_1 = 0.0$	0.363	0.393	0.908	0.902
	$\alpha_1 = 0.1$	0.595	0.636	0.885	0.872
	$\alpha_1 = 0.5$	0.974	0.925	0.613	0.591
$\alpha_0 = 4$	$\alpha_1 = 0.0$	0.197	0.223	0.920	0.945
	$\alpha_1 = 0.1$	0.349	0.396	0.936	0.955
	$\alpha_1 = 0.5$	0.761	0.826	0.892	0.881

from $\Theta = [0.05, 0.1] \times [2, 4] \times [0, 0.5] \times [0.9, 1]$, which covers $\tilde{\theta}$. The performance of $\tilde{\xi}$ under the true value of $\theta \in \Theta$ is measured by

$$\text{AEFF}(\tilde{\xi}|\theta) = \text{AMSE}(\tilde{\xi}|\theta)/\text{AMSE}(\xi_\theta^I|\theta) \in [0, 1],$$

where ξ_θ^I is the D_s^I-optimal design under θ, and $\text{AMSE}(\xi|\theta) = M(\xi|\theta)_{11}^{-1}$ is the (scaled) asymptotic mean square error (AMSE) of the prevalence estimator under ξ, which is also its (scaled) asymptotic variance.

Table 3 shows AEFF($\xi|\theta$) for some selected $\theta \in \Theta$, and Figure 3 shows AEFF($\xi|\theta$) for some θ randomly drawn from Θ. Under this parameter setting, we observe that the accuracies of the prespecified p_0 and α_1 are important factors within this range of parameters. Figure 4 further shows



(a) AEFF against p_0 (b) AEFF against α_0

(c) AEFF against α_1 (d) AEFF against p_2

Figure 3: The AEFF($\tilde{\xi}|\theta$) under different θ for 1000 draws.

how the true values of p_0 and α_1 affect the performance of $\tilde{\xi}$. We observe that when $\tilde{p}_0$ and $\tilde{\alpha}_1$ are misspecified on the same direction, especially when they are both over-specified, AEFF drops rapidly. Roughly speaking, when the true value of $\theta \in \Theta$ falls between the two dashed lines on Figure 4, $\tilde{\xi}$ performs well with its AEFF close to or greater than 80%.

6. Conclusion and discussion

In this work, we develop efficient group testing designs that accommodate real-world complexities including differential subject and assay costs,

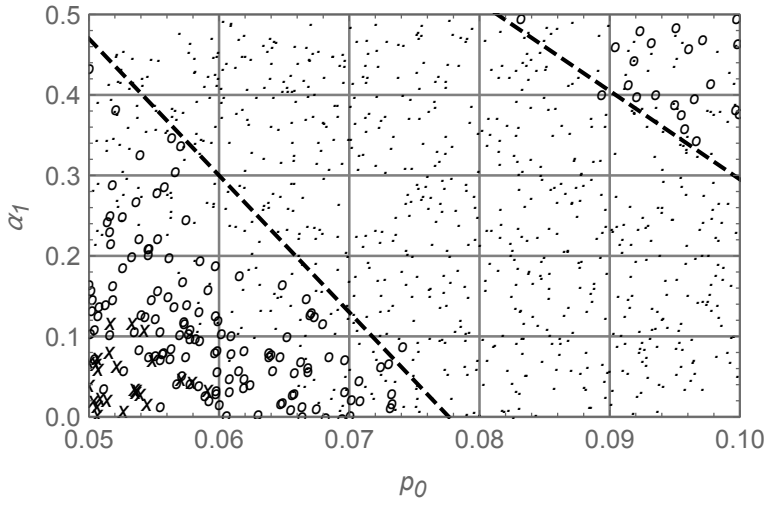


Figure 4: The AEFF($\tilde{\xi}|\theta$) vs. different true values of p_0 and α_1 ($\cdot$: AEFF $\geq$ 80%; $\circ$: AEFF $\in [50\%, 80\%)$; $\times$: AEFF $< 50\%$), where $\tilde{p}_0 = 0.07$ and $\tilde{\alpha}_1 = 0.3$.

and uncertain sensitivity and specificity that may have dilution effects. We characterize these designs and present an algorithm that is guaranteed to yield an optimal design on a discrete design space, as is encountered in practice. We found that accounting for subject costs yields designs with a smaller maximum group size compared to previously-published optimal designs in which the subjects were considered to be cost-free (Huang et al., 2017). Our results reveal that as the ratio of subject to assay costs increases even moderately, the largest group size of the resulting design and its proportion of trials drop rapidly, but its proportion of budget still

increases.

As a practical illustration, we provide examples addressing optimal allocation, with integer-valued trials at the optimal group sizes. Although the locally optimal designs depend on working parameters, our results based on a real-world setting show that the proposed designs are robust against misspecification of the working parameters and have good asymptotic efficiencies. When there are major concerns about possible misspecification of the working parameters, our optimal designs can be utilized with a multistage adaptive approach (Hughes-Oliver and Swallow, 1994). In the first stage, the working parameters may be specified using domain knowledge, and in subsequent stages, they are estimated from the previous stages. Alternatively, a Bayesian or minimax optimal design approach (Dette et al., 2014) can be adopted. These approaches seek designs either maximizing the D_s-optimality criterion (2.5) averaged over the parameters with respect to a prior distribution, or minimizing the largest possible variance of the prevalence estimator, respectively.

The most flexible model for group testing would allow the sensitivity and specificity to be estimated from the data, and potentially to vary with group size. However, the sensitivity and specificity parameters are nuisance parameters in practice, and are non-orthogonal to the primary parameter

of interest which is the prevalence. As a result, estimating these nuisance parameters increases the variance of the prevalence estimate, but eliminates any bias that would result from misspecifying them in a "plug-in" approach. The increase in variance is large for small numbers of trials, therefore it is unlikely to be favorable to estimate the sensitivity and specificity parameters in practice if the budget is small. However, if the budget is sufficiently large, the risk of bias due to misspecification dominates the increase in variance due to the additional parameter estimation. Our results therefore provide guidance to practitioners, suggesting that for smaller-scale research, a plug-in approach may be suitable, but researchers conducting larger studies should consider allowing the sensitivity and specificity parameters to be estimated from the data.

Increased interest in near real-time safety monitoring for disease epidemics, terror attacks, food safety, and environmental risks may provide new opportunities for group testing in the future. If cost considerations differ from the disease prevalence estimation that has dominated group testing to date, larger pools or larger total sample sizes may be practical, which could provide a setting where the marginal cost of estimating dilution effects along with prevalence would be modest. Our results may also be applied for evaluating the feasibility of such a procedure.

Acknowledgements

The first two authors were supported by the Ministry of Science and Technology, Taiwan, with grants no. MOST 105-2917-I-564-017 and MOST 103-2118-M-110-002-MY2, respectively. The third author was supported by the National Center of Theoretical Sciences, Taiwan.

Supplementary material

The supplementary material provides technical details for the proofs of the theorems and lemmas above. Some discussions on the D-optimal group testing designs under cost considerations are also provided.

References

Atkinson, A. C., Donev, A. N. and Tobias, R.D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press, Oxford.

Dette, H., Kiss, C., Benda, N. and Bretz, F. (2014). Optimal designs for dose finding studies with an active control. *J. Roy. Statist. Soc. Ser. B* **76**, 265–295.

Dorfman, R. (1943). The detection of defective members of large populations. *Ann. Math. Statist.* **14**, 436–440.

Gastwirth, J.L. (2000). The efficiency of pooling in the detection of rare mutations. *Amer. J. Hum. Genet.* **67**, 1036–1039.

Hughes-Oliver, J. M. and Swallow, W. H. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *J. Amer. Statist. Assoc.* **89**, 982–993.

Huang, S.-H., Huang, M.-N. L., Shedden, K. and Wong, W. K. (2017). Optimal group testing designs for estimating prevalence with uncertain testing errors. *J. Roy. Statist. Soc. Ser. B*, **79**, 1547–1563.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.* **2**, 849–879.

Liu, A., Liu, C., Zhang, Z. and Albert, P. S. (2012). Optimality of group testing in the presence of misclassification. *Biometrika* **99**, 245–251.

Liu, S.-C., Chiang, K.-S., Lin, C.-H., Chung, W.C., Lin, S.-H. and Yang, T.C. (2011). Cost analysis in choosing group size when group testing for Potato virus Y in the presence of classification errors. *Ann. Appl. Biol.* **159**, 491–502.

McMahan, C. S., Tebbs, J.M. and Bilder, C. R. (2012). Informative Dorfman Screening. *Biometrics* **68**, 287–296.

McMahan, C. S., Tebbs, J.M. and Bilder, C. R. (2013). Regression models for group testing data with pool dilution effects. *Biostatistics* **14**, 284–298.

Pilcher, C., Fiscus, S., Nguyen, T., Foust, E., Wolf, L., Williams, D., Ashby, R., O'Dowd, J., McPherson, J., Stalzer, B., Hightow, L., Miller, W., Eron, J., Cohen, M. and Leone, P. (2005). Detection of acute infections during HIV testing in North Carolina. *N. Engl. J. Med.* **352**, 1873–1883.

Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM, Philadelphia.

Turner, D.W., Stamey, J.D. and Young, D.M. (2009). Classic group testing with cost for grouping and testing. *Comput. Math. Appl.* **58**, 1930–1935.

Tu, X. M., Litvak, E., Pagano, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: Application to HIV screening. *Biometrika* **82**, 287–297.

Xie, M., Tatsuoka, K., Sacks, J. and Young, S. S. (2001). Group testing with blockers and synergism. *J. Amer. Statist. Assoc.* **96**, 92–102.

Zenios, S.A. and Wein, L. M. (1998). Pooled testing for HIV prevalence estimation: Exploiting the dilution effect. *Statist. Med.* **17**, 1447–1467.

Zhang, Z, Liu, C, Kim, S and Liu, A. (2014). Prevalence estimation subject to misclassification: the mis-substitution bias and some remedies. *Statist. Med.* **33**, 4482–4500.

Institute of Statistical Science, Academia Sinica, Nankang, Taipei 11529, Taiwan

E-mail: shhuang@stat.sinica.edu.tw

Department of Applied Math., National Sun Yat-sen University, Kaohsiung 80424, Taiwan

E-mail: lomn@math.nsysu.edu.tw

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA.

E-mail: kshedden@umich.edu