A Two-Stage Group Testing Model for Infections with Window Periods

Shaul K. Bar-Lev[*], Onno Boxma[†],
Wolfgang Stadje[‡] and Frank A. Van der Duyn Schouten[§]

Abstract

We present a two-stage group testing model for the detection of viruses in blood samples in the presence of random window periods. As usual, if a tested group is found to be positive, all its members are treated individually. The groups that were tested negative return for a second round after a certain time, new blood samples are taken and tested after pooling. The given system parameters are the size of the population to be screened, the incidence rates of the infections, the probability distributions of the lengths of the window periods, and the costs of group tests. The objective is to minimize the expected cost of running the system, which is composed of the cost of the conducted group tests and penalties on delayed test results and on misclassifications (noninfected persons declared to be positive and, more importantly, persons whose infections have not been identified). By an appropriate choice of the group size and the waiting time for the second round of testings one wants to optimize the various trade-offs involved. We derive in closed form all the probabilistic quantities occurring in the objective function and the constraints. Several numerical examples are given. The model is also extended to the case of several types of viruses with different window periods.

[*]Department of Statistics, University of Haifa, Haifa 31905, Israel (bar-lev@stat.haifa.ac.il)

[†]EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (boxma@win.tue.nl)

[‡]Department of Mathematics and Computer Science, University of Osnabrück, 49069 Osnabrück, Germany (wolfgang@mathematik.uos.de)

[§]Center for Economic Research, Tilburg University, 5000 LE Tilburg, The Netherlands (f.a.vdrduynschouten@uvt.nl)

1 Introduction

In this paper we consider a two-stage group testing model for the detection of viruses in blood samples in the presence of window periods.

Due to the high cost of advanced techniques like Nuclear Acid Testing (NAT), pooling methods have been frequently adopted when a large number of blood samples has to be screened for hepatitis B (HBV), hepatitis C (HCV), human immunodeficiency virus (HIV), or syphilis, for example in blood banks or in mass screenings.

A serious problem of testing for viral diseases is the presence of *window periods*, defined as the period elapsing from the time a person is infected by some virus until antibodies can be detected. Examples of average window periods for some viruses are: 22 days for HIV, 60 for HBV and 70 for HCV, but in individual cases window periods can be substantially longer.

In this paper we suggest and study the following model. Blood samples of a large number of individuals have to be tested for one or several viral diseases. Blood samples are taken, pooled in groups of equal size (which is a decision variable) and then tested. If a tested group is found to be positive, all persons in it are treated individually. In order to take into account the window periods, the other groups return for a second round of testing after a certain time (a second decision variable), new blood samples are taken and tested after pooling, using the same groups as in the first stage. The given system parameters are the size of the population to be tested, the incidence rates of the infections, the probability distributions of the (random) lengths of the window periods, and the costs of group tests. The objective is to minimize the expected cost of running the system, which is composed of the cost of the conducted group tests and penalties on delayed test results and on misclassifications (noninfected persons declared to be positive and, more importantly, persons whose infections have not been identified). By an appropriate choice of the group size and the waiting time for the second round of testings one wants to optimize the various trade-offs involved.

As a classical cost-efficient method to classify items from some finite population into different categories, group testing has been applied in various areas, first of all for blood testing to detect syphilis, HIV or other diseases [12, 6, 5, 9, 14, 16], but also in genetics [10, 11, 15], quality control for industrial production systems [13, 1], drug discovery [18], and communication networks [17]. A key reference is the monograph [4]. In [2] a more detailed discussion of the literature and a classification of group testing models according to various dichotomies are given. Several studies deal with false results within a grouping framework, e.g. [6, 14, 8, 3]. Two-stage Bayesian

procedures have been considered in [8, 7], proposing protocols that allow to detect false-negative results in the second stage that might have passed the first stage unnoticed.

The paper is organized as follows. In Section 2 we present the model in detail for the case of only one kind of virus. We formulate the optimization problem and derive in closed form all the probabilistic quantities occurring in the objective function and the constraints. In Section 3 we extend the model to the case of several types of viruses with different window periods. Section 4 is devoted to numerical examples.

2 Model I: single cause of contamination

2.1 Description and assumptions

For simplicity we call a single blood sample an item. We first consider the possibility of an item's infection by only one virus with a random window period; the case of infections by more than one virus will be discussed later. We make the following **assumptions**.

(i) *Individuals to be tested.* The population consists of N fresh items which are testable in groups of size m. We only consider group sizes m that divide N. This assumption avoids some computational and analytic complexity and causes only a negligible loss of generality in practical situations.

(ii) *Bernoulli assumption.* The expected proportion q of good items in the population is assumed to be known in advance (and will usually be close to 1). Every item is good (not infected) with probability q independently of the others. We set $p = 1 - q$.

(iii) *Test results.* For every group test there are two possible outcomes: 'clean', implying that no virus can be detected at the time of testing, or 'contaminated', implying that at least one item in the tested group has to be infected. Under this assumption, outcomes like 'false negative' or 'false positive' for tested groups are excluded.

(iv) *Window periods.* The window period of the virus is assumed to be a random variable taking values in the nonnegative integers (counting the time units until detectability), and having a known distribution. We denote by α_j, $j = 0, 1, 2, \ldots$, the probability that a given item has a window period of length j at the beginning of stage 1.

(v) *Objective.* We want to separate the population into clean and con-

taminated groups by means of pooled testing as accurately and quickly as possible by minimizing some cost function. The individuals belonging to groups found contaminated are immediately treated individually because of the suspected presence of viruses.

(vi) *The group testing procedure.* Every group is tested one or two times. One test is mandatory. A group which is found contaminated in stage 1 is not further group tested; the members of such a group are immediately called in for further individual testing to identify the infected persons. A group that is found clean at the first stage can still be contaminated as it may contain infected items which have not yet passed their window periods. Accordingly, in order to improve the quality of the testing procedure, the members of all groups declared noncontaminated at stage 1 will be called in after some time r (a decision variable) and have their blood again pooled (using the groups from the first stage) and tested. Only if found again to be clean, such a group will be finally declared good; otherwise the persons in this group will be treated individually. It is assumed that at each stage all groups are tested at the same time. The cost for testing a group of size m (in either stage) is $c(m)$; it may thus depend on the group size.

The model parameters

- N (population size),

- p (infection probability),

- $\{c(m)\}$ (costs of testing a group of size m),

- $(\alpha_j)_{j \in \mathbb{Z}_+}$ (probability distribution of the duration of the window period)

are assumed to be given. Our decision variables are the group size m and the waiting time r before conducting the second stage tests. Note that in the case of an unbounded window period distribution misclassifications of bad items cannot be avoided completely: however large r is chosen, the probability of not detecting a bad item will always be positive.

Clearly, a larger m leads to less expensive group testing but more wrong classifications, which may be costly. A larger r may lead to the detection of more bad items but increases the waiting times of items classified as good in both stages. An optimal selection of m and r will have to cope with these trade-offs and find the right balance.

Remark. The items transferred to the second stage of group testing are exactly the ones which are either not infected or infected with a positive remaining window period. They are interchangeable so that it does not matter

4

how the groups in the second stage are put together. One may also use a different group size for them and take it as a third decision variable. Since this extension would lead to more complicated formulas without adding crucial insight or requiring new methodology, we have made the assumption that the groups in the second stage will be the same as in the first stage (if transferred).

2.2 The underlying distributions

Let $l = N/m$ denote the number of groups, assuming without loss of generality that N/m is integer. Let

$$A_r = \sum_{j=r+1}^{\infty} \alpha_j; \qquad (2.1)$$

A_r is the probability that the window period is larger than r. Let Z_i be the number of clean items in group i, $i = 1, \dots, l$, that are not classified as good (because their group is found to be contaminated in stage 1 or stage 2) and let $Z = \sum_{i=1}^{l} Z_i$ be their total number in the set of all l groups. Similarly, let W_i be the number of bad items in group i that are finally (wrongly) classified as 'good' and let $W = \sum_{i=1}^{l} W_i$ be their total number. Knowing the distributions of these random variables is crucial for the selection of the decision parameters.

Theorem 1

$$\mathbb{P}(Z_i = k) = \binom{m}{k} q^k p^{m-k} \left(1 - A_r^{m-k} \right), \quad k = 1, \dots, m-1, \qquad (2.2)$$

$$\mathbb{P}(Z_i = 0) = [q + pA_r]^m + p^m \left(1 - A_r^m \right), \qquad (2.3)$$

$$\mathbb{P}_Z = (\mathbb{P}_{Z_1})^{*l}, \qquad (2.4)$$

$$\mathbb{P}(W_i = k) = \binom{m}{k} p^k q^{m-k} A_r^k, \quad k = 1, \dots, m, \qquad (2.5)$$

$$\mathbb{P}(W_i = 0) = 1 - [q + pA_r]^m + q^m, \qquad (2.6)$$

$$\mathbb{P}_W = (\mathbb{P}_{W_1})^{*l}, \qquad (2.7)$$

where $(\mathbb{P}_Y)^{*l}$ *denotes the lfold convolution of* $\mathbb{P}_Y$ *with itself.*

Proof. $Z_1, \dots, Z_l$ and $W_1, \dots, W_l$ are iid so that (2.4) and (2.7) are obvious. Next, $Z_i = 0$ means that one of the following two events occurs: either (i) the ith group passes both stages successfully, which is the case if and only if each of the m items is either good or has a window period larger than r

5

(probability $(q + pA_r)^m$); or (ii) the ith group is found to be contaminated and all its items are bad (probability $p^m[1 - A_r^m]$). This yields (2.3).

For the event $\{Z_i = k\}$, $k \in \{1, \ldots, m - 1\}$, to occur, there have to be exactly k clean items among the m items chosen for the ith group and the window period of at least one of the $m - k \geq 1$ bad items in the group must have passed before or at time r. This yields (2.2). For the event $\{W_i = k\}$, $k \in \{1, \ldots, m\}$, to occur, there have to be exactly k bad items among the m items in the ith group and all the corresponding k window periods are larger than r. Finally, $W_i = 0$ means that all items in the ith group are good or there are exactly k bad items for some $k \in \{1, \ldots, m\}$ of which at least one has a window period not exceeding r. Adding the corresponding probabilities we obtain (2.6) because

$$\mathbb{P}(W_i = 0) = 1 - \sum_{k=1}^{m} \mathbb{P}(W_i = k)$$
$$= 1 + q^m - \sum_{k=0}^{m} \binom{m}{k} p^k q^{m-k} A_r^k$$
$$= 1 + q^m - [q + pA_r]^m.$$

■

Remark. Z_i cannot take the value m because a group consisting only of good items will never be classified as contaminated. Indeed, summing (2.2)-(2.3) over $k = 0, \ldots, m - 1$ easily yields $\sum_{k=0}^{m-1} \mathbb{P}(Z_i = k) = 1$. On the other hand, W_i takes every value $0, \ldots, m$ with positive probability.

Some related important probabilities are given in the next theorem.

Theorem 2 *Define the events*

B: a given group is not declared contaminated in stage 1;

C: all m items in a given group are clean;

D: a given group passes successfully the two consecutive stages.

Then we have

(i)
$$P(C) = q^m, \tag{2.8}$$

(ii)
$$\mathbb{P}(B) = [q + p(1 - \alpha_0)]^m, \tag{2.9}$$

(iii)
$$\mathbb{P}(D) = [q + pA_r]^m, \tag{2.10}$$

(iv)

$$\mathbb{P}(D \mid B) = \left(\frac{q + pA_r}{q + p(1 - \alpha_0)} \right)^m, \qquad (2.11)$$

(v)

$$\mathbb{P}(C \mid B) = \left(\frac{q}{q + p(1 - \alpha_0)} \right)^m, \qquad (2.12)$$

(vi)

$$\mathbb{P}(W = 0) = (1 - [q + pA_r]^m + q^m)^l. \qquad (2.13)$$

Proof. (i) - (iii) follow immediately from the independence assumptions, and (iv) - (v) are elementary conditional probabilities derived from (i) - (iii). (vi) is a special case of (2.7) combined with (2.6). ∎

2.3   The objective function

We now turn to the formulation of the cost minimization problem. Recall (cf. Section 2.1) that the model parameters N, p, $\{c(m)\}$ and $\{\alpha_j\}$ are assumed to be given, and that our decision variables are the group size m and the waiting time r before conducting the second stage tests. The following costs and rewards can be considered:

(a) After the testing, a random number Y of groups has successfully passed the two consecutive stages. Clearly, $Y \sim B(l, \mathbb{P}(D))$, where $\mathbb{P}(D)$ is given in (2.10). Let $\rho(r)$ be the penalty per item due to a delay of r time units before getting the test result; $\rho(r)$ is assumed to be some nondecreasing function. Then the total penalty caused by delays is $\rho(r)mY$.

(b) The total cost of testing is the sum of the cost of the l tests in stage 1 and of the random number of tests in stage 2 and is thus given by $c(m)[l + (l - Y_1)]$, where Y_1 is the number of groups which do not reach stage 2. Clearly, $Y_1 \sim B(l, \mathbb{P}(B))$.

(c) For each of the N items to be tested some fee $a > 0$ will be collected.

(d) For every good item which belongs to some group that is found contaminated (so that individual treatment is required) we introduce a penalty π. The total number of these items has been denoted by Z so that this type of misclassification leads to a total cost of πZ.

(e) A crucial cost is due to the bad items that are declared good; such a wrong decision can have disastrous consequences. When this is the case a constraint on $\mathbb{P}(W > 0)$ seems appropriate. In any case we can put a high penalty $b > 0$ on each misclassified bad item.

The total net reward is the difference of the revenue and the costs and thus given by

$$aN - \rho(r)mY - c(m)[2l - Y_1] - \pi Z - bW.$$

Since the revenue aN is considered to be fixed independently of the design variables m and r, we only have to take the total cost into account. We want to minimize its expected value, which can be written as

$$R(m,r) = \rho(r)m\mathbb{E}(Y) + c(m)[2l - \mathbb{E}(Y_1)] + \pi\mathbb{E}(Z) + b\mathbb{E}(W). \qquad (2.14)$$

Note that

$$\mathbb{E}(Y) = l\mathbb{P}(D) = \frac{N}{m}\left[q + pA_r\right]^m, \qquad (2.15)$$

$$\mathbb{E}(Y_1) = l\mathbb{P}(B) = \frac{N}{m}\left[q + p(1 - \alpha_0)\right]^m \qquad (2.16)$$

and, from Theorem 1,

$$\mathbb{E}(Z) = Nq\left[1 - \{q + pA_r\}^{m-1}\right], \qquad (2.17)$$

$$\mathbb{E}(W) = NpA_r\left[q + pA_r\right]^{m-1}, \qquad (2.18)$$

so that all the terms in (2.14) can be easily computed.

Due to the particular importance of avoiding misclassifications of bad items, i.e., overlooking infections, we will put a constraint on W. If one wants the total number of wrongly classified bad items to be very small with a high reliability, a natural constraint is $\mathbb{P}(W = 0) \geq 1 - \gamma$, where $\gamma \in (0, 1)$ is a preassigned level, which in practice should of course be close to 0. Another possibility is a constraint on the expected value of W, say $\mathbb{E}(W) \leq w$ for some small threshold value w.

Our general optimization problem can now be formulated in the following form:

$$\begin{cases} \textbf{Cost minimization:} \\ \min_{m,r} R(m,r) \\ subject\ to \\ \mathbb{P}(W = 0) \geq 1 - \gamma. \end{cases} \qquad (2.19)$$

Alternatively, the last constraint may be replaced by

$$\mathbb{E}(W) \leq w.$$

Note that also $\mathbb{P}(W = 0)$ is given explicitly in Theorem 2 (for $\mathbb{E}(W)$ see (2.18)). Therefore, we are left with numerical deterministic minimization problems.

If $\rho(m)$ and $c(m)$ are simple sequences, the cost function $R(m, r)$ is a linear combination of elementary functions. For example, taking $\rho(r) = ar$ and $c(m) = \gamma m^\eta$ (as in Section 4) with positive constants a, γ and η we obtain

$$R(m, r) = N\Big[ar(q + pA_r)^m + \gamma m^{\eta-1}[2 - (q + (1 - \alpha_0)p)^m]$$
$$+ \pi q\left[1 - (q + pA_r)^{m-1}\right] + bpA_r(q + pA_r)^{m-1}\Big].$$

One can try to first find, for fixed r, the minimum with respect to m (treating m as a continuous variable) and then, in a second stage, to minimize with respect to r, at least in the case of a geometric window size distribution (considered in our examples below). However, setting the derivative with respect to m equal to zero results in a transcendental equation for m and is thus analytically intractable (and also note that the constraint has not yet been taken into account). A numerical approach seems unavoidable even under the simplest assumptions.

3 The case of several viruses with different window periods

Our model can be generalized to the case of several viruses as possible causes of contamination. We develop here the case of two viruses.

Denote by $q_{11}, q_{12}, q_{21}, q_{22}$ the probabilities that the item contains both viruses or only virus 1 or only virus 2 or none of them, respectively. Let α_i (β_j) be the probability that a given item has a window period of length i (j), where $i, j \in \{0, 1, 2, \ldots\}$. We suppose that in case both viruses are present in the same item the durations of the two window periods are independent random variables. The other model assumptions are exactly as in Section 2, and we want to minimize the same objective function (given by (2.14)). Hence, we have to determine the probabilities $\mathbb{P}(B)$ and $\mathbb{P}(D)$ and the distributions $\mathbb{P}_Z = (\mathbb{P}_{Z_1})^{*l}$ and $\mathbb{P}_W = (\mathbb{P}_{W_1})^{*l}$ in the two-virus case.

Theorem 3 *Let $A_r = \sum_{i=r+1}^{\infty} \alpha_i$, $B_r = \sum_{i=r+1}^{\infty} \beta_i$. In the two-virus model*

we have

$$\mathbb{P}(B) = [q_{22} + q_{11}(1 - \alpha_0)(1 - \beta_0) + q_{12}(1 - \alpha_0) + q_{21}(1 - \beta_0)]^m,$$

$$(3.1)$$

$$\mathbb{P}(D) = [q_{22} + q_{11}A_rB_r + q_{12}A_r + q_{21}B_r]^m, \qquad (3.2)$$

$$\mathbb{P}(Z_i = k) = \sum_{k_{11},k_{12},k_{21} \geq 0: k_{11}+k_{12}+k_{21}=m-k} \binom{m}{k_{11}, k_{12}, k_{21}, k} q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{k_{21}} q_{22}^{k}$$
$$\times \left(1 - A_r^{k_{11}+k_{12}} B_r^{k_{11}+k_{21}}\right),$$
$$k = 1, \ldots, m-1$$
$$(3.3)$$

$$\mathbb{P}(Z_i = 0)) = \left[q_{22} + q_{11}A_rB_r + q_{12}A_r + q_{21}B_r\right]^m$$
$$+ \sum_{k_{11},k_{12} \geq 0: k_{11}+k_{12} \leq m} \binom{m}{k_{11}, k_{12}, m - k_{11} - k_{12}} q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{m-k_{11}-k_{12}}$$
$$\times \left(1 - A_r^{k_{11}+k_{12}} B_r^{m-k_{12}}\right), \qquad (3.4)$$

$$\mathbb{P}(W_i = k) = \sum_{k_{11},k_{12},k_{21} \geq 0: k_{11}+k_{12}+k_{21}=k} \binom{m}{k_{11}, k_{12}, k_{21}, m-k} q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{k_{21}} q_{22}^{m-k}$$
$$\times A_r^{k_{11}+k_{12}} B_r^{k_{11}+k_{21}}, \quad k = 1, \ldots, m \qquad (3.5)$$

$$\mathbb{P}(W_i = 0) = q_{22}^m + \sum_{k=1}^{m} \sum_{k_{11},k_{12},k_{21} \geq 0: k_{11}+k_{12}+k_{21}=k} \binom{m}{k_{11}, k_{12}, k_{21}, m-k}$$
$$\times q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{k_{21}} q_{22}^{m-k}[1 - A_r^{k_{11}+k_{12}} B_r^{k_{11}+k_{21}}]. \qquad (3.6)$$

Proof. A given group is not declared to be contaminated in stage 1 if and only if for each of its m items the following holds: it is either good (probability q_{22}) or contains at least one virus with positive window period (probability $q_{11}(1 - \alpha_0)(1 - \beta_0) + q_{12}(1 - \alpha_0) + q_{21}(1 - \beta_0)$). This proves (3.1). A given group passes successfully both stages if and only if each of its m items is either good or contains viruses whose window periods are all larger than r (probability $q_{11}A_rB_r + q_{12}A_r + q_{21}B_r$). This argument leads to (3.2).

The event $Z_i = 0$ occurs if and only if the ith group has the following property: either it passes both stages successfully, i.e., each of the viruses present has a window period larger than r, or the group is found to be contaminated and all its items are bad. The first case has probability $\mathbb{P}(D)$. The second case is equivalent to the existence of nonnegative integers k_{11}, k_{12}, k_{21} satisfying $k_{11} + k_{12} + k_{21} = m$ such that the following holds:

(a) k_{11} items carry both viruses, k_{12} items carry only virus 1, k_{21} items carry

10

only virus 2;

(b) at least one of these viruses has a window period of at most r time units.

For $k \in \{1, \ldots, m-1\}$ the event $Z_i = k$ occurs if and only if the ith group has the following property: there are nonnegative integers k_{11}, k_{12}, k_{21} such that $k_{11} + k_{12} + k_{21} = m - k$ such that (a) and (b) above hold and additionally

(c) the remaining k items are good.

The probability that among $k_{11} + k_{12} + k_{21}$ bad items as in (a) no virus has a window period of at most r time units is equal to

$$(A_r B_r)^{k_{11}} A_r^{k_{12}} B_r^{k_{21}},$$

and (3.3) and (3.4) follow easily (for (3.4) one has to use $k_{21} = m - k_{11} - k_{12}$).

For $k \in \{1, \ldots, m\}$ the event $W_i = k$ occurs if and only if the underlying group has the following property: there are nonnegative integers k_{11}, k_{12}, k_{21} such that $k_{11} + k_{12} + k_{21} = k$ and

(i) k_{11} items carry both viruses, k_{12} items carry only virus 1, k_{21} items carry only virus 2;

(ii) none of the viruses involved has a window period of at most r time units;

(iii) the remaining $m - k$ items are good.

The probability that among $k_{11} + k_{12} + k_{21}$ bad items as in (i) none has a window period of at most r time units is equal to

$$(A_r B_r)^{k_{11}} A_r^{k_{12}} B_r^{k_{21}},$$

and we obtain (3.5).

Finally, $W_i = 0$ means that in the group under consideration either (a) each item is good or (b) for some $k \in \{1, \ldots, m\}$ and some nonnegative integers k_{11}, k_{12}, k_{21} summing to k the group contains $m - k$ good items, k_{11} items carry both viruses, k_{12} items carry only virus 1, k_{21} items carry only virus 2, and at least one of the viruses present has a window period of at most r. Writing this decomposition in terms of probabilities yields (3.6). ∎

Remark. As can be seen in the proof, the assumption of independent window periods for the viruses in the same item is not necessary; we can easily write down the probabilities in the second lines of (3.3) and (3.5) in the case of dependent window periods as well.

Using Theorem 3 we can again express all probabilities and expected values in the cost function (2.14) in terms of the system parameters and the deci-

sion variables and find its minimum and the minimizing values of m and r numerically.

4 Numerical examples

We present a few numerical examples for the minimization of the cost function in the one-virus case studied in Section 2.

Example 1. We fix the model parameters introduced in Section 2.1 as follows:

$$N = 100, p = 0.01, \alpha_r = (0.95)^r 0.05, c(m) = 0.07 \ m^{1/2}.$$

Note that $A_r = (0.95)^{r+1}$. The cost parameters in Section 2.3 are selected as follows:

$$\rho(r) = r/10, \pi = 1, b = 10, \gamma = 0.01.$$

Figure 1 shows a plot of the cost $R(m,r)$ as a function of r for fixed values of m (divisors of 100). The lowest curve belongs to $m = 4$, the highest to m=50. We put the constraint $\mathbb{P}(W = 0) \geq 0.99$. Then the waiting period has to exceed a certain threshold (depending on the group size) to give an admissible procedure. The admissible points are displayed in bold face (at the right side of the graphs). The optimal admissible solution is $m = 25$ and $r = 85$. It is seen that the cost function $r \mapsto R(m,r)$ for $m = 25$ is considerably larger than for smaller group sizes in the inadmissible regions, but beats the other cost functions slightly in the admissible region.

Example 2. Take $N = 10,000$, $\gamma = 0.09$ and the other parameters as in Example 1. The solution for the unconstrained problem is $m = 125, r = 39$, while the constrained minimum is attained for $m = 250, r = 87$. Hence, to obtain the constrained minimum one has to double the group size and more than double the waiting period.

Example 3. Now we take

$$N = 10,000, p = 0.01, \alpha_r = (0.95)^r 0.05, c(m) = 0.2 \ m^{3/2}$$

$$\rho(r) = r/10, \pi = 1, b = 10$$

and consider the minimization problem under the constraint $\mathbb{E}(W) \leq w$. The results are illustrated by three-dimensional plots.

(i) $\mathbb{E}(W) \leq 6$: Figure 2 displays the objective function $R(m,r)$, where only values of m dividing N are considered. The solution of the unconstrained
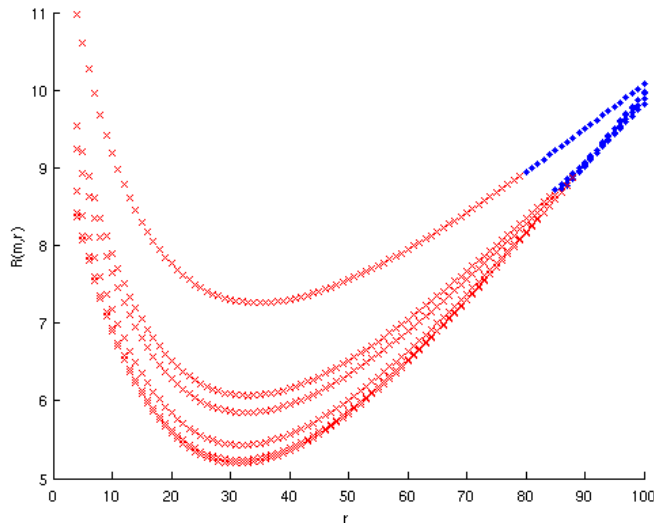
Figure 1: Objective function against r for fixed values of m

minimization problem is $m = 125, r = 39$. Since this pair is admissible, it is also optimal for the constrained problem.

(ii) $\mathbb{E}(W) \leq 1$: Figure 3 shows the objective function on the admissible region. The unconstrained optimum is of course $m = 125, r = 39$ as in (i) but this pair is now, under the sharper constraint $\mathbb{E}(W) \leq 1$, no longer admissible. The solution under this constraint is $m = 200, r = 53$. It is interesting that a sharper constraint can lead to a larger optimal group size. The increase in the length of the waiting period from 39 to 53 seems intuitive as one wants to avoid misclassifications of bad items more forcefully.

Example 4. Let

$$N = 10,000, p = 0.01, \pi = 1, b = 10, \alpha_r = (0.95)^r 0.05, \gamma = 0.09, c(m) = 0.04m^2.$$

The global minimum of the cost function is attained at $m = 1, r = 35$, while over the admissible region the values $m = 25, r = 64$ are optimal. In this example the cost of a group test grows quadratically in the group size so that without a constraint it is optimal to choose $m = 1$, i.e., not to form groups at all. However, when the constraint $\mathbb{P}(W = 0) \geq 0.91$ is introduced, the waiting period has to be extended drastically (from 35 to 64) and the group size 25 becomes optimal.

Remarks. (1) There are several choices possible for the window period distribution. However, since the minimization has to be carried out numerically, it seems difficult to get insight into the effect of the shape of this

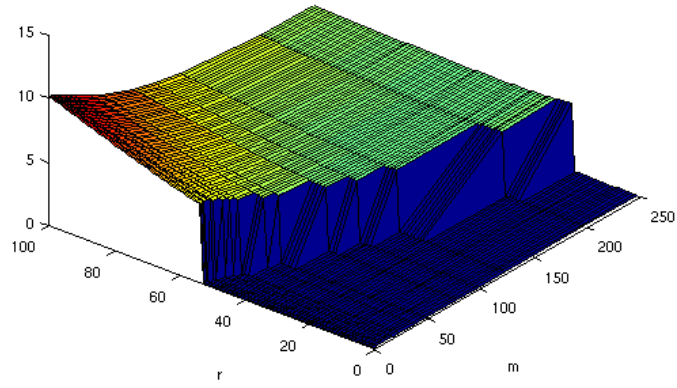Plot of the objective function in the admissible region.

Figure 2: Objective function in admissible area for $\mathbb{E}(W) \leq 6$

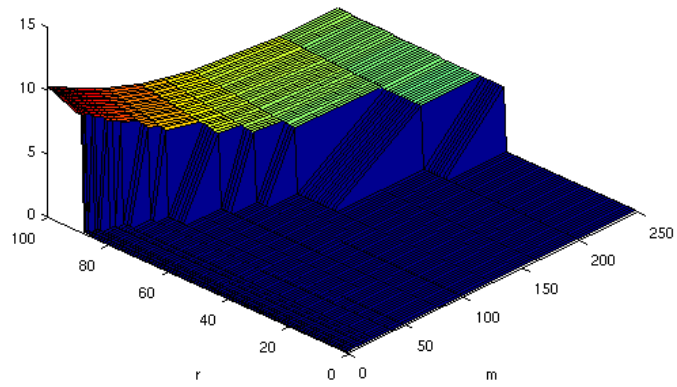Plot of the objective function in the admissible region.

Figure 3: Objective function in admissible area for $\mathbb{E}(W) \leq 1$

distribution. The model is based on truncating after r time units; therefore we conjecture that even for distributions with infinite mean the results would not be much different than for those having finite mean. We intend to pursue this point further.

(2) It may seem surprising that a sharper constraint can lead to a larger optimal group size (as in Example 4). It should be noted, though, that $\mathbb{P}(W = 0)$ does not always increase when the group size m decreases. As a simple example, take $N = 2$ and compare $\mathbb{P}(W = 0) = (1 - pA_r)^2$ (for $m = 1$) with $\mathbb{P}(W = 0) = 1 - (q + pA_r)^2 + q^2$ (for $m = 2$). It is easily seen that the latter probability is larger. Anyway, we cannot just consider *one* variable (m) in our constrained optimization problem. Both the constraint $\mathbb{P}(W = 0) \geq 1 - \gamma$ and the cost function $R(m, r)$ (representing various trade-offs with respect to m and r) depend on both m and r in a rather intricate way. In Example 4 the constraint is indeed violated for $(m, r) = (1, 35)$, so we must change m and/or r. If we just change r to satisfy the constraint, it affects $R(m, r)$, leading also to a different m. We finally end up with a quite different optimal (m, r): (25,64) instead of (1,35). Intuitively, if we have to increase the window size to achieve the desired reliability, we may thereafter have some freedom in choosing the group size without violating the constraint, and we then take m so as to obtain a small expected cost of testing, i.e., as large as it is admissible.

References

[1] Bar-Lev, S.K., Boneh, A. and Perry, D. (1990) Incomplete identification models for group-testable items. *Naval Research Logistics* **37**, 647-659.

[2] Bar-Lev, S.K., Stadje, W. and Van der Duyn Schouten, F.A. (2004) Multinomial group testing models with incomplete identification. *Journal of Statistical Planning and Inference* **135**, 384-401.

[3] Bar-Lev, S.K., Stadje, W. and Van der Duyn Schouten, F.A. (2006) Group testing procedures with incomplete identification and unreliable testing results. *Applied Stochastic Models in Business and Industry* **22**, 281-296.

[4] Du, D.-Z., and Hwang, F.K. (2000) *Combinatorial Group Testing and Its Applications*. 2nd. ed., World Scientific, Singapore.

[5] Gastwirth, J.L., and Johnson, W.O. (1994) Screening with cost-effective quality control: potential applications to HIV and drug testing. *Journal of the American Statistical Association* **89**, 972-981.

[6] Hammick, P.A. and Gastwirth, J.L. (1994) Group testing for sensitive characteristics: extension to higher prevalence levels. *International Statististical Review* **62**, 319-331.

[7] Hanson, T.E., Johnson, W.O. and Gastwirth, J.L. (2006) Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening. *Biostatistics* **7**, 41-57.

[8] Johnson, W.O. and Gastwirth, J.L. (2000) Dual group screening. *Journal of Statistical Planning and Inference* **83**, 449-473.

[9] Litvak, E., Tu, X.M. and Pagano, M. (1994) Screening for the presence of a disease by pooling sera samples. *Journal of the American Statistical Association* **89**, 424-434.

[10] Macula, A.J. (1999) Probabilistic nonadaptive group testing in the presence of errors and DNA library screening. *Annals of Combinatorics* **3**, 61-69.

[11] Macula, A.J. (1999) Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening. *Journal of Combinatorial Optimization* **2**, 385-397.

[12] Monzon, O.T., Paladin, F.J.E., Dimaandal, E., Balis, A.M., Samson, C., Mitchell, S. (1991) Relevance of antibody content and test format in HIV testing of pooled sera. *AIDS* **6**, 43-47.

[13] Sobel, M. and Groll, P.A. (1959) Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Technical Journal* **28**, 1179-1252.

[14] Tu, X.M., Litvak, E. and 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.

[15] Uhl, G., Liu, Q., Walther, D., Hess, J. and Naiman, D. (2001) Polysubstance abuse-vulnerability genes: genome scans for association using 1,004 subjects and 1,494 single-nucleotide polymorphisms. *American Journal of Human Genetics* **69** 1290-1300.

[16] Wein, L.M. and Zenios, S.A. (1996) Pooled testing for HIV screening: capturing the dilution effect. *Operations Research* **44**, 543-569.

[17] Wolf, J. (1985) Born again group testing: multiaccess communications. *IEEE Transactions on Information Theory* **IT31**, 185-191.

[18] Zhu, L., Hughes-Oliver, J. and Young, S. (2001) Statistical decoding of potent pools based on chemical structure. *Biometrics* **57**, 922-930.