EURANDOM PREPRINT SERIES

2014-020

December 18, 2014

Recycled Incomplete Identification Procedures for Blood Screening

S. Bar-Lev, O. Boxma, I. Kleiner, D. Perry, W. Stadje
ISSN 1389-2355

Recycled Incomplete Identification Procedures for Blood Screening

Shaul K. Bar-Lev[*], Onno Boxma[†], Igor Kleiner[‡], David Perry[§], Wolfgang Stadje[¶]

December 18, 2014

Abstract

The operation of blood bank systems is characterized by two crucial factors: testing procedures and perishability. We propose a new testing procedure that we term Recycled Incomplete Identification Procedure (RIIP). In RIIP, groups of pooled blood units which are found contaminated in a so-called ELISA test are divided into smaller subgroups and again group-tested by ELISA, and so forth, until finally a so-called PCR test is conducted for those subgroups which are found clean. We analyze and optimize the performance of RIIP, maximizing the profit associated with the procedure. Our numerical results suggest that it may indeed be profitable to do several cycles at ELISA.

1 Introduction

The operation of blood banks, worldwide, is aimed at the supply of uncontaminated human blood. In the laboratories of the Central Blood Services (CBS's), each donated blood unit goes through multiple tests. These are aimed at determining the unit's blood type and the presence of various pathogens which are able to cause transfusion-transmitted diseases, such as Hepatitis B (HBV), Hepatitis C (HCV), Human Immunodeficiency Virus (HIV) and Syphilis; see, e.g., [16], [19], [20], [24], [25], [27], [28]. Until recently, the routine testing was done with the ELISA (Enzyme Linked Immuno-Sorbent Assay) test that detects virus-specific antibodies in the blood. This test has high sensitivity and specificity but has a lower analytic detection limit, which affects the identification of positive blood samples very soon after HIV seroconversion, as it takes time to develop a high concentration of antibodies. The new PCR (Polymerase Chain Reaction) test can detect viral genetic material in the blood and has a much higher sensitivity and specificity. It increases the chances of early detection and decreases morbidity and mortality due to post-transfusion infections (see, e.g., [16], [25], [28]). However, PCR is very expensive relative to ELISA. Therefore, blood banks in the USA, Israel and some countries in Europe have established a screening protocol whereby all blood units are ELISA tested in groups and those which tested negative for ELISA are re-tested individually with PCR.

The operation of blood bank systems is characterized by two crucial factors: (i) testing procedures and (ii) perishability. Testing is necessary, since only clean blood units are used for blood transfusion;

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

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

[‡]Department of Statistics, University of Haifa, Haifa 31905 Israel (igkleiner@gmail.com)

[§]Department of Statistics, University of Haifa, Haifa 31905 Israel (dperry@stat.haifa.ac.il)

[¶]Institute of Mathematics, University of Osnabrück, Osnabrück, Germany (wstadje@uos.de)

groups found contaminated at ELISA, and units found contaminated at PCR, are discarded. Thousands of units of donated blood arrive daily at the central blood bank system for screening. Testing times and testing costs are associated with the screening process. In addition, each blood unit has an expiration date; after that it is perished. The controller faces a natural and well-motivated operations management problem. On the one hand, there is the need to make the testing period as short as possible; on the other hand, careful testing is required, which takes time and is costly. This raises the need to find more efficient group testing procedures, with the restriction of incomplete identification.

Since ELISA testing is relatively cheap, we propose a new screening process that we term Recycled Incomplete Identification group testing Procedure (RIIP), by which groups of pooled blood units, which are found contaminated in the previous ELISA cycle, are divided into smaller subgroups and again group-tested by ELISA, and so forth, until finally the PCR test is conducted for those subgroups which are found clean.

The goal of this paper is to provide an analysis and optimization of the performance of RIIP, in particular minimizing the costs (or maximizing the profit) associated with the test procedure.

In the next subsections we describe some features of a blood bank system. In Subsection 1.1 we describe the separation process of each blood unit into its three components along with their expiration dates and associated costs. The concept of group testing is reviewed in Subsection 1.2. In Subsection 1.3 we present the group testing procedures (complete and incomplete identification procedures) that are currently in use by blood banks. The further organization of the paper is outlined in Subsection 1.4.

1.1 Blood Components

Blood consists of several components: Red blood cells (RBC), plasma and platelets. Processing the whole blood units into the different components is done in parallel to the testing stage. Whole blood units are separated into different components, which have different biological functions, storage conditions and expiration dates. They will be supplied to different patients according to their medical needs:

1. *RBC* – This component, which is separated from the whole blood unit within $8-24$ hours from collection, can be used within 35 to 42 days, depending on whether additive solution is added. Mostly, the cost of an unmodified packed RBC unit for the hospitals is \$40 and that of a leukodepleted (solution added) unit is \$70.

2. *Plasma* – Plasma units are automatically made upon the production of an RBC unit. The corresponding cost for a plasma unit is around \$40; hospitals acquire 28% of all the plasma units produced.

3. *Platelets* – From each whole blood unit one random platelet unit is separated, which can be used for at most five days. The cost of producing random platelet units is about \$40/unit [7].

With respect to the above expiration dates, the following should be taken into account. On average it takes about 15 hours till blood samples arrive in the CBS after the moment they have been donated. The average processing time of an ELISA test is around 1 hour and costs around \$2.5 per average group (of blood units) of size 10, whereas that of the PCR is around 6 hours with an associated cost of about \$85 per blood unit. Such time constraints are vital for platelets' shelf-life, but less significant for RBC. However, such processing times should be taken into account in any blood screening procedure.

1.2 Group Testing

The issue of blood transfusion might be a question of life and death. This means that it is of paramount importance that all the blood units that enter the shelf are clean. Therefore, a necessary requirement is a meticulous inspection of all the blood units. However, since thousands of blood units (blood samples) arrive at the central blood bank every day, a natural screening procedure must be based on the idea of group testing; otherwise, the screening process will take too long and the costs of this process will be too high.

Group testing deals with the classification of the items of some population into two categories: 'good' and 'defective'. It is assumed that the items are group testable, i.e., for any subset of the population it is possible to carry out a simultaneous test (group test) with two possible outcomes: 'success' (also called 'clean', or 'negative'), indicating that all items in the subset are good, and 'failure' (also called 'contaminated', or 'positive'), indicating that at least one of the items in the subset is defective – without knowing which or how many are defective. A contaminated group can be subject to further screening, or be scrapped. Employing suitably designed procedures of this kind leads to a significant reduction of the number of required tests and thus of screening costs, under controlled probabilities of misclassification. A group testing procedure is therefore a cost-efficient technique. It has been applied in various areas, first and foremost, for blood screening to detect various viruses, for DNA screening, as well as for quality control for industrial production systems (see, e.g., [3]-[18], [21]-[23], [30]-[36]).

1.3 Complete versus Incomplete Identification

One may currently distinguish two types of identification testing procedures: complete and incomplete. The purpose of a *complete identification group testing procedure* is to classify each item in a given population as either clean or defective. This is done by testing groups of size m (a decision parameter) in the ELISA station only. If a group is found clean it is aggregated for blood transfusion purposes, otherwise, if it is found contaminated it is further re-tested by dividing it into subgroups. Such a procedure continues till each item in a given population is appropriately classified.

One could imagine two managerial reasons in which *complete identification procedures* are inefficient. The first one is that the tests are too expensive. Then a complete identification procedure leads to a high expected number of tests and, as a result, to high expected total testing costs. The second one is that the shelf-life is short (recall that the shelf-life of platelets is at most 5 days). Then the higher the number of tests, the shorter the residual shelf-life of clean items on the shelf.

The idea of an *Incomplete Identification group testing Procedure* (IIP) was first introduced in [3] for some industrial problem and was subsequently further developed for blood screening (e.g. [4]-[10]). An IIP starts as above by first testing groups of size m in the ELISA station. However, as opposed to the previous case, a group found contaminated is scrapped; otherwise, it is aggregated and sent to the PCR station for individual testing. Such a procedure is cost-wise rather efficient as it significantly decreases the number of tests (whether grouped or individual). It is particularly efficient when the prevalence rate of the "deficiency" (like the prevalence rate of, say, HIV in the population) is rather small.

In order to give some idea how real data are processed we mention the following. In Western European countries (as well as in the US, Israel and some other countries) about $35,000$ to $50,000$ blood donations are needed per 1 million persons per year. The following data for the years $2011-2012$ have been provided to us by the Israeli Central Blood Bank. The data describe per year the number of blood donations (blood units), the number of blood units found contaminated at the ELISA station (including HIV, HBV and HCV) and the number of units found clean at the ELISA station but then found contaminated at the PCR station (for the two years 2011 and 2012 the population size of Israel

was about $7,800,000$ and $7,900,000$, respectively).

Donations	**Year**	Cases confirmed	positive	by **ELISA**	
		HBV	HCV	HIV	**Total**
$294,117$	2011	126	62	13	201
$298,470$	2012	143	62	3	208

The table suggests that the estimated probability of finding a given unit to be contaminated at the ELISA station is approximately 0.00068 (later on such a probability will be denoted by ε). Note that this estimated probability is smaller than the prevalence rate of the contaminating virus(es) in the population as those infected persons who are aware of their situation usually do not donate blood samples. This means that if the probability of contaminated blood units is not known, for some reason, then the prevalence rate in the population can be used as an upper bound for such a probability. All of the positive ELISA units were also judged positive by PCR, but, in addition, there were another 12 units in 2011 and 13 in 2012 that were only found positive by PCR (but not by ELISA). Hence, the estimated probability of those units found clean by ELISA but then found contaminated by PCR is 0.00004 for both years 2011 and 2012 (later this probability will be denoted by γ).

1.4 Organization of the Paper

The paper is organized as follows. In Section 2 we describe the RIIP in more detail, including costs and times involved as well as related stopping times. We also define an appropriate objective function, aiming at maximizing the profit of the RIIP operation. Sections 3 and 4 are devoted to the derivation of the distributional behavior of the relevant stopping times and all the functionals occurring in the objective function. We introduce an underlying *unobservable* nonhomogeneous Markov chain, whose distribution will be determined in closed form, and show that all distributions of interest and thus the complete objective function can be expressed in terms of this Markov chain. It will turn out that the resulting formulas for these distributions are rather complicated, making the objective function quite intricate and formidable to handle analytically. Therefore we devote Section 5 to introducing approximations of the relevant functionals. These approximations ease the numerical evaluation of the objective function. Numerical examples are presented in Section 6, along with some sensitivity analysis of the involved parameters and decision variables. Section 7 contains conclusions.

2 The RIIP Model: Description, Stopping Times, Parameters and Objective Function

In the RIIP, blood samples are tested at two consecutive stations: first the ELISA station and subsequently the PCR station. In the sequel, we consider testing on a weekly basis. An initial supply of n blood units per week (according to the data in Subsection 1.3, n would be around 6000) is divided into k_1 groups of size m_1; m_1 is a decision variable. These groups are initially group-screened at the ELISA station. If a group is found clean it is sent to the PCR station for individual screening (i.e., each of the m_1 units in the group is screened individually). Otherwise, if the group is found contaminated, the m_1 units in the contaminated group are divided into k_2 subgroups of size m_2, all of which are recycled and resent to the ELISA station for further group-screening. The whole process then repeats itself for the next recycle (each contaminated subgroup is divided into k_3 subgroups of size m_3, which are recycled and resent to the ELISA station for further group-screening), and so forth. The reason for recycling at the ELISA station is the cost reduction, as the testing cost at the ELISA station is significantly smaller (\$2.5) than that at the PCR station (\$85 per blood unit, cf. Subsection 1.1). Such

a recycling at the ELISA station also makes sense from a time perspective, as the processing time at this station is around 1 hour per group whereas that at the PCR station is around $5-6$ hours. The aim of the testing is to satisfy a given demand of d blood units in a cost-efficient manner. For this one has to decide on the number of cycles at the ELISA station and the group sizes m_i.

In the sequel, we shall say that the blood testing process has r *cycles* if the last subdivision is in k_r groups of size m_r – so if one does only one group test, then $r = 1$. The efficient number of cycles depends on several factors.

First, the group sizes satisfy $m_1 > m_2 > ...$, and after a certain number of cycles, the number of items in the resulting subgroups becomes too small, making further recycling no longer worthwhile. Accordingly, the recycling is stopped at the smallest $m_i \geq \hat{m}$, where $\hat{m} \geq 2$ is the smallest permissible group size ($\hat{m}$ will be considered as a parameter of the objective function). There is also an upper bound $\tilde{m}$ on the number of samples that can be pooled together due to technical restrictions, so that $m_1 \leq \tilde{m}$; in practice usually no more than 50 samples are taken in one group.

Second, the processing times cause an upper bound on the number of cycles. We require that the total processing time of all cycles at the ELISA station will not exceed a predetermined time t_0. We assume that every test takes a time interval of fixed (constant) length, regardless of the batch size; every test takes T_{elisa} time units. Then the testing is terminated at the latest after the (t_0/T_{elisa})-th test.

Third, we assume that there is a prespecified upper bound c_0 for the total cost of the ELISA cycles and that the cost per test is constant, say C_{elisa}. Accordingly, the cost limit is reached after c_0/C_{elisa} tests.

Let N_h be the total number of tests after h cycles. According to the above constraints, the ELISA recycling process stops after the τ-th cycle where

$$\tau \doteq \sup\left\{h : m_h \geq \hat{m} \text{ and } N_h \leq \frac{t_0}{T_{elisa}} \text{ and } N_h \leq \frac{c_0}{C_{elisa}}\right\}. \tag{1}$$

We can equivalently write this stopping time as

$$\tau = \min[\tau_1, \tau_C], \tag{2}$$

where

$$\tau_1 \doteq \sup\{h : m_h \geq \hat{m}\}, \tag{3}$$

and

$$\tau_C \doteq \sup\{h : N_h \leq C\}, \tag{4}$$

with

$$C := \min(\frac{t_0}{T_{elisa}}, \frac{c_0}{C_{elisa}}). \tag{5}$$

We now formulate a profit objective function, which has to be maximized with respect to the decision parameters $m_1 > m_2 > \ldots$ The following notation will be used throughout:

- **Probability parameters**

ε - the probability that a unit is found contaminated by ELISA.

γ - the conditional probability that a unit is found contaminated by PCR given it was found clean by ELISA.

- **Counting random variables**

Again, the variables are considered on a weekly basis.

M - the total number of group tests at the ELISA station.

N - the total number of units (out of n) found clean by ELISA.

N^* - the total number of units (out of n) found clean by both ELISA and PCR. Note that $N^* \leq N \leq n$ and

$$\mathbb{E}\left(N^*\right) = (1-\gamma)\mathbb{E}\left(N\right). \tag{6}$$

- **Costs, penalties and rewards**

C_{pur} - the purchasing cost for a single unit (related to acquiring the n initial units).

C_{elisa} - the cost of a test for a batch of arbitrary size m at the ELISA station.

C_{pcr} - the cost of testing a single unit at the PCR station.

$C_{penalty}$ - the penalty cost for not satisfying a demand for a single unit which has been tested clean.

R_w - the reward for a satisfied demand unit.

$\hat{R}_w$ - the reward for any 'surplus' unit beyond the required demand.

This yields the following components of the **objective function**:

Cost components

- $C_{pur}n$ - the total cost for purchasing n units.
- $C_{elisa}M$ - the total cost for testing M groups at the ELISA station.
- $C_{pcr}N$ - the total cost for testing N units at the PCR station.
- $C_{penalty}\left[d - N^*\right] I_{(N^* \leq d)}$ - the total penalty for not satisfying the full demand d (here and in the sequel, $I_{(\cdot)}$ denotes an indicator function).

Reward components

- $R_w N^* I_{(N^* \leq d)}$ - the total reward for the satisfied (part of the) demand.
- $\hat{R}_w \left(N^* - d\right) I_{(N^* > D)}$ - the total reward for the demand surplus.

Thus the total reward is given by

$$R_w N^* I_{(N^* \leq d)} + \left[R_w d + \hat{R}_w \left(N^* - d\right)\right] I_{(N^* > d)}. \tag{7}$$

Combining all of the above, the (random) profit $P = P(m_1, m_2, ...)$ associated with the procedure is

$$\begin{aligned} P = {} & R_w N^* I_{(N^* \leq d)} + \left[R_w d + \hat{R}_w \left(N^* - d\right)\right] I_{(N^* > d)} \\ & - \left[C_{pur}n + C_{elisa}M + C_{pcr}N + C_{penalty}\left[d - N^*\right] I_{(N^* \leq d)}\right]. \end{aligned} \tag{8}$$

The objective function is then given by the expected profit $\tilde{P} = \mathbb{E}[P(m_1, m_2, ...)]$:

$$\begin{aligned}\tilde{P} = \mathbb{E}[R_w N^* I_{(N^* \leq d)}] + \mathbb{E}\left[\{R_w d + \hat{R}_w (N^* - d)\} I_{(N^* > d)}\right] \\ - \left\{C_{pur} n + C_{elisa}\mathbb{E}[M] + C_{pcr}\mathbb{E}[N] + C_{penalty}\mathbb{E}[(d - N^*) I_{(N^* \leq d)}]\right\}.\end{aligned} \tag{9}$$

We rewrite this as

$$\tilde{P} = P_1 + P_2 - P_3 - P_4 - P_5 - P_6, \tag{10}$$

where

$$\begin{cases} P_1 = \mathbb{E}[R_w N^* I_{(N^* \leq d)}], \\ P_2 = \mathbb{E}\left[\{R_w d + \hat{R}_w (N^* - d)\} I_{(N^* > d)}\right], \\ P_3 = C_{pur} n, \\ P_4 = C_{elisa}\mathbb{E}[M], \\ P_5 = C_{pcr}\mathbb{E}[N], \\ P_6 = C_{penalty}\mathbb{E}[(d - N^*) I_{(N^* \leq d)}]. \end{cases} \tag{11}$$

To maximize this objective function we have to find explicit expressions for all its ingredients, i.e., the expected values on the right side of (9). This is carried out in the next two sections by deriving the underlying distributions in closed form.

3 An Unobservable Underlying Markov Chain

In this section we introduce a Markov chain that captures the essence of the testing procedure. We determine its distribution in closed form, and we show that the distributions of all quantities of interest (M, N, N^* featuring in the profit objective function) can be expressed in terms of this Markov chain.

Recall that, after conditioning on the initial supply of a (binomially distributed) number of clean items in the population of size n, the basic RIIP model can be described as follows. Among n items r are contaminated and the rest is clean. In the first stage the items are split into k_1 groups, each of group size $m_1 = n/k_1$, which are pooled and tested together. Each of the groups found contaminated is then split into k_2 subgroups, each of size $m_2 = m_1/k_2$, which are then tested together. The procedure is iterated. Let us first assume that it is iterated until in the final stage only contaminated 'groups' of size 1 are left, which are then tested and the complete picture becomes known. Later we will truncate this procedure and use stopping times of the form (2), but we will see that this modification can easily be taken care of. Let p be the number of cycles, so that $n = \prod_{j=1}^{p} k_j$ and

$$m_1 > \cdots > m_{p-1} > m_p = 1.$$

For $j = 1, 2, ..., p$ let Y_l^j be the number of groups in the jth cycle that contain exactly l contaminated items. For $j = 0$ let Y^0 be the initial number of contaminated items. It is important to note that the Y_l^j, $l \geq 1$, are *not observable* during the testing process. Only Y_0^j and $\sum_{l=1}^{m_j} Y_l^j$, which denote the number of groups found clean or contaminated in the jth cycle, respectively, become known after this cycle. We will see that all distributions of interest to us and all cost and reward functionals we may want to consider can be expressed in terms of the *joint distribution of the sequence*

$$\left((Y_0^j, Y_1^j, \ldots, Y_{m_j}^j)\right)_{j=0,\ldots,p},$$

where $m_0 = 0$ and $Y_0^0 = Y^0$. This finite sequence is a *nonhomogeneous Markov chain*, so to determine its joint distribution we only have to derive all its one-step transition probabilities. This in turn can be reduced to some combinatorial considerations.

3.1 A Combinatorial Urn Problem

We now solve the following auxiliary urn problem. Consider an urn containing r red balls and $n-r$ white balls. Take out all balls in k groups of size $m = n/k$ according to the equidistribution. For $l = 0, 1, 2, ...$ let Y_l be the number of groups containing exactly l red balls; note that $Y_l = 0$ for $l > r$. Compute the probability

$$p(y_0, \dots, y_r \mid n, r, m) = \mathbb{P}(Y_0 = y_0, \dots, Y_r = y_r).$$

Solution: We only consider those values of $y_0, \dots, y_r$ for which the probability is positive. Thus the y_i's are nonnegative integers satisfying

$$y_0 + \cdots + y_r = k, \tag{12}$$
$$y_1 + 2y_2 + 3y_3 + \cdots + ry_r = r, \tag{13}$$
$$my_0 + (m-1)y_1 + (m-2)y_2 + \cdots + (m-r)y_r = n - r. \tag{14}$$

These equations indicate that there are k groups in total, the total number of red balls is r and that of white balls is $n - r$.

Of course $p(y_0, \dots, y_r \mid n, r, m)$ is proportional to the number of ways, say $h(y_0, \dots, y_r \mid n, r, m)$, to split the n balls accordingly.

In a first step, to achieve $Y_0 = y_0$, we have to take out y_0 groups of size m from the $n - r$ white balls. For this there are

$$h_0(y_0 \mid n, r, m) = \frac{1}{y_0!}\binom{n-r}{m}\binom{n-r-m}{m}\cdots\binom{n-r-(y_0-1)m}{m} \tag{15}$$

possibilities; we have to divide by $y_0!$ since we want an unordered set of groups.

How many possibilities are there to achieve $Y_l = y_l$, given that $Y_0 = y_0, \dots, Y_{l-1} = y_{l-1}$? We have to form y_l groups of size m each containing l red balls and $m - l$ white balls, where the white balls have to be chosen from the $n - r - y_0 m - ... - y_{l-1}(m - l + 1)$ remaining white balls and the red balls have to be chosen from the $r - y_1 - 2y_2 - ... - (l-1)y_{l-1}$ remaining red balls. The unordered number of ways for this is

$$h_l(y_l \mid y_0, ..., y_{l-1}, n, r, m) = \frac{1}{y_l!}\binom{r - y_1 - 2y_2 - ... - (l-1)y_{l-1}}{l}\binom{n - r - y_0 m - ... - y_{l-1}(m-l+1)}{m-l}$$
$$\times \cdots \times \binom{r - y_1 - 2y_2 - ... - (l-1)y_{l-1} - l(y_l - 1)}{l}\binom{n - r - y_0 m - ... - y_{l-1}(m-l+1) - (y_l-1)(m-l)}{m-l}. \tag{16}$$

Therefore,

$$h(y_0, \dots, y_r \mid n, r, m) = h_0(y_0 \mid n, r, m)\prod_{l=1}^{r} h_l(y_l \mid y_0, ..., y_{l-1}, n, r, m).$$

Note that the terms in Formula (16) are symmetric with respect to the red and white balls (they also consider all possible arrangements of white balls within a group and all possible arrangements of red balls). This also holds for (15), as becomes obvious after multiplying (15) by $(\binom{r}{r})^{y_0}$.

Formula (16) greatly simplifies after canceling out several factors. Notice that the first term of (15) (for y_1) starts with a factor $(n - r - y_0 m)!$, which cancels against the $(n - r - y_0 m)!$ term in the formula above. In this way, a lot of terms cancel when we consider $h = \prod_{j=0}^{r} h_j$, and we get:

$$h(y_0, \dots, y_r | n, r, m) = \frac{(n-r)!r!}{y_0! \dots y_r!} \quad \frac{1}{(m!0!)^{y_0}((m-1)!1!)^{y_1} \dots ((m-r)!r!)^{y_r}}. \tag{17}$$

Clearly,

$$p(y_0,\dots,y_r \mid n,r,m) = \frac{h(y_0,\dots,y_r \mid n,r,m)}{\sum_{z_0,\dots,z_r} h(z_0,\dots,z_r \mid n,r,m)},$$

where the sum in the denominator extends over all $z_0,\dots,z_r$ satisfying (12)-(14). In the ratio all factors only depending on n,r,m cancel and we arrive at

Theorem 1

$$p(y_0,\dots,y_r|n,r,m) = \frac{\prod_{i=0}^{r}\binom{m}{i}^{y_i}/y_i!}{\sum\prod_{i=0}^{r}\binom{m}{i}^{z_i}/z_i!}, \tag{18}$$

where the sum in the denominator runs over all values of $z_0,\dots,z_r$ satisfying (12)-(14).

With hindsight, Formula (18) is very natural; the ith term in the numerator reflects the number of ways of ordering i red balls in each of y_i urns, all urns containing m balls.

Let us denote the distribution on the set of tuples $(y_0,\dots,y_r)$ satisfying (12)-(14) with probability function (18) by $\mu(n,r,m)$.

3.2 Distribution of the Markov Chain

The Markov chain clearly starts with

$$\mathbb{P}(Y^0 = l) = \binom{n}{l}\varepsilon^l(1-\varepsilon)^{n-l},$$

since before we begin pooling, there is an initial supply of n items where each of them has, independently of the others, probability ε of being contaminated. Write $Y^j = (Y_0^j, Y_1^j, \dots, Y_{m_j}^j)$, $y^j = (y_0^j, y_1^j, \dots, y_{m_j}^j)$. The one-step transition probabilities, i.e., the conditional distributions $\mathbb{P}_{Y^{j+1}|Y^j}$, are given by

Theorem 2

$$\mathbb{P}(Y^{j+1} = y^{j+1} \mid Y^j = y^j) = \Big(\tilde{\prod}_{r=1}^{m_j}\mu(m_j,r,m_{j+1})^{*y_r^j}\Big)(\{y^{j+1}\}). \tag{19}$$

Here $\mu(n,r,m)^{*k}$ denotes the $k$fold convolution of $\mu(n,r,m)$ with itself, and the right-hand side of (19) is the probability of the one-point set $\{y^{j+1}\}$ under the convolution of the m_j probability measures $\mu(m_j,r,m_{j+1})^{*y_r^j}$, $r = 1,...,m_j$. Here $\tilde{\prod}_{r=1}^{m_j}$ denotes the convolution product.

To see (19), argue as follows. To have y_l^{j+1} groups in the $(j+1)$th cycle that have exactly l contaminated items, y_l^{j+1} must be the sum of the numbers of subgroups with l contaminated items formed by splitting the groups after the jth testing. At that time the remaining contaminated groups are of size m_j and are split into subgroups of size m_{j+1}. There are y_r^j groups containing exactly r contaminated items, $r = 0,...,m_j$. The numbers of newly formed subgroups with exactly l contaminated items from all these groups have to be added to obtain the number of such subgroups in the $(j+1)$th cycle. These numbers are independent random variables, which explains the convolutions. Eq. (19) is proved.

The joint distribution of the finite nonhomogeneous Markov chain is given by

Theorem 3

$$\mathbb{P}(Y^j = y^j \ \ for \ \ j = 0, ..., p) = \binom{n}{y_0} \varepsilon^{y_0} (1-\varepsilon)^{n-y_0} \Big(\prod_{j=0}^{p-1} \tilde{\prod}_{r=1}^{m_j} \mu(m_j, r, m_{j+1})^{*y_r^j} \Big) (\{y^{j+1}\}). \quad (20)$$

4 Exact Distributions of the Stopping Time and of M, N, N^*

The testing procedures are stopped when the subgroups are getting too small, say after τ_1 cycles (this means that $m_{\tau_1} \geq \hat{m} > m_{\tau_1+1}$, cf. (3)). We consider stopping times $\tau = \min(\tau_1, \tau_C)$, as defined in (1)-(5). Now observe that the total number of tested groups in the first j cycles is given by

$$\sum_{i=1}^{j} \sum_{l=0}^{m_i} Y_l^i.$$

Consequently, cf. (4),

$$\tau_C = \sup\{j \mid \sum_{i=1}^{j} \sum_{l=0}^{m_i} Y_l^i \leq C\}. \quad (21)$$

Thus the distribution of τ_C can be expressed in terms of the underlying Markov chain, whose distribution was obtained in Theorem 3. The same holds for the distributions of M, N and N^*.

Theorem 4

$$\mathbb{P}(\tau_C > j) = \mathbb{P}\left(\sum_{i=1}^{j} \sum_{l=0}^{m_i} Y_l^i \leq C \right), \quad (22)$$

$$\mathbb{P}(M = q) = \sum_{j=1}^{\tau_1 - 1} \mathbb{P}\left(\sum_{i=1}^{j} \sum_{l=0}^{m_i} Y_l^i = q, \sum_{i=1}^{j+1} \sum_{l=0}^{m_i} Y_l^i > C \right) + \mathbb{P}\left(\sum_{i=1}^{\tau_1} \sum_{l=0}^{m_i} Y_l^i = q \right), \quad q \leq C, \quad (23)$$

$$\mathbb{P}(N = h) = \sum_{j=1}^{\tau_1 - 1} \mathbb{P}\left(\sum_{i=1}^{j} m_i Y_0^i = h, \sum_{i=1}^{j} \sum_{l=0}^{m_i} Y_l^i \leq C < \sum_{i=1}^{j+1} \sum_{l=0}^{m_i} Y_l^i \right) + \mathbb{P}\left(\sum_{i=1}^{\tau_1} m_i Y_0^i = h, \sum_{i=1}^{\tau_1} \sum_{l=0}^{m_i} Y_l^i \leq C \right), \quad (24)$$

$$\mathbb{P}(N^* = s) = \sum_{z=s}^{n} \mathbb{P}(N = z) \binom{z}{s} (1-\gamma)^s \gamma^{z-s}. \quad (25)$$

Proof. Eq. (22) follows immediately from (21). Next, M (the total number of conducted group tests) can be decomposed as

$$M = \sum_{i=1}^{\tau} \sum_{l=0}^{m_i} Y_l^i,$$

yielding (23). Similarly, the number of items tested clean at the ELISA station during the iterations can be represented as

$$N = \sum_{i=1}^{\tau} m_i Y_0^i,$$

so that its distribution can also be obtained from (20). Eq. (24) is now easily checked. Finally, (25) follows from the law of total probability.

All the expected values appearing in the objective function $\tilde{P}$, which was introduced in (9), can now be expressed in terms of the underlying Markov chain. The resulting formulas are very lengthy.

5 An Approximation for Numerical Purposes

5.1 Some Realistic Assumptions and Related Approximations

The explicit formulas for the components of the objective function, as derived in the previous section, are very complicated. For the purposes of optimization it is hence important to come up with approximations for these components. A starting point for such approximations is the following consideration. Realistic numbers might be (cf. Subsection 1.3):

$$\begin{aligned} n &= 6720 \text{ per week}, \\ \epsilon &= \mathbb{P}(\text{unit contaminated}) = 6.8 \times 10^{-4}, \\ \gamma &= 4 \times 10^{-5}, \\ m_1 &= 48. \end{aligned}$$

So we have about 140 initial batches per week, and 7280 per year. The probability that a 48-batch is not contaminated is $(1-\epsilon)^{m_1} \approx 1 - m_1\epsilon \approx 0.97$. The probability that a 48-batch has one contaminated unit is $m_1\epsilon(1-\epsilon)^{m_1-1} \approx 0.03$. Thus on average this occurs approximately four times per week. The probability that a 48-batch has two contaminated units is $\frac{1}{2}m_1(m_1-1)\epsilon^2(1-\epsilon)^{m_1-2} \approx 6 \times 10^{-4}$. Therefore, on average this occurs approximately four times per year. From this one gets a feeling for the proportions.

Now looking at blood *units*, instead of batches, per week, the number of contaminated units in a week is binomially distributed with parameters n and ϵ, and for $n = 6720$ and $\epsilon = 6.8 \times 10^{-4}$ this is extremely accurately represented by a Poisson distribution with parameter $\lambda = 4.7$. This brings us to our

Approximation Assumption 1:
The distribution of the number of contaminated items per week can be approximated by a Poisson distribution with parameter λ, which can be considered as the arrival rate of contaminated items (per week), and which for given n equals $\lambda = n\epsilon$.

So with X the number of contaminated items in an arbitrary week:

$$\mathbb{P}(X = i) = \mathrm{e}^{-\lambda}\lambda^i/i!, \qquad i = 0, 1, \ldots, \tag{26}$$

where for the present example we may choose $\lambda = 4.7$.

For our objective function (9) we need approximative expressions for the terms

$$\mathbb{E}[N^* I_{(N^* \le d)}], \quad \mathbb{P}(N^* > d), \quad \mathbb{E}\left[(N^* - d)\, I_{(N^* > d)}\right], \quad \mathbb{E}[M], \quad \mathbb{E}[N], \quad \mathbb{E}[(d - N^*) I_{(N^* \le d)}]. \tag{27}$$

Now we introduce

Approximation Assumption 2:
One can ignore the event that at least two of the contaminated items are in the same initial batch (which is an event with probability of order ϵ^2).

We claim that both approximation assumptions are extremely accurate for realistic values of the contamination rate ϵ.

So now assume that initially there are $X = i$ contaminated items, all belonging to *different* batches. In this case there will be i contaminated batches in every recycling (each contaminated by one item) so that, when there are h cycles, the number of tests in this week equals $N_h = \frac{n}{m_1} + i(\frac{m_1}{m_2} + \cdots + \frac{m_{h-1}}{m_h})$ (remember that $k_z = \frac{m_{z-1}}{m_z}$ is the number of batches being tested in the zth cycle, $z = 2, 3, \dots$). It follows from (2) and (4) that the stopping times τ_1 and τ_C, and thus also τ, are constants (depending on i and on the batch sizes): we have $\tau \equiv h_i$, where $h_i = \min(\tau_1, \tau_C)$.

Let us now turn to the distribution of N, for given n. First of all, $P(N = n) = (1-\epsilon)^n \approx \mathrm{e}^{-\lambda}$; indeed, this is the case that all items of this week are clean. Secondly, $P(N = n - m_{h_1})$ equals the probability that this week there is exactly one contaminated item, i.e., $\mathbb{P}(X = 1)$. Hence $P(N = n - m_{h_1}) \approx \lambda \mathrm{e}^{-\lambda}$. Indeed, if we test the contaminated batch τ times (in smaller and smaller batches), then the number of items (out of the original m_1) that we do not send to PCR equals m_τ. Generally,

$$\mathbb{P}(N = n - i m_{h_i}) \approx \mathrm{e}^{-\lambda} \lambda^i / i!, \quad i = 0, 1, \dots \ . \tag{28}$$

This corresponds to the probability of having $X = i \geq 1$ contaminated items in this week; recall that we ignore the event that at least two of them are in the same initial batch (having probability of order ϵ^2).

Let us next consider the distribution of M, the number of groups tested in a week. With the same reasoning as above, again ignoring the event that at least two contaminated items appear in the same initial batch, we have:

$$\mathbb{P}(M = \frac{n}{m_1} + i(k_2 + \cdots + k_{h_i})) = \mathrm{e}^{-\lambda} \lambda^i / i!, \quad i = 0, 1, \dots \ . \tag{29}$$

Finally, the distribution of N^* can be approximately determined from (25). In view of the fact that $\gamma \approx 4 \times 10^{-5}$, we can very accurately approximate

$$\mathbb{P}(N^* = s) \approx \mathbb{P}(N = s) + (s+1)\gamma \mathbb{P}(N = s+1), \tag{30}$$

or even $\mathbb{P}(N^* = s) \approx \mathbb{P}(N = s)$; in fact, we propose to use the latter formula in the four terms in the objective function involving N^*. From (28)-(30) approximations for all terms in (27) are easily obtained. This yields a simple approach to the objective function that can be used for a tentative optimization.

The same reasoning can be used in the case when the stopping time is a *prespecified constant*, say h^0. Then

$$\mathbb{P}(N = n - i m_{h^0}) \approx \mathrm{e}^{-\lambda} \lambda^i / i!, \quad i = 0, 1, \dots \ . \tag{31}$$

This again corresponds to the probability of having $X = i$ contaminated items in this week. Taking means, we get

$$\mathbb{E}[N] \approx n - \lambda m_{h^0}. \tag{32}$$

Furthermore, in this case we get for the distribution of M

$$\mathbb{P}(M = \frac{n}{m_1} + i(k_2 + \cdots + k_{h^0})) \approx \mathrm{e}^{-\lambda} \lambda^i / i!, \quad i = 0, 1, \dots \tag{33}$$

In particular,

$$\mathbb{E}[M] \approx \frac{n}{m_1} + \lambda(k_2 + \cdots + k_{h^0}) = \frac{n}{m_1} + \lambda(\frac{m_1}{m_2} + \cdots + \frac{m_{h^0-1}}{m_{h^0}}). \tag{34}$$

5.2 Three Possible Approximation Cases

Consequently, taking into account Approximation Assumptions 1 and 2 above, we may distinguish (at least) three cases.

1. **Case I: the stopping time is a prespecified constant h^0**
   This is a natural and important case. We elaborate on this case in the sequel.

2. **Case II: $X = i$ is given**
   Here, we mean the following. Suppose we do a first cycle with groups of size m_1 (a decision variable). After this cycle, we count the number of contaminated groups. Suppose this number, X, equals i. Ignoring the possibility of having more than one contaminated item in the same group or subgroup, if we continue until we have done a total number of h_i cycles, we shall have $N = n - im_{h_i}$. Our decision variables are $m_1, m_2, \ldots, m_{h_i}$ (as well as h_i, unless we decide that we do as many cycles as possible, as long as $m_h \geq \hat{m}$ and $N_h \leq C$).

   Now, we claim that there are not so many cases to choose among. In general, the number of possibilities depends on the prime factorization of n. The numbers of groups, i.e., the k_i, can be generated by any partition of the prime factors of n, resulting in a huge number of grouping schemes. However, their number is drastically reduced by the upper bound on m_1. In our numerical examples we take $n = 6720$ and suppose that $m_1 \in \{16, 24, 32, 40, 48, 56, 64\}$, that $m_2 \in \{8, 12, 16, \ldots, m_1/2\}$ and that $m_3 \in \{4, 6, \ldots, m_2/2\}$. We may allow a fourth and even a fifth cycle, but that seems not realistic if $m_1 \leq 64$ and we make new groups always at least twice as small as the groups of the previous cycle. Furthermore, we cannot consider too small m_1-values, since $N_h \leq C$ says that $n/m_1 \leq C$. The small number of cases to be considered makes it very easy to search among all possible cases. Searching on the one hand means: checking whether the constraints regarding the stopping times are not violated. Searching on the other hand means: calculate the profit objective function, and take the largest one among those for which the stopping time constraints are not violated. In Case II, one has to adapt that profit objective function, given by formula (9), in an obvious way. We denote it by $\tilde{P}_i$ to emphasize its dependence on i, and calculate, e.g., its first term as follows:

$$R_w \mathbb{E}[N I_{(N \leq d)} | X = i] = R_w (n - i m_{h_i}) I_{(n - i m_{h_i} \leq d)}.$$

   In this way we get a value for $\tilde{P}_i$, for our given i, and for all combinations of $m_1, m_2, \ldots, m_{h_i}$ and h_i. As explained above, one can now take the combination, among the admissible ones, that yields the highest profit.

3. **Case III: the general case**
   One way to treat this case is to take case II (but with h_i replaced by a decision variable H that does not depend on the actual value of $X = i$), and multiply $\tilde{P}_i$ by $P(X = i)$ and sum over $i = 0, 1, \ldots, Z$. Here Z might be such that $P(X > Z) = \sum_{i=Z+1}^{\infty} \mathrm{e}^{-\lambda} \lambda^i / i! \leq 10^{-4}$, with, e.g., $\lambda = 4.7$. Do this again for all possible combinations of $m_1, m_2, \ldots, m_H$ and number of cycles H, where we allow H to take the values $1, 2, 3, 4, 5$, say, and where we check which combinations do not violate our stopping conditions.

Below we further elaborate on Cases I and III (Case II may be viewed as a special case of Case III).

Case I: the stopping time is a prespecified constant h^0

For simplicity, we assume that n is divisible by m_{h^0}. All functionals appearing in (11) can be simply obtained by using the approximations in (32) and (34). Indeed, by letting

$$\tilde{k} \doteq \frac{n-d}{m_{h^0}} \text{ and } F(t) = \mathbb{P}(X \leq t) = \sum_{i=0}^{[t]} \mathrm{e}^{-\lambda}\lambda^i/i!$$

(recall that X has a Poisson distribution with mean λ), the expressions for $P_1, ..., P_6$ are obtained as follows. First,

$$P_1 = \mathbb{E}[R_w N^* I_{(N^* \leq d)}] \simeq \mathbb{E}[R_w N I_{(N \leq d)}] = R_w \mathbb{E}[N I_{(N \leq d)}],$$

and for the righthand side note that $\mathbb{E}[N I_{(N \leq d)}] = \mathbb{E}[N] - \mathbb{E}[N I_{N > d)}]$, where $\mathbb{E}[N] = n - \lambda m_{h^0}$ and $\mathbb{E}[N I_{N>d)}] = n\mathbb{P}(X = 0) + (n - m_{h^0})\mathbb{P}(X = 1) + ...$

A straightforward computation yields

$$\mathbb{E}[N I_{(N \leq d)}] = \begin{cases} n - \lambda m_{h^0}, & \text{if } d \geq n \ (\tilde{k} \leq 0) \\ n(1 - \mathrm{F}(0)) - \lambda m_{h^0}, & \text{if } d < n \text{ and } 0 < \tilde{k} \leq 1 \\ n\left[1 - F(\tilde{k} - 1)\right] - \lambda m_{h^0}\left[1 - F(\tilde{k} - 2)\right], & i\text{if } d < n \text{ and } 2 \leq \tilde{k} \in \mathbb{N} \\ n\left[1 - F([\tilde{k}])\right] - \lambda m_{h^0}\left[1 - F([\tilde{k}] - 1)\right], & \text{if } d < n \text{ and } 1 < \tilde{k} \notin \mathbb{N}. \end{cases}$$

Next,

$$P_2 = \mathbb{E}\left[\{dR_w + \hat{R}_w\,(N^* - d)\} I_{(N^* > d)}\right] \simeq dR_w \mathbb{P}(N > d) - \hat{R}_w \mathbb{E}[(d - N) I_{(N > d)}],$$

where

$$dR_w \mathbb{P}(N > d) \simeq \begin{cases} 0, & \text{if } d \geq n \ (\tilde{k} \leq 0) \\ dR_w F(\tilde{k} - 1), & \text{if } d < n \text{ and } \tilde{k} \in \mathbb{N} \\ dR_w F([\tilde{k}]), & \text{if } d < n \text{ and } \tilde{k} \notin \mathbb{N}, \end{cases}$$

and

$$\mathbb{E}[(d - N) I_{(N > d)}] = \mathbb{E}(d - N) - \mathbb{E}[(d - N) I_{(N \leq d)}] = d - (n - \lambda m_{h^0}) - \mathbb{E}[(d - N) I_{(N \leq d)}].$$

Combining the above results we find that

$$\mathbb{E}[(d - N) I_{(N \leq d)}] \simeq \begin{cases} d - n + \lambda m_{h^0}, & \text{if } d \geq n \ (\tilde{k} \leq 0) \\ (d - n)(1 - F(0)) + \lambda m_{h^0}, & \text{if } d < n \text{ and } 0 < \tilde{k} \leq 1 \\ (d - n)\left[1 - F(\tilde{k} - 1)\right] + \lambda m_{h^0}\left[1 - F(\tilde{k} - 2)\right], & \text{if } d < n \text{ and } 2 \leq \tilde{k} \in \mathbb{N} \\ (d - n)\left[1 - F([\tilde{k}])\right] + \lambda m_{h^0}\left[1 - F([\tilde{k}] - 1)\right], & \text{if } d < n \text{ and } 1 < \tilde{k} \notin \mathbb{N} \end{cases}$$

and

$$\mathbb{E}[(d - N) I_{(N > d)} \simeq \begin{cases} 0, & \text{if } d \geq n \ (\tilde{k} \leq 0) \\ (d - n)F(0), & \text{if } d < n \text{ and } 0 < \tilde{k} \leq 1 \\ (d - n)F(\tilde{k} - 1) + \lambda m_{h^0} F(\tilde{k} - 2), & \text{if } d < n \text{ and } 2 \leq \tilde{k} \in \mathbb{N} \\ (d - n)F([\tilde{k}]) + \lambda m_{h^0} F([\tilde{k}] - 1), & \text{if } d < n \text{ and } 1 < \tilde{k} \notin \mathbb{N} \end{cases}$$

implying that

$$P_2 \simeq \begin{cases} 0, & \text{if } d \geq n \ (\tilde{k} \leq 0) \\ dR_w F(0) + \hat{R}_w (n - d)F(0), & \text{if } d < n \text{ and } 0 < \tilde{k} \leq 1 \\ dR_w F(\tilde{k} - 1) + \hat{R}_w[(n - d)F(\tilde{k} - 1) - \lambda m_{h^0} F(\tilde{k} - 2)], & \text{if } d < n \text{ and } 2 \leq \tilde{k} \in \mathbb{N} \\ dR_w F([\tilde{k}]) + \hat{R}_w[(n - d)F([\tilde{k}]) - \lambda m_{h^0} F([\tilde{k}] - 1)], & \text{if } d < n \text{ and } 1 < \tilde{k} \notin \mathbb{N}. \end{cases}$$

The other components of the profit function are given by

$$P_3 = C_{pur} n$$

$$P_4 = C_{elisa}\mathbb{E}[M] = C_{elisa}\left[\frac{n}{m_1} + \lambda\left(\frac{m_1}{m_2} + ... + \frac{m_{h^0-1}}{m_{h^0}}\right)\right]$$

$$P_5 = C_{pcr}\mathbb{E}[N] = C_{pcr}(n - \lambda m_{h^0})$$

$$P_6 = C_{penalty}(E[(d - N^*)I_{(N^* \leq d)}]) \simeq C_{penalty}(\mathbb{E}[(d - N)I_{(N \leq d)}].$$

Case III: the general case

For convenience, any possible assignment for the values of $h_0, m_1, ..., m_{h_0}$ will be called a strategy, denoted by $S = S(h_0, m_1, ..., m_{h_0})$. For any given initial values of parameters involved $(n, R_w, \hat{R}_w, ...)$, the number of strategies is finite. The set of possible strategies is denoted by $\mathcal{S}$.

For every strategy S the expected value of the related profit is given by

$$\tilde{P}(S) = \sum_{i=0}^{Z} \mathbb{P}(X = i) \times P_{i,S}, \tag{35}$$

where X, as defined above, is the random variable counting the number of contaminated blood units $(X \sim Poiss(\lambda))$, Z is determined by the constraint

$$\mathbb{P}(X > Z) = \sum_{i=Z+1}^{\infty} \mathrm{e}^{-\lambda}\lambda^i/i! \leq 10^{-4}, \tag{36}$$

and $\tilde{P}_{i,S}$ is the conditional expected profit when using strategy S given that $X = i$.

Before continuing we make the following comments:

1. For any given values of the parameters involved, there are not so many strategies. For example, for $n = 6720$, $m_1 \leq 64$ and minimal batch size $\hat{m} = 4$, there are no more than 99 strategies. In fact, we will provide all of them at the end of Section 6.

2. Following Approximation Assumption 2, we shall ignore the possibility of having more than one contaminated item in the same group or subgroup. If $X = i$ and we continue recycling until we have done a total number of h_i cycles, we then shall have $N = n - im_{h_i}$ and

$$M = M_i = \left[\frac{n}{m_1} + i\left(\frac{m_1}{m_2} + ... + \frac{m_{h_i-1}}{m_{h_i}}\right)\right]. \tag{37}$$

Our decision variables are $m_1, m_2, \ldots, m_{h_i}$ (as well as h_i, unless we decide that we do as many cycles as possible, as long as $m_h \geq \hat{m} = 4$ and $N_h \leq C$).

The small number of cases to be considered makes it easy to search among all possible cases. Searching on the one hand means: checking whether the constraints regarding the stopping times are not violated. Searching on the other hand means: calculate the profit objective function, and take the largest one among those for which the stopping time constraints are not violated.

Accordingly, one has to adapt the profit objective function in (9) in an obvious way. Indeed, for a given $X = i$, the expected profit $\tilde{P}_i$ is

$$\tilde{P}_i = P_{i,1} + P_{i,2} - P_{i,3} - P_{i,4} - P_{i,5} - P_{i,6},$$

where

$$P_{i,1} \simeq R_w(n - im_{h_i})I_{(n-im_{h_i} \le d)}$$

(as $X = i$ is given, the expectation turns out to be a constant), or

$$P_{i,1} \simeq \begin{cases} R_w(n - im_{h_i}), & \text{if } n - im_{h_i} \le d \\ 0, & \text{otherwise} \end{cases},$$

$$P_{i,2} \simeq (R_w d + \hat{R_w}(n - im_{h_i} - d))I_{(n-im_{h_i} > d)},$$

or

$$P_{i,2} \simeq \begin{cases} R_w d + \hat{R_w}(n - im_{h_i} - d), & \text{if } n - im_{h_i} > d \\ 0, & \text{otherwise} \end{cases},$$

$$P_{i,3} = C_{pur} \times n,$$

$$P_{i,4} = C_{elisa} \times M_i,$$

where M_i is given by (37),

$$P_{i,5} = C_{pcr}(n - im_{h_i}),$$

$$P_{i,6} \simeq C_{penalty}\left[d - (n - im_{h_i}))I_{(n-im_{h_i} \le d)}\right]$$

or

$$P_{i,6} \simeq \begin{cases} C_{penalty}(d - (n - im_{h_i})), & \text{if } n - im_{h_i} \le d \\ 0, & \text{otherwise} \end{cases}.$$

6 Numerical and sensitivity analysis

In this section we present a numerical and sensitivity analysis for Cases I and III. Since many different parameters are involved, one could consider a wide range of parameter values, studying the influence of each of them on the profit function. We mainly restrict ourselves to particular parameter values which seem to be realistic in the case of the Israeli Central Blood Bank. This allows us to focus on a few key aspects, viz.:
(i) For Case I, we study the influence of demand on profit.
(ii) For this case, we also consider the six components $P_1, \ldots, P_6$ of the profit function to get an impression of their relative contributions.
(iii) In Case III we study the effect of doing multiple tests (with ever smaller groups) at the ELISA station on the profit function.
(iv) Case III also allows us to study the effect of group sizes at the ELISA station on the profit function.

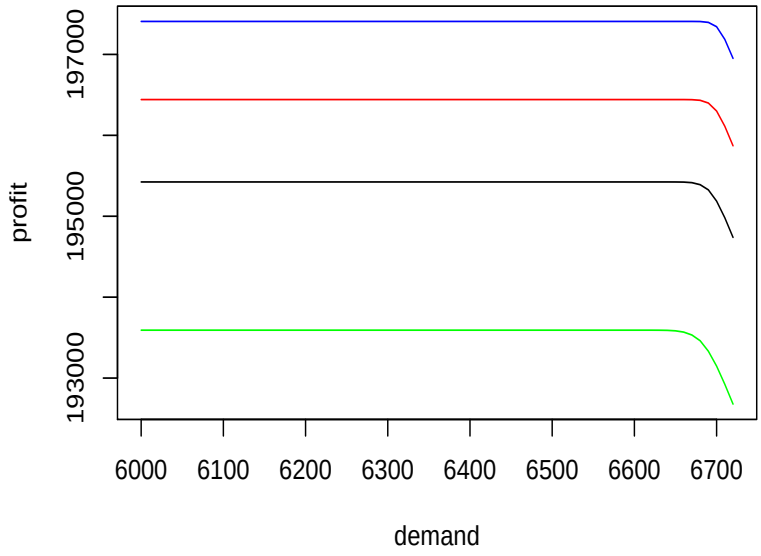

Figure 1: Profit vs. demand for $R_w = \hat{R}_w = 250$

6.1 Case I

We first plot profit versus demand for $h^0 = 4$ and for four cases of iterated batch sizes, given by the following table:

Case	m_1	m_2	m_3	m_4
a	48	24	12	6
b	60	30	15	5
c	64	32	16	8
d	64	16	8	4

Demand changes from 6000 to 6720 and the chosen parameter values are:
$n = 6720$; $\varepsilon = 0.00068$; $C_{pur} = 180$; $C_{elisa} = 2.5$; $C_{pcr} = 40$; $C_{penalty} = 25$.
We consider two cases for R_w and $\hat{R}_w$: $R_w = \hat{R}_w = 250$ and $R_w = 250 > \hat{R}_w = 230$.

Figures 1 and 2 display, respectively, plots of profit vs. demand for the two cases of R_w and $\hat{R}_w$. In both figures, the black curve displays case a, the red one case b, the green one case c and the blue one case d. The behavior of the four different strategies seems to be similar and the profit functions do not intersect.

Various other values of the parameters that have been considered show a pattern similar to the one presented. In particular, when $R_w = \hat{R}_w$ and $C_{penalty} = 0$, the profit stays constant as a function of the demand (indeed, notice that now $P_1 + P_2 = R_w \mathbb{E}[N]$ does not depend on d, and neither do $P_3, \ldots, P_6$).

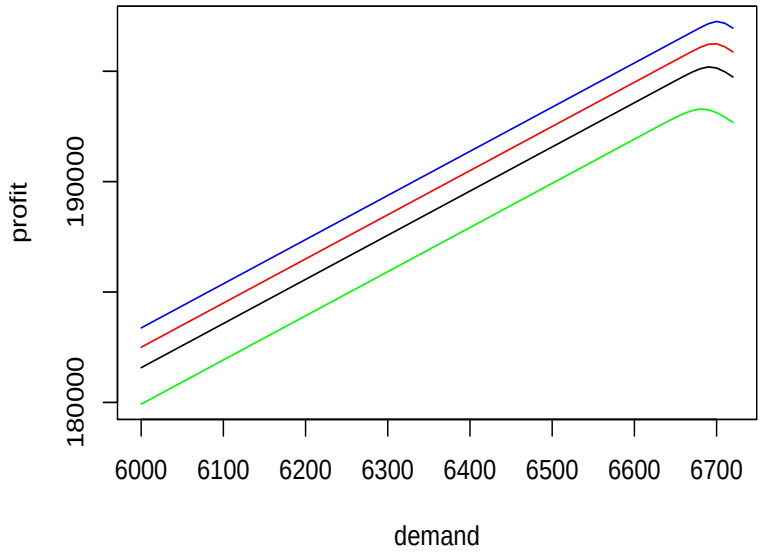

Figure 2: Profit vs. demand for $R_w = 250 > \hat{R}_w = 230$

Next, we study the impact on the cost and penalty components of the expected profit $\tilde{P}$ when slightly lowering the demand.

The parameter values considered are:
$\varepsilon = 0.00068; C_{pur} = 180; C_{elisa} = 2.5;\ C_{pcr} = 40; C_{penalty} = 10;\ R_w = \hat{R}_w = 250.$
h^0 is taken to be 4 and $m_1 = 48, m_2 = 24, m_3 = 12, m_4 = 4$.

We first consider the case where $n = d = 6720$, for which the components of $\tilde{P}$ are given in the following table.

$\tilde{P}$	$197,148.79$
P_1	$1,675,430.40$
P_2	0.00
P_3	$1,209,600.00$
P_4	429.97
P_5	$268,068.86$
P_6	182.78

Here, as expected, $P_2 = 0$, while P_4 and P_6 are almost negligible compared to the profit component P_1.

Now, for $n = 6720 > d = 6715$, a similar table gives

$\tilde{P}$	$197,197.79$
P_1	$1,578,520.29$
P_2	$96,910.11$
P_3	$1,209,600.00$
P_4	429.97
P_5	$268,068.86$
P_6	133.78

Here, too, P_4 and P_6 are almost negligible whereas the total reward for the satisfied demand and demand surplus increases significantly (as compared to the previous case). A similar picture has been observed for other demand values smaller than n.

6.2 Case III

Here, we focus on how the profit depends on the choice of the strategy, without imposing boundary constraints on the number of ELISA tests. We consider the following parameter values:
$n = 6720; d = 6700; \hat{m} = 4; \varepsilon = 0.00068; C_{pur} = 180; C_{elisa} = 2.5; C_{pcr} = 40; C_{penalty} = 10; R_w = 250;$ and $\hat{R}_w = 240$.

Figure 3 displays the expected profit vs. the maximum number of ELISA tests per strategy for different numbers of iterations $h^0 = 1, 2, ..., 5$. For this figure we use colors to distinguish between strategies having different numbers of iteration $h^0 = 1, 2, ..., 5$:

- Black points - for one-iteration strategies;
- Red points - for two-iteration strategies;
- Green points - for three-iteration strategies;
- Blue points - for four-iteration strategies;
- Turquoise points - for five-iteration strategies (only one point, corresponding to successive batch sizes $64, 32, 16, 8, 4$).

Each point in the graph presents the expected profit for one strategy.

Table 1 below displays a full list of all 99 possible strategies (without boundary constraints on the number of ELISA tests) for $n = 6720, \hat{m} = 4, m_1 \in \{64, 56, 48, 42, 40, 32, 28, 24, 21, 20, 16\}$ and $h^0 = 1, 2, ..., 5$. Strategies yielding a profit above 195,000 are boldfaced. It allows us to study the influence of the group sizes.

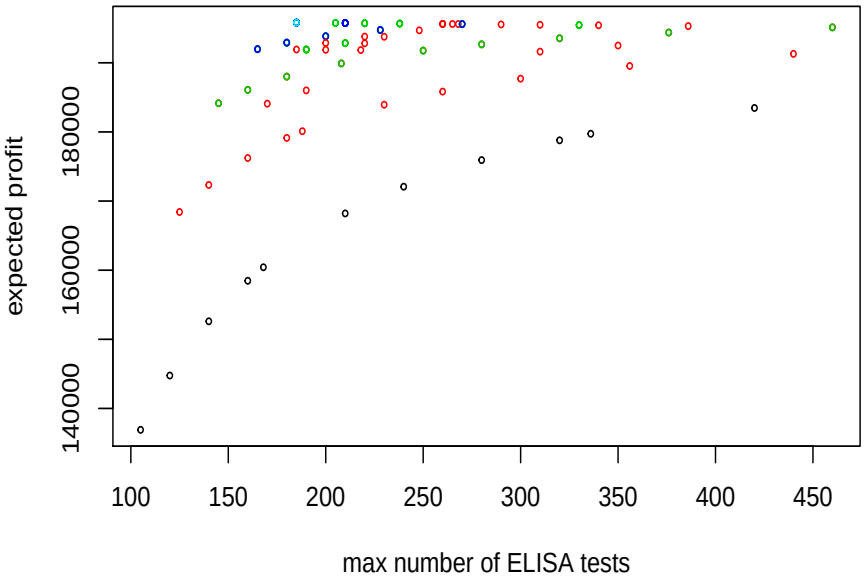

Figure 3: Profit vs. maximum number of ELISA tests

Table 1: Full list of all 99 possible strategies

	m_1	m_2	m_3	m_4	m_5	maximal number of ELISA tests	profit	mean number of ELISA tests
1	16.0	0.0	0.0	0.0	0.0	420.0	183277.6	420.0
2	20.0	0.0	0.0	0.0	0.0	336.0	179540.2	336.0
3	21.0	0.0	0.0	0.0	0.0	320.0	178593.5	320.0
4	24.0	0.0	0.0	0.0	0.0	280.0	175733.4	280.0
5	28.0	0.0	0.0	0.0	0.0	240.0	171886.8	240.0
6	32.0	0.0	0.0	0.0	0.0	210.0	168015.5	210.0
7	40.0	0.0	0.0	0.0	0.0	168.0	160228.1	168.0
8	42.0	0.0	0.0	0.0	0.0	160.0	158275.0	160.0
9	48.0	0.0	0.0	0.0	0.0	140.0	152406.0	140.0
10	56.0	0.0	0.0	0.0	0.0	120.0	144564.0	120.0
11	64.0	0.0	0.0	0.0	0.0	105.0	136709.6	105.0
12	16.0	4.0	0.0	0.0	0.0	460.0	**195015.7**	437.9
13	16.0	8.0	0.0	0.0	0.0	440.0	191119.6	429.0
14	20.0	4.0	0.0	0.0	0.0	386.0	**195199.4**	358.4
15	20.0	5.0	0.0	0.0	0.0	376.0	194237.7	353.9
16	20.0	10.0	0.0	0.0	0.0	356.0	189355.1	345.0
17	21.0	7.0	0.0	0.0	0.0	350.0	192329.3	333.5
18	24.0	4.0	0.0	0.0	0.0	340.0	**195313.5**	306.9
19	24.0	6.0	0.0	0.0	0.0	320.0	193390.2	297.9

20	24.0	8.0	0.0	0.0	0.0	310.0	191442.1	293.5
21	24.0	12.0	0.0	0.0	0.0	300.0	187521.2	289.0
22	28.0	4.0	0.0	0.0	0.0	310.0	**195387.9**	271.4
23	28.0	7.0	0.0	0.0	0.0	280.0	192503.0	257.9
24	28.0	14.0	0.0	0.0	0.0	260.0	185647.5	249.0
25	32.0	4.0	0.0	0.0	0.0	290.0	**195437.6**	245.9
26	32.0	8.0	0.0	0.0	0.0	250.0	191591.0	227.9
27	32.0	16.0	0.0	0.0	0.0	230.0	183749.1	219.0
28	40.0	4.0	0.0	0.0	0.0	268.0	**195492.2**	212.8
29	40.0	5.0	0.0	0.0	0.0	248.0	194555.3	203.9
30	40.0	8.0	0.0	0.0	0.0	218.0	191670.4	190.4
31	40.0	10.0	0.0	0.0	0.0	208.0	189722.3	185.9
32	40.0	20.0	0.0	0.0	0.0	188.0	179907.5	177.0
33	42.0	6.0	0.0	0.0	0.0	230.0	193613.6	191.4
34	42.0	7.0	0.0	0.0	0.0	220.0	192651.9	186.9
35	42.0	14.0	0.0	0.0	0.0	190.0	185821.2	173.5
36	42.0	21.0	0.0	0.0	0.0	180.0	178940.9	169.0
37	48.0	4.0	0.0	0.0	0.0	260.0	**195512.0**	193.8
38	48.0	6.0	0.0	0.0	0.0	220.0	193638.4	175.9
39	48.0	8.0	0.0	0.0	0.0	200.0	191715.1	166.9
40	48.0	12.0	0.0	0.0	0.0	180.0	187818.9	157.9
41	48.0	16.0	0.0	0.0	0.0	170.0	183897.9	153.5
42	48.0	24.0	0.0	0.0	0.0	160.0	176031.2	149.0
43	56.0	4.0	0.0	0.0	0.0	260.0	**195512.0**	182.8
44	56.0	7.0	0.0	0.0	0.0	200.0	192701.6	155.9
45	56.0	8.0	0.0	0.0	0.0	190.0	191739.9	151.4
46	56.0	14.0	0.0	0.0	0.0	160.0	185895.7	137.9
47	56.0	28.0	0.0	0.0	0.0	140.0	172135.0	129.0
48	64.0	4.0	0.0	0.0	0.0	265.0	**195499.6**	176.7
49	64.0	8.0	0.0	0.0	0.0	185.0	191752.3	140.9
50	64.0	16.0	0.0	0.0	0.0	145.0	183960.0	122.9
51	64.0	32.0	0.0	0.0	0.0	125.0	168226.4	114.0
52	16.0	8.0	4.0	0.0	0.0	460.0	**195015.7**	437.9
53	20.0	10.0	5.0	0.0	0.0	376.0	194237.7	353.9
54	24.0	8.0	4.0	0.0	0.0	330.0	**195338.3**	302.4
55	24.0	12.0	4.0	0.0	0.0	330.0	**195338.3**	302.4
56	24.0	12.0	6.0	0.0	0.0	320.0	193390.2	297.9
57	28.0	14.0	7.0	0.0	0.0	280.0	192503.0	257.9
58	32.0	8.0	4.0	0.0	0.0	270.0	**195487.2**	236.9
59	32.0	16.0	4.0	0.0	0.0	270.0	**195487.2**	236.9
60	32.0	16.0	8.0	0.0	0.0	250.0	191591.0	227.9
61	40.0	8.0	4.0	0.0	0.0	238.0	**195566.6**	199.4
62	40.0	10.0	5.0	0.0	0.0	228.0	194605.0	194.9
63	40.0	20.0	4.0	0.0	0.0	238.0	**195566.6**	199.4
64	40.0	20.0	5.0	0.0	0.0	228.0	194605.0	194.9
65	40.0	20.0	10.0	0.0	0.0	208.0	189722.3	185.9
66	42.0	14.0	7.0	0.0	0.0	210.0	192676.7	182.4
67	42.0	21.0	7.0	0.0	0.0	210.0	192676.7	182.4
68	48.0	8.0	4.0	0.0	0.0	220.0	**195611.3**	175.9
69	48.0	12.0	4.0	0.0	0.0	210.0	**195636.1**	171.4
70	48.0	12.0	6.0	0.0	0.0	200.0	193688.0	166.9

71	48.0	16.0	4.0	0.0	0.0	210.0	**195636.1**	171.4
72	48.0	16.0	8.0	0.0	0.0	190.0	191739.9	162.4
73	48.0	24.0	4.0	0.0	0.0	220.0	**195611.3**	175.9
74	48.0	24.0	6.0	0.0	0.0	200.0	193688.0	166.9
75	48.0	24.0	8.0	0.0	0.0	190.0	191739.9	162.4
76	48.0	24.0	12.0	0.0	0.0	180.0	187818.9	157.9
77	56.0	8.0	4.0	0.0	0.0	210.0	**195636.1**	160.4
78	56.0	14.0	7.0	0.0	0.0	180.0	192751.2	146.9
79	56.0	28.0	4.0	0.0	0.0	210.0	**195636.1**	160.4
80	56.0	28.0	7.0	0.0	0.0	180.0	192751.2	146.9
81	56.0	28.0	14.0	0.0	0.0	160.0	185895.7	137.9
82	64.0	8.0	4.0	0.0	0.0	205.0	**195648.5**	149.8
83	64.0	16.0	4.0	0.0	0.0	185.0	**195698.1**	140.9
84	64.0	16.0	8.0	0.0	0.0	165.0	191802.0	131.9
85	64.0	32.0	4.0	0.0	0.0	205.0	**195648.5**	149.8
86	64.0	32.0	8.0	0.0	0.0	165.0	191802.0	131.9
87	64.0	32.0	16.0	0.0	0.0	145.0	183960.0	122.9
88	32.0	16.0	8.0	4.0	0.0	270.0	**195487.2**	236.9
89	40.0	20.0	10.0	5.0	0.0	228.0	194605.0	194.9
90	48.0	16.0	8.0	4.0	0.0	210.0	**195636.1**	171.4
91	48.0	24.0	8.0	4.0	0.0	210.0	**195636.1**	171.4
92	48.0	24.0	12.0	4.0	0.0	210.0	**195636.1**	171.4
93	48.0	24.0	12.0	6.0	0.0	200.0	193688.0	166.9
94	56.0	28.0	14.0	7.0	0.0	180.0	192751.2	146.9
95	64.0	16.0	8.0	4.0	0.0	185.0	**195698.1**	140.9
96	64.0	32.0	8.0	4.0	0.0	185.0	**195698.1**	140.9
97	64.0	32.0	16.0	4.0	0.0	185.0	**195698.1**	140.9
98	64.0	32.0	16.0	8.0	0.0	165.0	191802.0	131.9
99	64.0	32.0	16.0	8.0	4.0	185.0	**195698.1**	140.9

Finally, we may draw the following **conclusions** from the figures and tables:

1. The strategy in which no iterations are used (a single ELISA test) is dominated by strategies using iterations.
2. A strategy with at most one recycle is frequently dominated by strategies with more iterations.
3. The five-cycle strategy is best, although the difference with respect to second-best is almost negligible.
4. The correlation between the maximum number of ELISA tests and the profit induced by each of the iterations ($h^0 = 1, 2, ..., 5$) seems to be strictly positive and rather large.
5. The actual choice of the group sizes makes a significant difference.
6. If we impose an ELISA cost constraint, then we simply need to cut off a part of the graph by the vertical line $x = C_0/C_{elisa}$.
7. If we impose a time constraint (i.e., a constraint on the total ELISA testing time), we need to cut off the graph by a vertical line accordingly.

7 Conclusions

We have proposed a new testing procedure that we called Recycled Incomplete Identification Procedure (RIIP). In RIIP, groups of pooled blood units which are found contaminated in a so-called ELISA test are divided into smaller subgroups and again group-tested by ELISA, and so forth, until finally a so-called PCR test is conducted for all items in those subgroups which are found clean.
We have introduced an underlying unobservable Markov chain that captures the essence of the testing procedure, and we have determined its complete distribution in closed form. we have shown that the distributions of the main quantities of interest in the testing procedure, like the number of groups tested and the number of items found clean, can be expressed in terms of this Markov chain.
In view of the complexity of the resulting expressions, we have proposed simple approximations for the distributions of the main quantities of interest, based on the observations that the distribution of the number of contaminated items is approximately (with high accuracy) Poisson and that the probability that a group contains more than one contaminated item is negligibly small. These observations allow us to accurately evaluate a profit objective function for a wide range of different testing strategies. The numerical evaluation gives considerable insight into these strategies. In particular, our numerical experiments strongly suggest that, for realistic parameter values, it is profitable to do one or more re-tests of contaminated groups at the ELISA station. The actual group sizes also make a significant difference.

Acknowledgment The authors gratefully acknowledge several discussions with Professor Eilat Shinar, director of the Israeli Central Blood Bank. The research of Shaul Bar-Lev, Igor Kleiner, David Perry and Wolfgang Stadje was supported in part by grant No. I-1184-31.4/2012 from the German-Israel Science Foundation and grant No. 1071/14 from the Israel Science Foundation.

References

[1] Abolnikov, L. and Dukhovny, A. (2003) Optimization in HIV screening problems, *Journal of Applied Mathematics and Stochastic Analysis*, 16, (4) 361–374.

[2] Balcioglu, B., Kopach, R. and Carter, M. (2008) Tutorial on a red blood cell inventory management system with two demand rates. *European Journal of Operational Research* 185, 1051-1059.

[3] Bar-Lev, S.K., Boneh, A. and Perry, D. (1990) Incomplete Identification Models for Group-Testable Items. *Naval Research Logistics*, 37, 647-659.

[4] Bar-Lev,S.K., Parlar, M. and Perry, D. (1995) Optimal Sequential Decisions for Incomplete Identification of Group-Testable Items. *Sequential Analysis*, 14(1), 41-57.

[5] Bar-Lev, S.K., Stadje, W. and Van der Duyn Schouten, F.A. (2003) Hypergeometric group testing with incomplete information. *Prob. Engin. Inf. Sci.*, 17, 335-350.

[6] Bar-Lev, S.K., Stadje, W. and Van der Duyn Schouten, F.A. (2004) Optimal group testing with processing times and incomplete identification. *Methodology Comp. Prob.*, 6, 55-72.

[7] Bar-Lev, S.K., Stadje, W. and Van der Duyn Schouten, F.A. (2005) Multinomial group testing models with incomplete identification. *Journal of Statistical Planning and Inference*, 135, 384-401.

[8] 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.

[9] Bar-Lev, S.K., Parlar, M., Perry, D. and Van der Duyn Schouten, F.A (2007) Application of bulk queues to group testing models with incomplete identification. *European Journal of Operational Research*, 183(1), 226-237.

[10] Bar-lev, S.K., Blanc, J.P.C. , Boxma, O.J., Janssen, G. and Perry, D. (2013) Tandem Queues with Impatient Customers for Blood Screening Procedures. *Methodology Comp. Prob.*, 15, 423-451.

[11] Chick, S.E. (1996) Bayesian models for limiting dilution assay and group test data. *Biometrics*, 52, 1055-1062.

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

[13] 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.

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

[15] 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.

[16] Hourfar, M.K., Jork, C., Schottstedt, V., Weber-Schehl, M., Brixner, V., Busch, M.P., Geusendam, G., Gubbe, K., Mahnhardt, C., Mayr-Wohlfart, U., Pichl, L., Roth, W.K., Schmidt, M., Seifried, E. and Wright, D.J. (2008) Experience of German Red Cross blood donor services with nucleic acid testing: results of screening more than 30 million blood donations for human immunodeficiency virus-1, hepatitis C virus, and hepatitis B virus. *Transfusion*, 48(8), 1558-1566.

[17] Hughes-Oliver, J. and Rosenberger, W. (2000) Efficient estimation of the prevalence of multiple rare traits. *Biometrika*, 87, 315–327.

[18] Johnson, W.O. and Gastwirth, J.L. (2000) Dual group screening. *J. Statist. Planning Infer.*, 83, 449-473.

[19] Kantanen, M.L., Koskela, P. and Leinikki, P. (1996) Unlinked anonymous HIV screening of pregnant women in low-prevalence population. *Scand. J. Infectious Dis.*, 28, 3-7.

[20] Kline, R.L., Frother, T.A., Brookmeyer, R. et al. (1989) Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *J. Clin. Microbiol.*, 27, 1449-1452.

[21] Litvak, E., Tu, X.M. and Pagano, M. (1994) Screening for the presence of a disease by pooling sera samples. *J. Amer. Statist. Assoc.*, 89, 424-434.

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

[23] Macula, A.J. (1999) Probabilistic nonadaptive and two-stage group testing with relatively small pools and DNA library screening. *J. Combin. Optim.*, 2, 385-397.

[24] Monzon, O.T., Paladin, F.J.E., Dimaandal, E., Balis, A.M., Samson, C. and Mitchell, S. (1992) Relevance of antibody content and test format in HIV testing of pooled sera. *AIDS*, 6, 43-48.

[25] Schottstedt, V., Tuma, W., Bünger, G., and Lefèvre, H. (1998). PCR for HCV and HIV-1 experiences and first results from routine screening programme in a large blood transfusion service. *Biologicals*, 26, 101-104.

[26] Sebastian, H.W., Yates, N., Wilding, R. and Cotton, S. (2012) Blood inventory management: Hospital best practice, *Transfusion Medicine Reviews*, 26, 153-163.

[27] Steiner, M.E., Assmann, S.F., Levy, J.H., Marshall, J., Pulkrabek, S., Sloan, S.R., Triulzi, D. and Stowell, C.P. (2010) Addressing the question of the effect of RBC storage on clinical outcomes: the Red Cell Storage Duration Study (RECESS) (Section 7). *Transfus Apher Sci.* 43, 107-116.

[28] Stramer, S.L., Glynn, S.A., Kleinman, S.H., Strong, D.M., Caglioti, S., Wright, D.J., Dodd, R.Y. and Busch, M.P. (2004) Detection of HIV-1 and HCV infections among antibody-negative blood donors by nucleic acid-amplification testing. *The New England Journal of Medicine*, 351(8), 760-768.

[29] Tebbs, J.M., McMahan, C.S. and Bilde, C.R. (2013) Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project. *Biometrics*, 69 (4), 1064-1073.

[30] 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.

[31] 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. *Amer. J. Human Genet.*, 69, 1290-1300.

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

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

[34] Wolf, J. (1985) Born again group testing: multiaccess communications. *IEEE Transactions on Information Theory*, 31, 185-191.

[35] Yamamura, K. and Ishimoto, M. (2009) Optimal sample size for composite sampling with subsampling, when estimating the proportion of pecky rice grains in a field. *Journal of Agricultural, Biological, and Environmental Statistics*, 14, 135-153.

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