Group testing: revisiting the ideas*

Viktor Skorniakov[1], **Remigijus Leipus**[1], **Gediminas Juzeliūnas**[2], **Kęstutis Staliūnas**[3,4,5]

[1] Institute of Applied Mathematics, Faculty of Mathematics and Informatics, Vilnius University, Naugarduko 24, Vilnius LT-03225, Lithuania.

[2] Institute of Theoretical Physics and Astronomy, Vilnius University, Saulėtekio 3, Vilnius LT-10257, Lithuania.

[3] Laser Reseach Center, Faculty of Physics, Vilnius University, Saulėtekio 9, Building 3, Vilnius LT-10222, Lithuania.

[4] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona 08010, Spain.

[5] Departamento de Física, Universitat Politècnica de Catalunya, Campus Diagonal Nord, Edifici B5. C. Jordi Girona, 1-3, Barcelona 08034, Spain.

Received: date / Revised: date / Published online: data

Abstract

The task of identification of randomly scattered 'bad' items in a fixed set of objects is a frequent one, and there are many ways to deal with it. 'Group testing' (GT) refers to the testing strategy aiming to effectively replace the inspection of single objects by the inspection of groups spanning more than one object. First announced by Dorfman in 1943, the methodology has underwent vigorous development, and though many related research still takes place, the ground ideas remain the same. In the present paper, we revisit two classical GT algorithms: the Dorfman's algorithm and the halving algorithm. Our fresh treatment of the latter and expository comparison of the two is devoted to dissemination of GT ideas which are so important in the current COVID-19 induced pandemic situation.

Keywords: Group testing; Quick Sort algorithm; COVID-19.

1 Introduction

The task of identification of bad items in a given set of objects arises quite often. For example, consider identification of: a) the infected patients in a fixed cohort; or b) the defective items in the production batch. Usually, this identification task is a composite problem and spans many subtasks. One of such subtasks can be described as 'an efficient utilization of resources devoted to testing of investigated objects'. It turns out that, among plenitude of context dependent methods designed for the solution of this subtask, the appropriately chosen testing plan plays

*Second author is supported by grant S-COV-20-4 from the Research Council of Lithuania. Early version of the paper was entitled "Optimal group based testing strategy for detecting infected individuals: comparison of algorithms"

an exceptional role since it alone can reduce the testing costs substantially. This is the contextual target of the present paper. To be more precise, we focus on testing strategies widely known under the name of Group (or Pooled Sample) Testing (in what follows, we make use of an abbreviation GT). The core idea underlying GT strategy is an observation that, in many cases, the testing of single items can be replaced by the testing of a group spanning more than one item. Though it is difficult to trace back the exact date and inventor of this cornerstone idea (for a good historical account see [13], Ch. 1), without doubt much of the credit goes to the pioneering work of Dorfman [12]. In that paper, the Blood Testing problem was described and the following scheme was suggested. Given N individual blood samples, pool them and test for the presence of an infection in the pooled sample; in case of the negative test – finish; in case of the positive test – retest each single patient. The rationale behind is clear: if the prevalence of the infection is low, one usually ends up with a single test applied to the pool instead of N tests applied individually.

Since appearance of Dorfman's work [12] in 1943, GT ideas were evolving in many directions and found important applications in molecular biology, quality control, computer science and other fields. Digging into the literature, one can observe that it is indeed very widespread across different disciplines. Because of this reason, some developments were overlapping and rediscovered by researchers working in the different fields. Our personal familiarity with the field also underwent this route: attracted by potential applications in the context of COVID-19 epidemics, we have rediscovered some well known facts. Nonetheless, the attained experience and understanding of the importance of the tool inspired us to write a promotional paper on the topic. This is the main intent of the paper: we believe that, in the current pandemic situation, the spread of GT ideas and attraction of other researchers to the field is an important and meaningful task. We do not propose novel GT schemes or methodological improvements. Our presentation is primarily devoted to those unfamiliar with the subject aiming to provide a quick lightweight introduction 'by example' without delving into details yet giving a flavor of the topic as a whole. Choosing a mathematical journal we, first of all, were interested in the dissemination within the *mathematically oriented* community. Secondly, while getting familiar with the topic, we have encountered a lot of papers where the subject was treated without sufficient mathematical rigorousness. We therefore felt that our rigorous treatment of the GT Scheme **H** (see Section 2), unseen (or at least unobserved) by us, was a missing item in the existing literature. Finally, after submission of the initial version of the paper, we have discovered that our Proposition 2.2 adds some new information to what is known about classical Dorfman's scheme (see the comments in Section 2).

The remaining part of the paper is organized as follows. In Section 2, we provide some preliminaries, then describe and contrast two classical GT schemes. In Section 3, we give an accompanying discussion highlighting some relevant issues and skim through the related literature. Appendices A and B contain some mathematical derivations and tables.

Because of COVID and the exemplary nature, we attach the whole presentation to the biomedical context.

2 Two classical GT schemes

Consider the following setup. Assume that the prevalence of some disease (the fraction of the infected individuals) is equal to $p \in (0, 1)$ in an infinite (or large enough) population. A cohort spanning N independent individuals has to be tested and infected patients have to be identified. To achieve the goal, samples are collected from each individual. The applied test performs equally well for individual and for pooled samples: a situation might occur, e.g., when the test indicates the presence of the infection in the blood sample and there is no difference whether the latter is obtained from a single individual or from a pooled cohort of samples. For the situation described, physicians can choose different testing strategies. Let us assume that the following are three possible choices[†].

Scheme A: Test each patient's sample.

Scheme D: Conduct testing of the pooled sample. Test each member of the cohort separately only in case of detected infection in the pooled sample.

Scheme H:

Step 1. Test pooled sample of the whole cohort. Proceed to *Step 2.*

Step 2. If the test is positive, proceed to *Step 3*, otherwise, finish testing cohort.

Step 3. Divide the cohort into two parts consisting of the first and second halves respectively. Apply the whole algorithm to the two obtained parts recursively.

Although it is not obvious at the first glance, schemes **D** and **H** can be much more efficient as compared to **A** provided prevalence p is low enough. To give a rigorous justification (along with the concept of *efficiency*), let us formally define the underlying model.

Consider the sample of N individuals. Put $X_i = 1$, provided the test of i-th individual is positive, and $X_i = 0$, otherwise. Let $S = S_N = X_1 + \cdots + X_N$ be the total number of infected individuals in the sample and let $T = T_N$ be the total number of tests applied to the cohort.

We start with the scheme **D**. The test is applied once if the result is negative, and it is further applied to each of N individuals otherwise, i.e.

$$T \quad = \quad 1 + N \cdot \mathbb{1}\{S > 0\}.$$

The above implies that $X_1, \ldots, X_N$ are independent identically distributed (i.i.d.) random variables each having Bernoulli distribution $\mathrm{Be}(p)$. Therefore, S has the

[†]The notations for a short reference of GT schemes come from Dorfman (**D**); Halving (**H**). **A** reflects the most naive and straightforward option.

Binomial distribution $\mathrm{Bin}(N,p)$. Consequently, an average number of tests per cohort is

$$\mathrm{E}T \;=\; 1+N\mathrm{P}(S>0) \;=\; 1+N(1-q^N), \tag{1}$$

where $q := 1-p$. An average number of tests per individual, say $t = t(N)$, is

$$t(N) \;=\; \frac{\mathrm{E}T}{N} \;=\; \frac{1}{N}+1-q^N. \tag{2}$$

Consider a function $t\colon (0,\infty)\mapsto(0,\infty)$ given in (2). By equating its derivative to 0, we see that the stationary points solve equation

$$\frac{1}{N^2} \;=\; -q^N\ln q \;\text{ or, equivalently, }\; N \;=\; q^{-N/2}\Big(\ln\frac{1}{q}\Big)^{-1/2}, \tag{3}$$

which is a fixed point equation for $g(N) = q^{-N/2}\big(\ln(1/q)\big)^{-1/2}$, $N\in(0,\infty)$, and, hence, can be easily solved iteratively. It is further not difficult to prove that, for p in the region enclosing $(0,0.2)$, there exists a unique solution $N_p>0$ of (3) which is a minimizer of $t(N)$ (see Proposition 2.2 below). Then, turning back to economic/biomedical interpretation, we conclude that, having a cohort of $\lfloor N_p\rfloor$ (here and in the sequel, $\lfloor y\rfloor$ stands for an integer part of $y\in\mathbb{R}$) individuals, Scheme **D** results in a lowest average number of tests per person which is possible when applying scheme of this type for a population having prevalence p. Scheme **A**, in contrast, always has a constant number of tests 1 per person. Therefore, an average (absolute) gain attained applying Scheme **D** instead of Scheme **A** is given by the difference

$$G_p \;=\; 1-t(N_p) \;=\; q^{N_p}-\frac{1}{N_p}. \tag{4}$$

Right panel in Figure 1 shows the graph of $p\mapsto 100G_p$, $p\in(0,0.2)$, which is an average gain measured by the number of tests saved per 100 individuals. The corresponding values are provided in Table 1 (see Appendix B). An accompanying graph of $p\mapsto N_p$ (see the left panel of the Figure 1) demonstrates dependence of an optimal sample size on p. To obtain a fast numerical evidence, assume that N is bounded away from zero and $pN\to 0$. Then from (3) it follows that the optimal sample size satisfies

$$N \;\sim\; \frac{1}{\sqrt{p}} \quad\text{and}\quad t(N)=\frac{1}{N}+1-(1-p)^N \;\sim\; \frac{1}{N}+pN \;\sim\; 2\sqrt{p}.$$

Hence, assuming that p is small enough for $pN\approx 0$ to hold, the above implies that

$$G_p \;\approx\; 1-2\sqrt{p}.$$

For example, if $p=0.01$, then we have $G_p\approx 0.8$, i.e. an approximate average gain is 80% or so.

Now let us switch to the Scheme **H**. Its main features are summarized in the following proposition (for the proof see Appendix A).

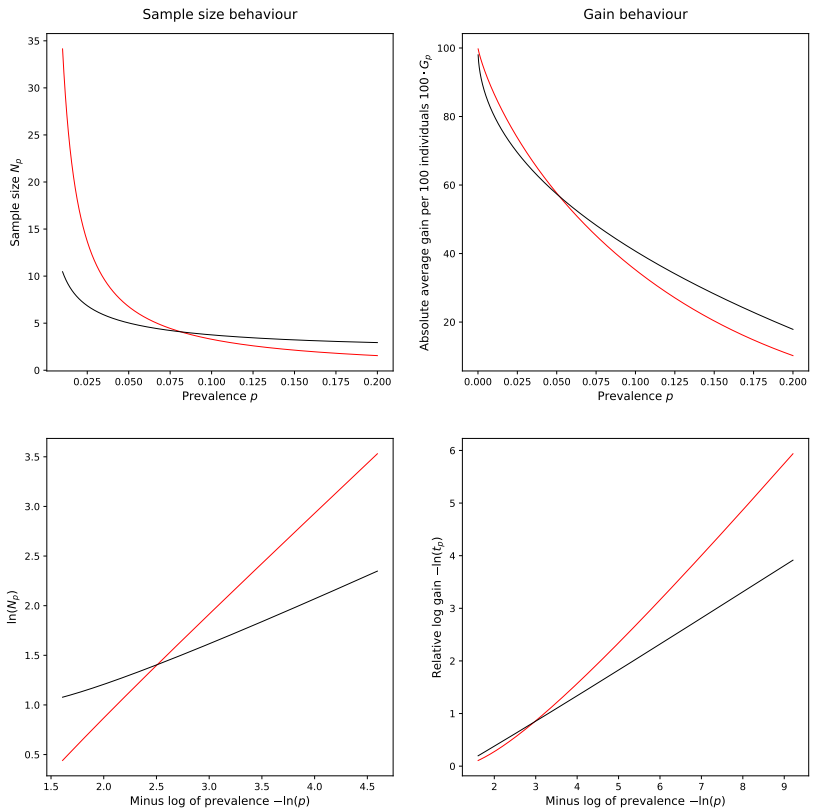

Figure 1: Scheme **H** (red) vs. Scheme **D** (black) on the linear and the log-log scale.

Proposition 2.1. *Assume the Scheme **H**. Then*
(i) *an average number of tests per person is given by*

$$t(N) \;=\; \frac{1}{N} + 2\sum_{k=1}^{\log_2 N} \frac{1-q^{2^k}}{2^k} \;=\; \frac{1}{N} + 2\log_2 N \int_0^1 \Big(\int_q^1 x^{2^{1+\lfloor v \log_2 N \rfloor}-1} \mathrm{d}x \Big) \mathrm{d}v; \tag{5}$$

(ii) *an average number of tests per person in the case of an infinitely large cohort*

is

$$t(\infty) \;=\; \lim_{N\to\infty} t(N) \;=\; 2\sum_{k=1}^{\infty} \frac{1-q^{2^k}}{2^k};$$

(iii) *for a fixed* $p \in (0,1)$*, function* $t : \mathbb{N} \mapsto (0,\infty)$ *admits at most two minimizers* N_p*: the value* $N = N_p$ *corresponding to optimal sample size is either* $\lfloor \frac{1}{2\log_2(1/q)} \rfloor$ *or* $\lfloor \frac{1}{2\log_2(1/q)} \rfloor + 1$.

Inspection of the results in the statement of the proposition leads to a quick comparison of Scheme **H** with **A** and **D**. Indeed, consider first the limit in (ii). Obviously,

$$t(\infty) \;\leq\; 2\Big(1 - \frac{q^2}{2} - \frac{q^4}{4}\Big).$$

Hence, for $q \approx 1$ (or alternatively $p \approx 0$), $t(\infty) < 1$. The latter means that, when the prevalence is low, this scheme always outperforms common sequential Scheme **A**. Again, to gain a quick quantitative insight, assume that p is small enough for $pN \approx 0$ to hold. Then turning to (iii) and taking a 'continuous' (undiscretized) version of N_p equal to $\frac{\ln 2}{2\ln(1/q)}$ yields relationships (see Remark A.1, eq. (9))

$$N_p \;\approx\; \frac{\ln 2}{2p} \quad \text{and} \quad t(N_p) \;\approx\; \frac{2p}{\ln 2} + 2p\log_2\Big(\frac{\ln 2}{2p}\Big) \;\approx\; -2p\log_2 p. \tag{6}$$

Therefore, an approximation to an average gain $G_p = 1 - t(N_p)$ is $1 + 2p\log_2 p$. Taking, e.g., $p = 0.01$ results in $G_{0.01} \approx 0.867$. Considering analogous example given for Scheme **D**, we see that the gain has an increase close to 7%. In fact, this is not surprising (for a visual comparison of Schemes **D** and **H** on the linear and the log-log scales, see Figure 1, and, for the numerical one, Table 1 and Table 2 in Appendix B) since, for Scheme **D**, we had $G_p \approx 1 - 2\sqrt{p}$ and $p\log_2 p/\sqrt{p} \to 0$ as $p \to 0$. Equality (6), however, exhibits some magic flavor. To see this, note that, for $p \approx 0$, entropy I_p of $X \sim \mathrm{Be}(p)$ is asymptotically equivalent to $p\log_2 p$ since

$$\begin{aligned} I_p \;&=\; p\log_2 p + (1-p)\log_2(1-p) \;=\; p\log_2 p - p(1-p) + o(p) \\ &=\; p\log_2 p(1 + o(1)). \end{aligned}$$

Consequently, (6) means that the optimal average number of tests per one individual scales like entropy of the prevalence of the infection. Keeping in a view the above relationship, it is not surprising that the significant number of works ([8], [19], [1]) have approached the testing problem from the Information Theory perspective. In the next section we provide additional comments regarding connections with the Information Theory. Here we end up with the previously mentioned Proposition 2.2, which is proved in Appendix A.

Proposition 2.2. *Let $p \in A := (0, 1-\mathrm{e}^{-4/\mathrm{e}^2})$ be fixed. Consider function $g_p(N) = g(N) = q^{-N/2}(\ln(1/q))^{-1/2}$, $N > 0$. It admits a unique fixed point N_p which minimizes $t(N)$ given by* (2).

The above proposition can be viewed as a counterpart of Proposition 2.1. Note, that it does not contain analytical expression of an optimal sample size. The latter was given by Samuels [35] and is either $1 + \lfloor p^{-1/2} \rfloor$ or $2 + \lfloor p^{-1/2} \rfloor$. Of most importance is that Samuels [35] not only provided the analytical expression of the optimal sample size but has also shown that, for the case of Scheme **D**, the optimal sample size equals to 1 for $p > 1 - (1/3)^{1/3} \approx 0.31$. This, in turn, is in agreement with the fundamental fact of the GT theory discovered by Ungar [42]: if $p \geq \frac{3-\sqrt{5}}{2} \approx 0.38$, then there does not exist an algorithm that is better than individual one-by-one testing.

An interesting detail here is that our proof given in Appendix A differs from that of Samuels and leads to an exact analytical expression for the range of p (the set A above), where $g_p(N)$ has a unique minimizer.

3 Discussion

Since its appearance, the Dorfman's Scheme **D** was rigorously investigated by many authors (we refer to [40], [39], [23], [35] to name a few). Talking about Scheme **H**, the situation is a bit different. To our best knowledge, the reference [46] is the only work close to ours both in nature of investigations and results. However, in that paper, the authors focus on the treatment of an asymptotic regime of Scheme **H** when $p \to 0$. Majority of other references encountered by us provide instructions suitable for the practical application of Scheme **H** with a brief and nonrigorous theoretical background. For example, in the present context, it was currently afresh discussed by Gollier and Gossner [18], Mentus et al. [29] and Shani-Narkiss et al. [37]. For an older reference discussing the case of nonhomogeneous population (i.e., the one in which the probability of being infected p may vary across individuals) and containing quite a large body of applied literature on halving algorithm (i.e., Scheme **H**) we refer to [4].

One should have noticed that halving, constituting the core of the Scheme **H**, yields another link to the Information and Algorithm Theory in addition to the one already mentioned[‡] in Section 2. Namely, in its essence, the Scheme **H** is nothing more than the Quick Sort (QS) algorithm designed to sort a set containing keys of two types (bad and good ones). It is well known that QS yields the best (up to the constant multiple) possible average performance among comparison-based algorithms: to sort an array having N non-constant (i.e., random) items, the smallest average number of comparisons is of the order $N \ln N$ [10], and all

[‡]the discussed appearance of entropy in formula (6), in fact, is a simple conclusion following from Shannon's Coding Theory; a bit more on the connections with that theory can be found in the Appendix H of the Supplementary Material of the reference [27]

randomized 'divide-and-conquer' type algorithms (with QS being one among the rest) have expected time asymptotically equivalent to QS, which randomly splits sorted set into two equal subsets [11]. Our formula (5) is just a confirmation of the well known fact. To see this, note that, in the context of sorting task, (5) presents an average number of comparisons per item. Though the order is correct, we are inclined to think that the multiplier $\int_0^1 \big(\int_q^1 x^{2^{1+\lfloor v \log_2 N \rfloor}-1} \mathrm{d}x \big) \mathrm{d}v$ can be improved by making use of QS modification (or another comparison-based algorithm) designed to sort items with a small number of possible values (in our case, there are just two values: 'sick' and 'healthy'). On the other hand, as already mentioned above, the order is optimal since though there are algorithms which can beat QS when sorting integers, e.g., [2], [41], they operate under different, i.e., noncomparison-based, mode. In our case, however, comparison is predefined by the setting of the problem at hand: we assume that biomedical tests can only be carried out by making use of comparison.

Though biomedical context is very frequent in applications, there are many others including engineering, environmental sciences, information theory, etc. (see [32], [21], [22], [30], [26], [15], [24], [28], [25], [6]). This 'real life' contextual diversity brings many constraints to take into account despite the fact that the standard binomial setting, considered in Section 2, quite often can be regarded as a good starting approximation. To get a full understanding of the matter touched, below is a shortlist of key issues with a brief description of each.

- Heterogeneity of population. The prevalence of disease may depend on other factors (e.g., age and gender).
- Imperfectness of the test. The test can have sensitivity and/or specificity below 1.
- Dilution effect. Pooling can reduce testing accuracy substantially. If this is the case, it is necessary to impose upper bound on the number of pooled samples.
- Implementation costs. In Section 2, we silently assumed that implementation of the considered schemes only involves retesting related costs. However, it may involve others as well.
- Dependence. It can happen that tested individuals are somehow related.

All these underpinnings have to be addressed carefully. Take, for example, the last one. From results presented in Section 2, one can infer that the application of the GT procedures is most effective when the prevalence is low ($p \approx 0$). In such case, under classical assumption $pN \to \lambda > 0$, the number of infected individuals S_N can be well approximated by the Poisson distribution $\mathrm{Pois}(pN)$, and the approximation remains quite accurate irrespectively of the nature of the dependence exhibited by summands (see [7], [43], [3] and references therein for

results of this kind with possible extensions beyond the classical setting). It is therefore reasonable to assume that, after switching to Poisson approximation, at least some of the existing schemes can be carried over to the dependent case. Clearly, additional restrictions call for new theoretical investigations.

The set of directions of such investigations can be significantly appended by including other methods and GT related tasks. More concretely, the schemes considered in Section 2 broadly fall into the class of probabilistic GT schemes. Another widely adopted paradigm is called *combinatorial approach*. Within its framework, one does not assume any random mechanism and tries to make use of combinatorial methods in order to identify d bad items in the given group of $N \geq d$ objects (see monographs [13], [14]). Speaking about tasks, up to now, we have focused only on the identification of bad items (or infected patients) under assumption that the prevalence p is known. In addition to the literature devoted to this task, there is a huge body of literature dealing with an estimation (both point and interval) of p from pooled samples observations as well as testing issues (see, e.g., [34], [16], [20] and references therein).

We hope that our discussion complies well with our initial goal stated in the introduction. To emphasize the relevance of similar promotional discussions in the present context, we point out a huge burst of papers devoted to similar problems (see, e.g., [5], [17], [31], [33], [36], [38], [44], [45]). Besides that, we also note that some countries have already successfully applied pooling methodology for testing of the SARS-CoV-2 virus.[§]

A Appendix. Technical details

Proof of Proposition 2.1. (i) For $1 \leq i \leq j \leq N = 2^n$, let $M_{ij} = \{X_i, X_{i+1}, \ldots, X_j\}$ and let $S(i,j) = \sum_{k=i}^{j} X_k$ be the number of infected individuals in the cohort M_{ij}. Let $\mathbb{1}_{i,j}$ denote an indicator function taking value 1 if there is at least one infected individual in the group $M_{i,j}$, i.e. $\mathbb{1}_{i,j} = \mathbb{1}\{S(i,j) > 0\}$. Also, let $T(i,j)$ be *the total number of tests applied to the cohort M_{ij} after the initial pooled test*. By the description of the testing Scheme **H**, applying recursive

[§]According to Wikipedia [9], 'In Israel, researchers at Technion and Rambam Hospital developed a method for testing samples from 64 patients simultaneously, by pooling the samples and only testing further if the combined was positive. Pool testing was then adopted in Israel, Germany, Ghana, South Korea, Nebraska, China and the Indian states Uttar Pradesh, West Bengal, Punjab, Chhattisgarh, and Maharashtra.' Also see 'List of countries implementing pool testing strategy against COVID-19' therein.

equations, we have

$$
\begin{aligned}
T &= 1+T(1,N) \;=\; 1+\mathbb{1}_{1,N}(1+T(1,N/2))+\mathbb{1}_{1,N}(1+T(N/2+1,N)) \\
&= 1+2\cdot\mathbb{1}_{1,N}+\mathbb{1}_{1,N}\big(T(1,N/2)+T(N/2+1,N)\big) \\
&= 1+2\cdot\mathbb{1}_{1,N}+\mathbb{1}_{1,N}\big(2\cdot\mathbb{1}_{1,N/2}+\mathbb{1}_{1,N/2}\big(T(1,N/4)+T(N/4+1,N/2)\big) \\
&\quad +2\cdot\mathbb{1}_{N/2+1,N}+\mathbb{1}_{N/2+1,N}\big(T(N/2+1,N/2+N/4)+T(N/2+N/4+1,N)\big)\big) \\
&= 1+2\cdot\mathbb{1}_{1,N}+2(\mathbb{1}_{1,N/2}+\mathbb{1}_{N/2+1,N})+\mathbb{1}_{1,N/2}\big(T(1,N/4)+T(N/4+1,N/2)\big) \\
&\quad +\;\mathbb{1}_{N/2+1,N}\big(T(N/2+1,N/2+N/4)+T(N/2+N/4+1,N)\big) \\
&= \;\ldots\; = \;1+2\cdot\mathbb{1}_{1,N}+2\big(\mathbb{1}_{1,N/2}+\mathbb{1}_{N/2+1,N}\big)+{+}2\big(\mathbb{1}_{1,N/4}+\mathbb{1}_{N/4+1,N/2} \\
&\quad +\;\mathbb{1}_{N/2+1,N/2+N/4}+\mathbb{1}_{N/2+N/4+1,N}\big)+\cdots+2\big(\mathbb{1}_{1,2}+\mathbb{1}_{3,4}+\cdots+\mathbb{1}_{N-1,N}\big).
\end{aligned}
$$

Taking expectations yields

$$
\begin{aligned}
\mathrm{E}T &= 1+2\sum_{j=0}^{n-1}2^j\mathrm{P}\{S(1,2^{n-j})>0\} \;=\; 1+2\sum_{j=0}^{n-1}2^j(1-q^{2^{n-j}}) \\
&= 1+2\cdot 2^n\sum_{k=1}^{n}\frac{1-q^{2^k}}{2^k}. \qquad (7)
\end{aligned}
$$

Hence, the first equality in (i). For the second one, take the last sum above and continue as follows:

$$
\begin{aligned}
\sum_{k=1}^{n}\frac{1-q^{2^k}}{2^k} &= \sum_{k=1}^{n}\int_q^1 x^{2^k-1}\mathrm{d}x \;=\; \sum_{k=1}^{n}\int_k^{k+1}\Big(\int_q^1 x^{2^{\lfloor y\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}y \\
&= \int_1^{n+1}\Big(\int_q^1 x^{2^{\lfloor y\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}y \;=\; n\int_0^1\Big(\int_q^1 x^{2^{1+\lfloor vn\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}v \\
&= \log_2 N\int_0^1\Big(\int_q^1 x^{2^{1+\lfloor v\log_2 N\rfloor}-1}\mathrm{d}x\Big)\mathrm{d}v.
\end{aligned}
$$

(ii) By (7),

$$
t(N) \;=\; \frac{\mathrm{E}T}{N} \;=\; \frac{1}{N}+2\sum_{k=1}^{\log_2 N}\frac{1-q^{2^k}}{2^k} \;\to 2\sum_{k=1}^{\infty}\frac{1-q^{2^k}}{2^k} \quad\text{as}\quad N\to\infty. \qquad (8)
$$

(iii) Since $N=N(n)=2^n$, by the second equality in (8),

$$
\Delta_n \;:=\; t(N(n+1))-t(N(n)) \;=\; \frac{1}{2N}-\frac{1}{N}+2\frac{1-q^{2N}}{2N} \;=\; \frac{1-2q^{2N}}{2N}.
$$

Clearly, $q^{2N}\downarrow 0$ as $N\to\infty$. Therefore, there exist no more than two $N_p\in\mathbb{N}$ such that $\Delta_n\le 0$ for all $N\le N_p$ and $\Delta_n\ge 0$ for all $N\ge N_p$, and $t(N_p)$ attains its minimal value at N_p. To find N_p, we first solve $1-2q^{2N}=0$ with respect to N, and then choose from the two nearest integers (i.e., $\lfloor N\rfloor,\lfloor N\rfloor+1$) the one which minimizes t_N. $\square$

Remark A.1. Note that, if $N \geq 1$ and $pN \to 0$, then for $t(N)$ in (5) it holds

$$t(N) \;=\; \frac{1}{N} + 2p\log_2 N\Big(1 + O\Big(\frac{pN}{\log_2 N}\Big)\Big). \tag{9}$$

To see this, it suffices to use the following bounds

$$1 - pn \;\leq\; (1-p)^n \;\leq\; 1 - pn + \frac{n(n-1)}{2}\,p^2, \quad n \geq 1.$$

Proof of Proposition 2.2. Step 1 (fixed points). Define

$$v \;:=\; \frac{2}{\ln(1/\sqrt{q})}, \quad f(N) \;:=\; N - g(N) = N - \frac{\sqrt{v}}{2}\,\mathrm{e}^{2N/v}. \tag{10}$$

Then derivative $f'(N) = 1 - \mathrm{e}^{2N/v}/\sqrt{v} = 0$ which is equivalent to $N = (v\ln\sqrt{v})/2$. Note that $f'(N) \to -\infty$ as $N \to \infty$. Moreover, $f'(0) > 0$ since $1 - (1/\sqrt{v}) > 0$ $\iff p < 1 - \mathrm{e}^{-4}$ which is satisfied for any $p \in A$. Therefore, f attains maximal value at

$$N_{\max} \;:=\; \frac{v\ln\sqrt{v}}{2} \tag{11}$$

and

$$f_{\max} \;=\; N_{\max} - \frac{\sqrt{v}}{2}\,\mathrm{e}^{\frac{2}{v}N_{\max}} \;=\; \frac{v}{2}\big(\ln\sqrt{v} - 1\big) > 0$$

since $\ln\sqrt{v} - 1 > 0 \iff p < 1 - \mathrm{e}^{-4/\mathrm{e}^2}$. On the other hand, $f(0) = -\sqrt{v}/2 < 0$ and $f(N) \to -\infty$. Consequently, f has two zeroes: $N_p \in (0, N_{\max})$ and $\widetilde{N}_p \in (N_{\max}, \infty)$. The latter means that g has two fixed points.

Step 2 (minimizer). In this step, we show that N_p from Step 1 is the minimizer for $t(N)$ given in (2). By (3),

$$t''(N_p) \;=\; \frac{2}{N_p^3} - q^{N_p}(\ln q)^2 \;=\; -q^{N_p}\ln q\Big(\frac{2}{N_p} + \ln q\Big).$$

Hence,

$$t''(N_p) > 0 \iff \frac{2}{N_p} + \ln q > 0 \iff \frac{v}{2} > N_p. \tag{12}$$

From Step 1 (see (11)) it follows that $N_{\max}/(v/2) = \ln\sqrt{v} > 1$, i.e. $v/2 \in (0, N_{\max})$. Therefore, (12) holds if and only if $f(v/2) > 0$. The latter reads as $(v - \mathrm{e}\sqrt{v})/2 > 0$ and is equivalent to $p < 1 - \mathrm{e}^{-4/\mathrm{e}^2}$ showing that N_p (being the critical point of t) is indeed the announced minimizer. Finally, note that the above analysis also implies that $\widetilde{N}_p$ from Step 1 is the maximizer of $t(N)$ which affirms the uniqueness of the minimizer. □

Remark A.2. One can also show that $p \mapsto N_p$ is strictly decreasing and continuous on A. However, the latter properties seem to be of less importance and we omit the details.

B Appendix. Tables

In the tables below, the following information is provided:

- Column 'N_p' shows an optimal sample size corresponding to p ranging in an interval given in the column 'Range of p'.
- Column 'Range of $100G_p$' shows an average gain (as defined in the main body of the paper) per 100 individuals corresponding to values of p and N_p given in the two leading columns. The highest gain corresponds to the lowest p in the corresponding interval. For example, in Table 1, the first line should be interpreted as follows: for $p \in [0.1865, 0.2000]$, optimal sample size N_p is equal to 2; if $p = 0.2000$, then average gain per 100 individuals $100G_p$ is equal to 16.1782; if $p = 0.1865$, then $100G_p = 14.000$; for intermediate values of p, the value of $100G_p$ lies in $[14.0000, 16.1782]$.

Table 1: Performance of Scheme **D**

N_p	Range of p	Range of $100G_p$	N_p	Range of p	Range of $100G_p$
2	0.1865 – 0.2000	14.0000 – 16.1782	22	0.0020 – 0.0021	90.9350 – 91.1457
3	0.0855 – 0.1864	20.5225 – 43.1472	23	0.0019 – 0.0019	91.3723 – 91.3723
4	0.0506 – 0.0854	44.9721 – 56.2450	24	0.0017 – 0.0018	91.6016 – 91.8321
5	0.0336 – 0.0505	57.1747 – 64.2917	25	0.0016 – 0.0016	92.0759 – 92.0759
6	0.0239 – 0.0335	64.8434 – 69.8233	26	0.0015 – 0.0015	92.3261 – 92.3261
7	0.0179 – 0.0238	70.1977 – 73.8374	27	0.0014 – 0.0014	92.5843 – 92.5843
8	0.0140 – 0.0178	74.1163 – 76.8337	28	0.0013 – 0.0013	92.8517 – 92.8517
9	0.0112 – 0.0139	77.0523 – 79.2489	29	0.0012 – 0.0012	93.1296 – 93.1296
10	0.0091 – 0.0111	79.4383 – 81.2637	30	0.0011 – 0.0011	93.4188 – 93.4188
11	0.0076 – 0.0090	81.4428 – 82.8596	32	0.0010 – 0.0010	93.7241 – 93.7241
12	0.0065 – 0.0075	83.0288 – 84.1396	33	0.0009 – 0.0009	94.0421 – 94.0421
13	0.0055 – 0.0064	84.2998 – 85.3889	35	0.0008 – 0.0008	94.3806 – 94.3806
14	0.0048 – 0.0054	85.5569 – 86.3428	38	0.0007 – 0.0007	94.7426 – 94.7426
15	0.0042 – 0.0047	86.5106 – 87.2152	41	0.0006 – 0.0006	95.1303 – 95.1303
16	0.0037 – 0.0041	87.3879 – 87.9915	45	0.0005 – 0.0005	95.5524 – 95.5524
17	0.0033 – 0.0036	88.1708 – 88.6533	50	0.0004 – 0.0004	96.0195 – 96.0195
18	0.0030 – 0.0032	88.8385 – 89.1800	58	0.0003 – 0.0003	96.5507 – 96.5507
19	0.0027 – 0.0029	89.3683 – 89.7296	71	0.0002 – 0.0002	97.1814 – 97.1814
20	0.0024 – 0.0026	89.9265 – 90.3079	100	0.0001 – 0.0001	98.0049 – 98.0049
21	0.0022 – 0.0023	90.5176 – 90.7183			

Table 2: Performance of Scheme **H**

N_p	Range of p	Range of $100G_p$	N_p	Range of p	Range of $100G_p$
1	0.1592 – 0.2000	10.2068 – 18.1573	59	0.0058 – 0.0058	91.4813 – 91.4813
2	0.1092 – 0.1591	18.1801 – 32.0493	60	0.0057 – 0.0057	91.5996 – 91.5996
3	0.0830 – 0.1091	32.0828 – 41.7999	61	0.0056 – 0.0056	91.7184 – 91.7184
4	0.0670 – 0.0829	41.8413 – 48.8858	62	0.0055 – 0.0055	91.8377 – 91.8377
5	0.0562 – 0.0669	48.9333 – 54.2777	64	0.0054 – 0.0054	91.9575 – 91.9575
6	0.0484 – 0.0561	54.3303 – 58.5384	65	0.0053 – 0.0053	92.0779 – 92.0779
7	0.0424 – 0.0483	58.5953 – 62.0600	66	0.0052 – 0.0052	92.1987 – 92.1987
8	0.0378 – 0.0423	62.1207 – 64.9241	67	0.0051 – 0.0051	92.3202 – 92.3202
9	0.0341 – 0.0377	64.9881 – 67.3443	69	0.0050 – 0.0050	92.4422 – 92.4422
10	0.0311 – 0.0340	67.4112 – 69.3911	70	0.0049 – 0.0049	92.5648 – 92.5648
11	0.0285 – 0.0310	69.4607 – 71.2322	72	0.0048 – 0.0048	92.6880 – 92.6880
12	0.0264 – 0.0284	71.3044 – 72.7691	73	0.0047 – 0.0047	92.8118 – 92.8118
13	0.0245 – 0.0263	72.8434 – 74.2009	75	0.0046 – 0.0046	92.9362 – 92.9362
14	0.0229 – 0.0244	74.2775 – 75.4396	76	0.0045 – 0.0045	93.0612 – 93.0612
15	0.0215 – 0.0228	75.5181 – 76.5498	78	0.0044 – 0.0044	93.1869 – 93.1869
16	0.0202 – 0.0214	76.6301 – 77.6042	80	0.0043 – 0.0043	93.3132 – 93.3132
17	0.0191 – 0.0201	77.6863 – 78.5153	82	0.0042 – 0.0042	93.4402 – 93.4402
18	0.0181 – 0.0190	78.5990 – 79.3593	84	0.0041 – 0.0041	93.5678 – 93.5678
19	0.0172 – 0.0180	79.4445 – 80.1325	86	0.0040 – 0.0040	93.6962 – 93.6962
20	0.0164 – 0.0171	80.2193 – 80.8312	88	0.0039 – 0.0039	93.8253 – 93.8253
21	0.0157 – 0.0163	80.9193 – 81.4518	91	0.0038 – 0.0038	93.9552 – 93.9552
22	0.0150 – 0.0156	81.5412 – 82.0814	93	0.0037 – 0.0037	94.0858 – 94.0858
23	0.0144 – 0.0149	82.1721 – 82.6285	96	0.0036 – 0.0036	94.2172 – 94.2172
24	0.0138 – 0.0143	82.7204 – 83.1829	98	0.0035 – 0.0035	94.3493 – 94.3493
25	0.0133 – 0.0137	83.2760 – 83.6506	101	0.0034 – 0.0034	94.4824 – 94.4824
26	0.0128 – 0.0132	83.7448 – 84.1237	104	0.0033 – 0.0033	94.6162 – 94.6162
27	0.0124 – 0.0127	84.2190 – 84.5063	108	0.0032 – 0.0032	94.7509 – 94.7509
28	0.0119 – 0.0123	84.6025 – 84.9897	111	0.0031 – 0.0031	94.8866 – 94.8866
29	0.0115 – 0.0118	85.0871 – 85.3808	115	0.0030 – 0.0030	95.0231 – 95.0231
30	0.0112 – 0.0114	85.4792 – 85.6767	119	0.0029 – 0.0029	95.1607 – 95.1607
31	0.0108 – 0.0111	85.7759 – 86.0750	123	0.0028 – 0.0028	95.2992 – 95.2992
32	0.0105 – 0.0107	86.1752 – 86.3764	128	0.0027 – 0.0027	95.4388 – 95.4388
33	0.0102 – 0.0104	86.4775 – 86.6804	133	0.0026 – 0.0026	95.5794 – 95.5794
34	0.0099 – 0.0101	86.7822 – 86.9868	138	0.0025 – 0.0025	95.7211 – 95.7211
35	0.0096 – 0.0098	87.0896 – 87.2959	144	0.0024 – 0.0024	95.8640 – 95.8640
36	0.0094 – 0.0095	87.3996 – 87.5035	150	0.0023 – 0.0023	96.0081 – 96.0081
37	0.0091 – 0.0093	87.6078 – 87.8172	157	0.0022 – 0.0022	96.1534 – 96.1534
38	0.0089 – 0.0090	87.9224 – 88.0279	164	0.0021 – 0.0021	96.3001 – 96.3001
39	0.0087 – 0.0088	88.1337 – 88.2398	173	0.0020 – 0.0020	96.4481 – 96.4481

Continued on next page

Table 2 – *Continued from previous page*

N_p	Range of p	Range of $100G_p$	N_p	Range of p	Range of $100G_p$
40	0.0085 – 0.0086	88.3463 – 88.4532	182	0.0019 – 0.0019	96.5976 – 96.5976
41	0.0083 – 0.0084	88.5603 – 88.6678	192	0.0018 – 0.0018	96.7486 – 96.7486
42	0.0081 – 0.0082	88.7757 – 88.8839	203	0.0017 – 0.0017	96.9012 – 96.9012
43	0.0079 – 0.0080	88.9924 – 89.1014	216	0.0016 – 0.0016	97.0555 – 97.0555
44	0.0077 – 0.0078	89.2107 – 89.3203	230	0.0015 – 0.0015	97.2116 – 97.2116
45	0.0076 – 0.0076	89.4303 – 89.4303	247	0.0014 – 0.0014	97.3696 – 97.3696
46	0.0074 – 0.0075	89.5408 – 89.6515	266	0.0013 – 0.0013	97.5297 – 97.5297
47	0.0072 – 0.0073	89.7627 – 89.8743	288	0.0012 – 0.0012	97.6920 – 97.6920
48	0.0071 – 0.0071	89.9863 – 89.9863	314	0.0011 – 0.0011	97.8567 – 97.8567
49	0.0070 – 0.0070	90.0987 – 90.0987	346	0.0010 – 0.0010	98.0241 – 98.0241
50	0.0068 – 0.0069	90.2115 – 90.3247	384	0.0009 – 0.0009	98.1943 – 98.1943
51	0.0067 – 0.0067	90.4383 – 90.4383	433	0.0008 – 0.0008	98.3677 – 98.3677
52	0.0066 – 0.0066	90.5524 – 90.5524	494	0.0007 – 0.0007	98.5448 – 98.5448
53	0.0064 – 0.0065	90.6669 – 90.7819	577	0.0006 – 0.0006	98.7260 – 98.7260
54	0.0063 – 0.0063	90.8973 – 90.8973	692	0.0005 – 0.0005	98.9120 – 98.9120
55	0.0062 – 0.0062	91.0132 – 91.0132	866	0.0004 – 0.0004	99.1039 – 99.1039
56	0.0061 – 0.0061	91.1295 – 91.1295	1155	0.0003 – 0.0003	99.3030 – 99.3030
57	0.0060 – 0.0060	91.2463 – 91.2463	1732	0.0002 – 0.0002	99.5119 – 99.5119
58	0.0059 – 0.0059	91.3636 – 91.3636	3465	0.0001 – 0.0001	99.7360 – 99.7360

Acknowledgment. We would like to thank two anonymous referees for insightful and constructive comments, which helped to improve the preliminary version of the paper.

References

[1] M. Aldridge, O. Johnson, J. Scarlett, Group testing: An information theory perspective, *Foundations and Trends® in Communications and Information Theory*, **15**(3–4):196–392, 2019.

[2] A. Andersson, T. Hagerup, S. Nilsson, R. Raman, Sorting in linear time?, *Journal of Computer and System Sciences*, **57**(1):74–93, 1998.

[3] A. D. Barbour, L. Holst, S. Janson, *Poisson Approximation*, Clarendon Press, Oxford, 1992.

[4] M. S. Black, C. R. Bilder, J. M. Tebbs, Group testing in heterogeneous populations by using halving algorithms, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**(2):277–290, 2012.

[5] A. Z. Broder, R. Kumar, A note on Double Pooling Tests, arXiv:2004.01684 [cs, math, stat], 2020, available from: `http://arxiv.org/abs/2004.01684`.

[6] T. Bui, M. Kuribayashi, M. Cheraghchi, I. Echizen, Efficiently decodable non-adaptive threshold group testing, *IEEE Transactions on Information Theory*, **65**:5519–5528, 2019.

[7] L. H. Y. Chen, Poisson approximation for dependent trials, *Annals of Probability*, **3**:534–545, 1975.

[8] P. Chen, L. Hsu, M. Sobel, Entropy-based optimal group-testing procedures, *Probability in the Engineering and Informational Sciences*, **1**:497–509, 1987.

[9] Wikipedia contributors, COVID-19 testing, 2021, available from: `https://en.wikipedia.org/wiki/COVID-19_testing`. [Online; accessed 14 February, 2021]

[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts, 3rd edition, 2009.

[11] B. C. Dean, A simple expected running time analysis for randomized "divide and conquer" algorithms, *Discrete Applied Mathematics*, **154**(1):1–5, 2006.

[12] R. Dorfman, The detection of defective members of large populations, *The Annals of Mathematical Statistics*, **14**(4):436–440, 1943.

[13] D. Du, F. Hwang, *Combinatorial Group Testing and its Applications*, Volume 12 of *Series on Applied Mathematics*, World Scientific, Singapore; River Edge, NJ, 2nd edition, 2000.

[14] D. Du, F. Hwang, *Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing*, Volume 18 of *Series on Applied Mathematics*, World Scientific, New Jersey, 2006.

[15] J. W. Fahey, P. J. Ourisson, F. H. Degnan, Pathogen detection, testing, and control in fresh broccoli sprouts, *Nutrition Journal*, **5**(13):1–6, 2006.

[16] European Centre for Desease Prevention and Control, Methodology for estimating point prevalence of SARS-CoV-2 infection by pooled RT-PCR testing, 2020.

[17] C. Gollier, Optimal group testing to exit the Covid confinement, Preprint, Toulouse School of Economics, 2020, available from: `https://www.tse-fr.eu/optimal-group-testing-exit-covid-confinement`.

[18] C. Gollier, O. Gossner, Group testing against Covid-19, *Covid Economics*, Issue **2**:32–42, 2020.

[19] L. Hsu, New procedures for group-testing based on the Huffman lower bound and Shannon entropy criteria, in N. Flournoy, W. F. Rosenberger (Eds.), *Adaptive Designs*, Volume 25 of *Lecture Notes – Monograph Series*, pp. 249–262, Institute of Mathematical Statistics, 1995.

[20] S.-H. Huang, M.-N. L. Huang, K. Shedden, W. K. Wong, Optimal group testing designs for estimating prevalence with uncertain testing errors, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(5):1547–1563, 2017.

[21] N. Johnson, S. Kotz, R. Rodriguez, Statistical effects of imperfect inspection sampling iii. Screening (group testing), *Journal of Quality Technology*, **20**:98–124, 1988.

[22] N. Johnson, S. Kotz, R. Rodriguez, Statistical effects of imperfect inspection sampling iv: Modified Dorfman screening procedures, *Journal of Quality Technology*, **22**:128–137, 1990.

[23] J.-K. Lee, M. Sobel, Dorfman and $R1$-type procedures for a generalized group-testing problem, *Mathematical Biosciences*, **15**:317–340, 1972.

[24] J. T. Lennon, Diversity and metabolism of marine bacteria cultivated on dissolved DNA, *Applied and Environmental Microbiology*, **73**(9):2799–2805, 2007.

[25] K. Li, D. Precup, T. J. Perkins, Pooled screening for synergistic interactions subject to blocking and noise, *PLoS ONE*, **9**(1), 2014.

[26] E. Litvak, X. M. Tu, M. Pagano, Screening for the presence of a disease by pooling sera samples, *Journal of the American Statistical Association*, **89**(426):424–434, 1994.

[27] Y. Malinovsky, P. S. Albert, Revisiting nested group testing procedures: New results, comparisons, and robustness, *The American Statistician*, **73**(2):117–125, 2019.

[28] S. May, A. Gamst, R. Haubrich, C. Benson, D. M. Smith, Pooled nucleic acid testing to identify antiretroviral treatment failure during HIV infection, *JAIDS Journal of Acquired Immune Deficiency Syndromes*, **53**(2):194–201, 2010.

[29] C. Mentus, M. Romeo, C. DiPaola, Analysis and applications of adaptive group testing method for COVID-19, `https://www.medrxiv.org/content/10.1101/2020.04.05.20050245v2`, 2020.

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

[31] L. Mutesa, P. Ndishimye, Y. Butera, J. Souopgui, A. Uwineza, R. Rutayisire, E. L. Ndoricimpaye, E. Musoni, N. Rujeni, T. Nyatanyi, E. Ntagwabira, M. Semakula, C. Musanabaganwa, D. Nyamwasa, M. Ndashimye, E. Ujeneza, I. E. Mwikarago, C. M. Muvunyi, J. B. Mazarati, S. Nsanzimana, N. Turok, W. Ndifon, A pooled testing strategy for identifying SARS-CoV-2 at low prevalence, Nature, **589**:276–280, 2021.

[32] M. S. Nagi, L. G. Raggi, Importance to "airsac" disease of water supplies contaminated with pathogenic Escherichia coli, *Avian Diseases*, **16**(4):718–723, 1972.

[33] K. R. Narayanan, A. Heidarzadeh, R. Laxminarayan, On accelerated testing for COVID-19 using group testing, arXiv:2004.04785v1 [cs.IT], 2020, available from: `http://arxiv.org/abs/2004.01684`.

[34] N. A. Pritchard, J. M. Tebbs, Estimating disease prevalence using inverse binomial pooled testing, *Journal of Agricultural, Biological, and Environmental Statistics*, **16**(1):70–87, 2011.

[35] S. M. Samuels, The exact solution to the two-stage group-testing problem, *Technometrics*, **20**:497–500, 1978.

[36] E. Seifried, S. Ciesek, Pool testing of SARS-CoV-2 samples increases test capacity, `https://eurekalert.org/pub_releases/2020-03/guf-pto033020.php`, 2020.

[37] H. Shani-Narkiss, O. D. Gilday, N. Yayon, I. D. Landau, Efficient and practical sample pooling for High-Throughput PCR diagnosis of COVID-19, `https://www.medrxiv.org/content/10.1101/2020.04.06.20052159v2`, 2020.

[38] N. Sinnott-Armstrong, D. L. Klein, B. Hickey, Evaluation of group testing for SARS-CoV-2 RNA, `https://www.medrxiv.org/content/10.1101/2020.03.27.20043968v1`, 2020.

[39] M. Sobel, Optimal group testing, in *Proceedings of the Colloquium on Information Theory*, Organized by the Bolyai Mathematical Society, Debrecen (Hungary), 1967, pp. 411–488.

[40] M. Sobel, P. A. Groll, Group testing to eliminate efficiently all defectives in a binomial sample, *Bell System Technical Journal*, **38**:1179–1252, 1959.

[41] M. Thorup, Randomized sorting in $O(n \log \log n)$ time and linear space using addition, shift, and bit-wise Boolean operations, *Journal of Algorithms*, **42**(2):205–230, 2002.

[42] P. Ungar, The cutoff point for group testing, *Communications on Pure and Applied Mathematics*, **13**(1):49–54, 1960.

[43] V. Čekanavičius, *Approximation Methods in Probability Theory*, Springer, New York, 2016.

[44] J. Žilinskas, A. Lančinskas, M. R. Guarracino, Pooled testing with replication as a mass testing strategy for the COVID-19 pandemics, *Scientific Reports*, **11**:3459, 2021.

[45] I. Yelin, N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, R. Kishony, Evaluation of COVID-19 RT-qPCR test in multi-sample pools, `http://medrxiv.org/lookup/doi/10.1101/2020.03.26.20039438`, 2020.

[46] N. Zaman, N. Pippenger, Asymptotic analysis of optimal nested group-testing procedures, *Probability in the Engineering and Informational Sciences*, **30**:547–552, 2016.