

# Pool Testing with Dilution Effects and Heterogeneous Priors

Gustavo Quinderé Saraiva\*

July 29, 2023

## Abstract

The Dorfman *pooled testing* scheme is a process in which individual specimens (e.g., blood, urine, swabs, etc.) are pooled and tested together; if the merged sample tests positive for infection, then each specimen from the pool is tested individually. Through this procedure, laboratories can reduce the expected number of tests required to screen the population, as individual tests are only carried out when the pooled test detects infection. Several different partitions of the population can be used to form the pools. In this study we analyze the performance of *ordered partitions*, those in which subjects with similar probability of infection are pooled together. We derive sufficient conditions under which ordered partitions outperform other types of partitions in terms of minimizing the expected number of tests, the expected number of false negatives, and the expected number of false positive classifications. These sufficient conditions can be easily verified in practical applications, once the dilution effect has been estimated. We also propose a measure of equity and present conditions under which this measure is maximized by ordered partitions.

## Highlights

- This study derives conditions under which it may be desirable to group together patients with similar probability of infection when implementing the Dorfman pooled testing procedure. It is shown that, depending on the dilution effect, this way of pooling subjects minimizes the expected number of tests required to screen the population, as well as the expected number of false positives and false negatives.
- This study also proposes a measure of equity and derives conditions under which this pooling method maximizes equity.
- Two case studies are conducted showing how these results can be applied to real data.

## 1 Introduction

For many infectious diseases it is common practice to screen the population through a *pool testing* scheme (also known as *group testing*) a process in which specimens (e.g., blood, urine, swabs) from different subjects are pooled and tested together. One of the simplest versions of pooled testing is the Dorfman screening procedure (due to Dorfman [1943]). In this procedure, pooled samples are tested together, and whenever a pooled test detects infection, each specimen from that group is tested individually. Compared to testing subjects individually, this procedure may potentially reduce the overall expected number of tests required to screen a population, as subjects are only tested individually in the event the pooled test detects infection. Perhaps because of its simplicity, this type of pool testing scheme is the one most implemented in practical applications (e.g., McMahan, Tebbs and Bilder [2012]).

The origin of the pooled testing literature is usually attributed to the seminal work of Dorfman [1943], which suggested pooling blood samples from the US military to detect syphilis in soldiers during World War II. Since then, the field has evolved to produce several new applications, including the detection of other infectious diseases such as Chlamydia (e.g., McMahan, Tebbs and Bilder [2012]) and HIV (e.g., Nguyen et al. [2019]), the detection of defective parts in production lines (e.g., Sobel and Groll [1959]), the detection of data tampering using one-way hash functions (e.g., Goodrich, Atallah and Tamassia [2005]) and the allocation of transmission time slots to

---

\*Business School, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, Santiago, Chile. E-mail address: gsaraiva@uc.cl

users in multiaccess channels (e.g., Chlebus [2001]). Recently, the field has regained new interest as pool testing techniques can be used to reduce the costs of COVID-19 screening. As of now, many countries, including the US, Germany, China and Chile, have used pool testing to screen for COVID-19 (e.g., Basso et al. [2022], Basso et al. [2023], Grobe et al. [2020] and Fan [2020]).

For tractability reasons, most theoretical research on pooled testing works in an environment that does not allow dilution effects, i.e., they assume that increasing the number of infected within a pool does not increase the probability of detecting infection in the pool (see Kim et al. [2007] for a thorough literature review). But in reality, dilution effects have proven to be non-negligible in several practical applications, including in the detection of COVID-19 using RT-PCR tests (e.g., Bateman et al. [2020] and Yelin et al. [2020]).<sup>1</sup> Small but non-negligible dilution effects have been reported in other applications, such as when pooling samples to test for chlamydia and gonorrhea, for pool sizes less than or equal to 10 (Morre et al. [2000] and Kacena et al. [1998b]).

Most research dedicated to finding optimal pool sizes also works in an environment where each subject has the same probability of infection (e.g., Kim et al. [2007]) when in reality, the probability of infection can be highly dependent on subjects' sociodemographic and clinical characteristics. For example, the prevalence of chlamydia and other sexually transmitted diseases in the U.S. varies considerably with age and other demographics.<sup>2</sup> For blood screening, the probability of HIV infection from first-time donors in the U.S. is approximately 7 times higher than that from repeat donors (Zou, Stramer and Dodd [2012]). Because some of subjects' attributes can be collected by the tester, the tester could use some of this information to determine the optimal partition used to form the pools.

In this study we analyze the performance of *ordered partitions*, those in which subjects with similar probability of infection are pooled together. We show that, provided that the dilution effect is not too strong and a technical condition is met regarding the concavity of the function governing the dilution effect, then, for any arbitrary partition, there always exists an ordered partition with the same pool size configuration that yields a lower expected number of tests, another (potentially different) ordered partition with same pool size configuration that yields a lower expected number of false negatives, and yet another ordered partition with same pool size configuration that yields a lower expected number of false positives. As a corollary, if these conditions are met and one is only considering partitions in which all pools have the same size, ordered partitions will always outperform any other partition with the same pool size configuration in all of the three attributes mentioned earlier: expected number of tests, expected number of false positives and expected number of false negatives. This result has practical applications to situations in which reconfiguring the testing machine for different pool sizes is time-consuming or impractical. For cases in which pool sizes are allowed to be heterogeneous, we show that we can use the algorithm proposed by Aprahamian, Bish and Bish [2019] to find the optimal ordered partition as well as a lower bound to the minimum cost. We also propose a simple method to find the optimal ordered partition in conjunction with the optimal cutoff point for the biomarker reading above which infection is detected, which can vary with the pool size.

We also characterize ordered partitions in terms of equity. Ideally, one would want to implement a testing scheme that is fair in the sense that it provides equitable expected payoffs to subjects. This is important because, depending on the matching criteria used to form the pools, subjects belonging to certain groups can end up with a disproportionately high probability of being misclassified (either with a false negative or a false positive classification). We show that conditional that all pools have the same size, ordered partitions do not always yield the most equitable allocation, even if they minimize both types of classification errors. But we find an instance in which ordered partitions are guaranteed to generate the most equitable allocation regardless of the concavity of the function governing the dilution effect and regardless of the distribution of priors: when all pools are comprised of only two subjects, and subjects either only care about false negative errors, or they only care about false positive errors. For bigger pools and situations in which subjects care about both types of classification errors, we can use the information on the dilution effect and the distribution of priors to build an upper bound to the maximum level of equity that can be obtained by any partition.

We apply our results to the detection of chlamydia and hepatitis B through pooled testing. To do so, we first estimate the dilution effect for assays used to detect those diseases, as well as the prevalences of those diseases for different demographic groups, using data from previous studies. We then use those estimates to verify whether the

---

<sup>1</sup>Bateman et al. [2020] estimate that the probability of detecting COVID-19 from an infected subject is 6% lower when his sample is diluted with the sample of 4 healthy subjects. The percentage reduction in the precision of the test is 8% when the infected sample is further diluted with 9 non-infected samples, and 18% when the infected sample is diluted with 49 non-infected specimens.

<sup>2</sup>See section 9 for details.

hypotheses from our theoretical results are satisfied, and conduct numerical exercises comparing the performance of other heuristics, such as random partitions (i.e., those in which subjects are assigned to pools randomly). We find that, in general, the conditions that guarantee that ordered partitions perform well in terms of minimizing the expected number of tests and the expected number of false positives are met. Though the sufficient conditions that guarantee that ordered partitions perform well in terms of minimizing the expected number of false negatives are not always met, our simulations indicate that ordered partitions still tend to outperform other heuristics considered in practical applications, such as random pooling. Our simulations also indicate that ordered partitions have a better performance than random partitions in terms of maximizing equity.

## 2 Related Literature

This paper is related to the work of Hwang [1975], McMahan, Tebbs and Bilder [2012] and Aprahamian, Bish and Bish [2019], which propose algorithms to determine the optimal partition used to perform Dorfman testing, so as to either minimize the expected number of tests (Hwang [1975] and McMahan, Tebbs and Bilder [2012]) or a convex combination of the expected number of tests and both types of classification errors (Aprahamian, Bish and Bish [2019]). But different from our setting, they follow the literature convention by assuming that pooled testing is not subject to dilution effects. Hwang [1975] and McMahan, Tebbs and Bilder [2012] have shown that, in the absence of dilution effects, there exists an ordered partition that minimizes the expected number of tests required to screen the population. Moreover, in the absence of dilution effects, the probability that any infected subject is incorrectly diagnosed as not infected does not depend on who the subject is matched with in the pool. So, in the absence of dilution effects, the matching criteria used to form the pools does not affect the expected number of false negatives. As to the other type of classification error, Aprahamian, Bish and Bish [2019] have shown that, in the absence of dilution effects, there exists an ordered partition that minimizes the expected number of false positives. So these results indicate that ordered partitions perform well in all of the three dimensions considered: expected number of tests, expected number of false positives and expected number of false negatives. We extend these results by showing that, when dilution effects are present but are sufficiently small, there exists an ordered partition that minimizes the expected number of tests and another ordered partition that minimizes the expected number of false positives, conditional on a given set of pool sizes.

This paper is also related to the work of Hwang [1976] and Wein and Zenio [1996] which vie to find optimal pool sizes for the Dorfman screening when pooled testing is subject to dilution effects. Their work, however, follow the literature convention by assuming that the probability of infection is homogeneous across the population, whereas in our environment we allow subjects to have heterogeneous probability of infection.

Most research in the literature that simultaneously allows for dilution effects and heterogeneous priors are the ones dedicated to estimating the probability of infection conditional on the results of pooled tests, such as the work of Wang, McMahan and Gallagher [2015], Warasi et al. [2017] and Mokalled et al. [2021]. To the best of my knowledge, the only theoretical research that attempt to formulate optimal pooling schemes under the presence of both dilution effects and heterogeneous priors are the work of El-Amine, Bish and Bish [2017], Aprahamian, Bish and Bish [2020] and Aprahamian, Bish and Bish [2018]. El-Amine, Bish and Bish [2017] and Aprahamian, Bish and Bish [2020] propose a testing scheme in which subjects with similar probability of infection are matched together to form the pools. However, for tractability reasons they assume that the dilution effect only depends on the pool size, not on the number of subjects infected within the pool.

More in line with our approach, Aprahamian, Bish and Bish [2018] work in an environment in which the dilution effect is affected by the proportion of infected subjects within the pool. They present conditions regarding the concavity of the dilution function which guarantee that ordered partitions minimize the expected number of tests and the expected number of false negative classifications. We show that these conditions can be relaxed by using the fact that only a discrete number of subjects is ever tested, so we only need to present conditions regarding concavity of the dilution function at the discrete level. One practical application of this stronger result is that it implies that, for any realistic dilution function (more precisely, if increasing the proportion of infected subjects within the pool increases the probability of detecting infection), ordered partitions will always minimize the expected number of false negatives conditional that all subjects are pooled into groups of size 2 (an analogous result follows for false positive classifications). This strengthened result also allows us to show analytically that, if the average viral load of infected subjects is sufficiently high compared to non-infected subjects, the dilution effect will not be “too strong”, which implies that ordered partitions are optimal in terms of minimizing the expected

number of tests (and expected number of false positives). Aprahamian, Bish and Bish [2018] also require pool sizes to be equal, whereas we derive results that apply to situations in which pool sizes are heterogeneous. Finally, Aprahamian, Bish and Bish [2018] only derive results regarding the expected number of tests and expected number of false negatives, while we also derive conditions under which ordered partitions are optimal in terms of minimizing the expected number of false positives.

Our chlamydia case study follows Aprahamian, Bish and Bish [2019] very closely, with the exception that we allow the existence of dilution effects. Aprahamian, Bish and Bish [2018] perform a similar case study using the same dataset and allowing the existence of dilution effects. But they restrict all pool sizes to be homogeneous, whereas we employ the methodology proposed by Aprahamian, Bish and Bish [2019] to find the optimal ordered partition when pool sizes are allowed to be heterogeneous. Moreover our results are not directly comparable with Aprahamian, Bish and Bish [2018], as they analyze the optimal ordered partition under an adaptive array testing, whereas we study the optimal ordered partition under Dorfman testing.

Our Hepatitis B case study appears to be the first attempt to compute the optimal ordered partition to screen subjects for HBV infection for the purpose of blood transfusion. Though our dataset has been extensively analyzed in other studies, such as Wang, McMahan and Gallagher [2015], Warasi et al. [2017] and Mokalled et al. [2021], these have been primarily focused in estimating the probability of someone being infected as a function of the result of a pooled testing scheme, not in finding the optimal ordered partition to implement Dorfman testing. We also propose a simple methodology to compute the optimal ordered partition in conjunction with a cutoff point for the biomarker, above which infection is detected, where we allow the cutoff point to be pool size dependent. The optimal cutoff point is also a function of the distribution of OD readings among infected and non infected specimens, as well as on the cost of implementing a test, the cost of getting a false negative result and the cost of getting a false positive result. Of these costs, the one that is arguably the hardest to estimate is the cost of a false negative, which we assume to be equal to the cost of infecting someone with the Hepatitis B virus (HBV). We used a simple Markov model similar to the one employed in Jackson et al. [2003] and Birkmeyer et al. [1993] to estimate this cost. These papers study the potential benefits associated with allogenic blood transfusion, taking into account the possibility that the screening for blood borne diseases, such as Hepatitis B, Hepatitis C and HIV, may not be 100% sensitive. Though they incorporate dilution effects in their calibration, Jackson et al. [2003] and Birkmeyer et al. [1993] take the pooled testing scheme as given, and are not focused in finding optimal partitions to test subjects, nor in finding optimal cutoff points for the biomarker.

To the best of my knowledge, the only theoretical research on pooled testing that addresses equity concerns when implementing ordered partitions is the work from Aprahamian, Bish and Bish [2019]. Dilution effects are not allowed in their environment, however. This implies that if subjects only care about false negative classifications, then conditional on a subject being infected, his expected payoff is not affected by the set of subjects with whom he was pooled, only by whether he was pooled or not. But in our case, when dilution effects are present, the probability of infection of other subjects within a pool does affect a subject's probability of receiving a false negative result. In such a case, we show that ordered partitions do not necessarily yield the most equitable allocation. But using information on the dilution effect and the configuration of probabilities of infection, we show how to derive an upper bound to the optimal level of equity that can be achieved by any partition. To derive this upper bound, we require the usage of a welfare function that is slightly different than the one used in Aprahamian, Bish and Bish [2019]: they employed the  $\alpha$ -fairness specification, whereas we use an *utilitarian max-min* welfare function.

### 3 Environment

Suppose that there is a population  $S = \{1, 2, \dots, n\}$  of subjects to be tested. Each subject can either be infected or not infected (we use the terms *not infected* and *healthy* interchangeably). Each subject  $i \in S$  is infected with probability  $q_i \in [0, 1]$ . Without loss of generality, we assume throughout the paper that

$$q_1 \leq q_2 \leq \dots \leq q_n.$$

If a subject is individually tested and he is not infected, the test will classify him as healthy with probability  $S_p \in [0, 1]$ . An infected subject who is individually tested is classified as infected with probability  $S_e \in [0, 1]$ . Using the terminology from clinical trials,  $S_p$  corresponds to the *specificity* of the test, while  $S_e$  corresponds to its *sensitivity*. We assume that  $S_e > 1 - S_p$ , so that whenever a test detects infection, the subject is more likely to be infected compared to the case in which no infection is detected.

Table 1: Notation and Abbreviations.

Notation	Description
$k$	Pool size.
$I$	Number of infected within a pool.
$h(I, k)$	Dilution function: it returns the probability that infection is detected in a pool with $k$ subjects with exactly $I \leq k$ of them infected.
$S_e$	Sensitivity of individual testing (i.e., $S_e = h(1, 1)$ ).
$S_p$	Specificity of individual testing (i.e., $S_p = 1 - h(0, 1)$ ).
$S$	Population of all subjects to be tested.
$n$	Number of subjects in the population (i.e., $n =  S $ ).
$q_i$	Probability that subject $i \in S$ is infected.
$q$	Vector of probabilities of infection: $(q_1, q_2, \dots, q_n)$ .
$\Omega$	A partition of $S$ .
$G_g$	A group of subjects, i.e., a subset of $S$ .
$P_{G_g}(I)$	Probability that exactly $I$ of the $ G_g $ subjects in $G_g$ are infected.
$\mathbb{E}[T(\Omega)]$	Expected number of tests obtained after implementing partition $\Omega$ .
$\mathbb{E}[FN(\Omega)]$	Expected number of false negatives obtained after implementing partition $\Omega$ .
$\mathbb{E}[FP(\Omega)]$	Expected number of false positives obtained after implementing partition $\Omega$ .
$\mathbb{E}[C(\Omega)]$	Expected cost obtained after implementing partition $\Omega$ .
$u_i(\Omega)$	Expected utility from subject $i \in S$ when partition $\Omega$ is implemented.
$\pi_\alpha(u_1, u_2, \dots, u_n)$	The utilitarian max-min welfare function: $\alpha \min_{i \in S}(u_i) + (1 - \alpha) \sum_{i \in S} u_i$ .
Abbreviation	Description
OR	Optimal ordered partition (i.e., the ordered partition $\Omega$ that minimizes $\mathbb{E}[C(\Omega)]$ ).
$\widehat{OR}$	Optimal ordered partition ignoring the existence of dilution effects.
R1	Optimal random partition with homogeneous pool sizes.
R2	Optimal random partition with homogeneous pool sizes and cutoffs above which subjects are individually tested.
IT	Individual testing.
LB	Lower bound to the minimum expected cost $\mathbb{E}[C(\Omega)]$ .
HBV	Hepatitis B Virus.
OD	Optical Density: a biomarker used to detect HBV infection.

In a Dorfman procedure, the subjects to be tested,  $S$ , are pooled into disjoint groups, and the samples from subjects belonging to the same group are amalgamated and tested together. If the test detects infection for the pooled sample, then each subject within the pool is tested individually. If a group is comprised of only one subject, then this subject is only tested individually, without a followup test.

Several criteria can be used to form the pools. One way is to group subjects randomly, without taking into account their prior probability of infection, into pools of equal size. But such pooling method does not take advantage of the information on subjects' prior probability of infection, nor does it take into account the fact that, when a subject's probability of infection is sufficiently high, it may be preferable to test that subject individually.

So next we consider a class of pooling schemes in which subjects are ordered from lowest to highest probability of infection, and those with similar probability of infection are pooled together. Formally, a partition  $\Omega = \{G_1, G_2, \dots, G_m\}$  of  $S$  is said to be an *ordered partition* if, for any  $g, w \in \{1, 2, \dots, m\}$  with  $g \neq w$  we have that either  $q_i \geq q_j$  for all  $i \in G_g$  and all  $j \in G_w$  or  $q_i \leq q_j$  for all  $i \in G_g$  and all  $j \in G_w$ . In an *ordered pooling* scheme subjects are grouped according to an ordered partition.

We will compare this method of grouping subjects with alternative ones in terms of the expected number of tests they require to diagnose the entire population  $S$ , how many false negatives and false positives they generate (i.e., how many subjects are misclassified) and how equitable they are. But in order to make those comparisons, we must first make assumptions regarding how the probability of detecting infection in a pooled test is affected by the number of infected subjects within the pool.

Let  $h(I, k)$  be the probability of detecting infection in a pooled sample collected from  $k \in \mathbb{N}$  subjects, conditional that exactly  $I \in \{0, 1, 2, \dots, k\}$  of those subjects are infected. From our definition of sensitivity and specificity,

we must have  $h(1, 1) = S_e$  and  $h(0, 1) = 1 - S_p$ . We will occasionally refer to  $h$ , as the *dilution function*. For each  $k \in \mathbb{N}$ , we assume that  $h(\cdot, k)$  is (weakly) increasing, i.e., the more infected subjects there are in the pool, the more likely the pooled test will detect infection, which is arguably a very mild and reasonable assumption.

**Assumption 1**  $h(I, k)$  is increasing in  $I$  (i.e., the more infected subjects there are in the group, the more likely the pooled test will detect infection for that group).

Most of the literature assumes that pooled samples are not susceptible to dilution, so that the probability of detecting infection in a pooled sample that has at least one infected subject is the same as the probability of detecting infection from an individual test of an infected subject. This corresponds to the case in which

$$h(I, k) = \begin{cases} S_e, & \text{if } I > 0 \\ 1 - S_p, & \text{if } I = 0. \end{cases} \quad (1)$$

But in general,  $h$  could assume different formats. The faster  $h(I, k)$  converges to  $h(k, k)$  as  $I$  approaches  $k$ , the lower is the dilution effect. Figure 1 depicts different functions  $h(\cdot, k)$  that satisfy assumption 1. The lower and more convex functions correspond to cases in which the dilution effect is stronger.

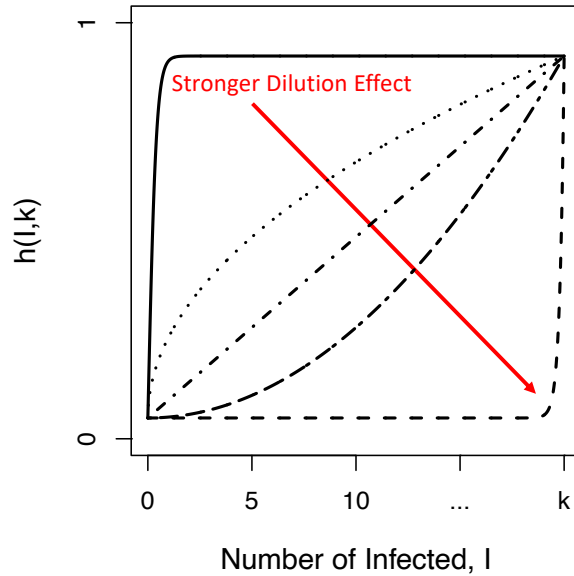


Figure 1: Different dilution functions  $h(\cdot, k)$ , for a given  $k \in \mathbb{N}$ .

Concavity of  $h(I, k)$  with respect to  $I$  (i.e., concavity of  $h(\cdot, k)$ ) is a sufficient condition for many of the results presented in this paper, as concavity of  $h(I, k)$  with respect to  $I$  implies that the dilution effect is not too strong. But because the number of subjects within each group is a discrete variable, we only require concavity at the discrete level. So for most part of the paper we will use the following definition of concavity:

**Definition 1** We say  $h(\cdot, k)$  is discrete-concave if, for any  $I \in \{0, 1, 2, \dots, k\}$ , we have that  $h(I + 1, k) - h(I - 1, k) \leq 2h(I, k)$ .

Notice that, if  $h(\cdot, k)$  is concave, it is discrete-concave, but the reverse is not necessarily true. This is important because, sometimes, depending on how the dilution function is estimated, one can find  $h(\cdot, k)$  to be convex for values of  $I$  between  $(0, 1)$ , even though the function is discrete-concave (see example 3 from section 8).

Table 1 summarizes the notation that will be used throughout the paper.

## 4 Expected Number of Tests

Because testing is costly, ideally one would want to implement a partition that minimizes the expected number of tests required to screen the population. For a given partition  $\Omega \equiv \{G_1, G_2, \dots, G_{n/k}\}$  of the population  $S$ , we denote  $T(\Omega)$  as the number of tests performed on the population after implementing the Dorfman procedure using this partition to determine the pools, and  $\mathbb{E}[T(\Omega)]$  as its corresponding expectation.

It is straightforward to show that, for any partition  $\Omega \equiv \{G_1, G_2, \dots, G_m\}$  of  $S$ , the expected number of tests required to screen the population using the Dorfman procedure is given by

$$\mathbb{E}[T(\Omega)] \equiv \sum_{G_g \in \Omega} T_{G_g}, \quad (2)$$

where

$$T_{G_g} \equiv \begin{cases} 1, & \text{if } |G_g| = 1 \\ 1 + |G_g| \sum_{I=0}^{|G_g|} h(I, k) P_{G_g}(I), & \text{if } |G_g| > 1 \end{cases} \quad (3)$$

and

$$P_{G_g}(I) \equiv \sum_{\substack{G \subseteq G_g \\ s.t. |G|=I}} \prod_{i \in G} q_i \prod_{j \in G_g \setminus G} (1 - q_j), \quad (4)$$

i.e.,  $T_{G_g}$  is the expected number of test associated with group  $G_g$ , and  $P_{G_g}(I)$  is the probability that group  $G_g$  has exactly  $I$  infected subjects.

When  $h(\cdot, k)$  is discrete-concave for all  $k \in \mathbb{N}$ , we show that, for any partition  $\Omega$ , we can find an ordered partition  $\Omega^*$  that preserves the pool sizes of  $\Omega$ , and generates a weakly lower expected number of tests. Moreover, this partition is such that, whenever it individually tests a subject  $i$  with probability of infection  $q_i$ , it also individually tests all subjects with higher probability of infection. This result is very similar to the one presented in Aprahamian, Bish and Bish [2018], with the difference that we only require discrete-concavity as opposed to concavity, and our result also applies to cases in which pool sizes differ.

**Theorem 1** *Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be an arbitrary partition of  $S$ . Suppose that  $h(\cdot, |G_g|)$  is discrete-concave for every  $G_g \in \Omega$ . Then, there exists an ordered partition  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  such that*

- a)  $|G_g^*| = |G_g|$  for all  $g \in \{1, 2, \dots, m\}$ .
- b) Whenever a subject  $i \in S$  with probability of infection  $q_i$  is individually tested under  $\Omega^*$ , subjects with a probability of infection higher than  $q_i$  are also individually tested under  $\Omega^*$ .
- c)  $\mathbb{E}[T(\Omega^*)] \leq \mathbb{E}[T(\Omega)]$ .

Intuitively, the parts a and c from theorem 1 can be explained as follows. Suppose that we were to pool subjects randomly. In this case, each group would have a high probability of having at least one infected subject. A discrete-concave dilution function  $h(\cdot, k)$  implies that the dilution effect is not too strong. If the dilution effect is sufficiently small, pooled tests would detect infection for most groups, resulting in many individual follow-up tests being conducted. If, on the other hand, ordered pooling was implemented, only the groups with high probability of infection would be likely to have at least one infected subject, and therefore to test positive for the disease, thus resulting in a lower number of follow-up tests.

As to part b from theorem 1, we have that, adding more infected subjects into a pool increases the probability that the pooled test detects infection, thus triggering followup tests for all the subjects within that pool. So to minimize the expected number of tests, those with highest probability of infection should be the ones, if any, allocated for individual testing.

Notice that theorem 1 requires  $h(\cdot, k)$  to be discrete-concave for all different pool sizes from the original partition. This hypothesis will be satisfied if, for instance, the dilution function is assumed to depend only on the proportion of infected subjects within the pool,  $I/k$ , and such dilution function is concave. As an example, consider the following class of dilution functions previously used in the literature (e.g., Burns and Mauro [1987] and Aprahamian, Bish and Bish [2018])

$$h(I, k) = (1 - S_p) + (S_p + S_e - 1) \left( \frac{I}{k} \right)^\delta, \quad (5)$$

where  $\delta \geq 0$ . As long as  $\delta \leq 1$ , the function  $h(\cdot, k)$  will be concave. In practice, the concavity of such function can be easily calibrated using empirical data on pooled tests (see section 9).

Theorem 1 guarantees that, for any partition  $\Omega$  we can always find an ordered partition  $\Omega^*$  that generates a weakly lower expected number of tests than  $\Omega$  and preserves the pool sizes of  $\Omega$ . So according to this theorem, if  $\Omega = \{\{3, 4\}, \{1, 2, 5\}\}$ , the ordered partition that generates a lower expected number of tests could be  $\{\{1, 2\}, \{3, 4, 5\}\}$  or  $\{\{1, 2, 3\}, \{4, 5\}\}$  (or both). Later, in section 6 we discuss how to find the optimal ordered partition by adopting the algorithm proposed by Aprahamian, Bish and Bish [2019].

**Corollary 1** *Suppose that  $h(\cdot, k)$  is discrete-concave. Let  $\Omega$  be a partition of  $S = \{1, 2, \dots, n\}$  with  $|G_g| = k \forall G_g \in \Omega$ . Then, the following ordered partition of  $S$*

$$\Omega^* = \{\{1, 2, \dots, k\}, \{k+1, k+2, \dots, 2k\}, \dots, \{n-k+1, n-k+2, \dots, n\}\}$$

is such that  $\mathbb{E}[T(\Omega^*)] \leq \mathbb{E}[T(\Omega)]$ .

## 5 Expected Number of False Negatives and False Positives

A false negative occurs when an infected subject is incorrectly classified as healthy, and a false positive occurs when a healthy subject is incorrectly classified as infected. For a given partition  $\Omega$  of  $S$ , we denote  $FN(\Omega)$  and  $FP(\Omega)$  as the total number of false negatives and false positives, respectively, obtained after implementing the Dorfman procedure using the partition  $\Omega$ .

It can be show that,<sup>3</sup> for any partition  $\Omega = \{G_1, G_2, \dots, G_m\}$  of  $S$ , the expected number of false negatives and the expected number of false positives obtained after implementing the Dorfman procedure using this partition is given by

$$\mathbb{E}[FN(\Omega)] = \sum_{G_g \in \Omega} FN_{G_g} \quad (6)$$

and

$$\mathbb{E}[FP(\Omega)] = \sum_{G_g \in \Omega} FP_{G_g}, \quad (7)$$

respectively, where, for each  $G_g \in \Omega$

$$FN_{G_g} \equiv \begin{cases} (1 - S_e)q_i, & \text{if } G_g = \{i\} \\ \sum_{I=0}^{|G_g|} P_{G_g}(I)I[1 - h(I, |G_g|)S_e], & \text{if } |G_g| > 1 \end{cases},$$

$$FP_{G_g} \equiv \begin{cases} (1 - S_p)(1 - q_i), & \text{if } G_g = \{i\} \\ \sum_{I=0}^{|G_g|} P_{G_g}(I)h(I, |G_g|)(|G_g| - I)(1 - S_p) & \text{if } |G_g| > 1 \end{cases}$$

and  $P_{G_g}(I)$  is the probability that group  $G_g \in \Omega$  has exactly  $I$  infected subjects (see equation (4)).

We now derive sufficient conditions under which ordered partitions minimize each type of classification error.<sup>4</sup>

**Hypothesis 1** *Suppose that the dilution function  $h(\cdot, k)$  is such that, for all  $I \in \{1, 2, \dots, k-1\}$ ,*

$$\frac{I+1}{2I}h(I+1, k) + \frac{I-1}{2I}h(I-1, k) \geq h(I, k).$$

**Theorem 2** *Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be an arbitrary partition of  $S$ . Suppose that hypothesis 1 holds for every pool size  $k \in \{k' \in \mathbb{N}; k' = |G_g|, \text{ for some } G_g \in \Omega\}$ . Then, there exists an ordered partition  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  such that*

<sup>3</sup>Details provided in the online Appendix.

<sup>4</sup>Aprahamian, Bish and Bish [2018] had previously shown that, when all the pools have the same size  $k \geq 2$ , and the following condition holds

$$I \frac{\partial^2 h(I, k)}{\partial I^2} + 2 \frac{\partial h(I, k)}{\partial I} \geq 0 \quad \forall I \in [0, k],$$

then grouping subjects according to an ordered partition minimizes the expected number of false negatives. In the online Appendix we show that this condition implies that hypothesis 1 holds, but the converse is not necessarily true.

a)  $|G_g^*| = |G_g|$  for all  $g \in \{1, 2, \dots, m\}$ ,

b)  $\mathbb{E}[FN(\Omega^*)] \leq \mathbb{E}[FN(\Omega)]$ .

Hypothesis 1 can be interpreted as requiring that the dilution function  $h(\cdot, k)$  is not “too concave”. In fact, hypothesis 1 is satisfied whenever  $h(\cdot, k)$  is convex, as convexity of  $h(\cdot, k)$  implies that

$$\frac{I+1}{2I} h(I+1, k) + \frac{I-1}{2I} h(I-1, k) \geq h(I+1/I, k) \geq h(I, k).$$

But convexity is not a necessary condition for hypothesis 1 to hold. Indeed, for any  $\delta \in (0, 1)$ , the dilution function introduced in equation (5) is not convex, and yet it satisfies hypothesis 1.

Though Aprahamian, Bish and Bish [2019] have shown that, in the absence of dilution effects, allocating those with a higher probability of infection to be individually tested minimizes false negative classifications, this is not necessarily the case in the presence of dilution effects, as illustrated by example 1.

**Example 1** Suppose that the dilution function is given by expression 5, with  $\delta = 1/2$ ,  $S_e = .97$  and  $S_p = .95$ . Then, if  $S = \{1, 2, 3\}$  and  $q_1 = .1$ ,  $q_2 = .9$  and  $q_3 = .99$  we have that

$$\mathbb{E}[FN(\{\{1\}, \{2, 3\}\})] = 0.143 < 0.303 = \mathbb{E}[FN(\{\{1, 2\}, \{3\}\})].$$

Intuitively, when there are dilution effects it is not necessarily optimal to allocate a subject with high probability of infection for individual testing, as allocating this subject into a pool will reduce the probability of false negative from those within the pool.

We now derive sufficient conditions under which ordered pooling minimizes false positive classifications. Those conditions will go in the opposite direction of the conditions we required for ordered pooling to minimize false negative classifications: instead of requiring the dilution function not to be “too concave”, we now require it not to be “too convex”.

**Hypothesis 2** Suppose that the dilution function  $h(\cdot, k)$  is such that, for all  $I \in \{1, 2, \dots, k-1\}$ ,

$$\frac{k-I-1}{2(k-I)} h(I+1, k) + \frac{k-I+1}{2(k-I)} h(I-1, k) \leq h(I, k).$$

**Theorem 3** Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be an arbitrary partition of  $S$ . Suppose that hypothesis 2 holds for every pool size  $k \in \{k' \in \mathbb{N}; k' = |G_g|, \text{ for some } G_g \in \Omega\}$ . Then, there exists an ordered partition  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  such that

a)  $|G_g^*| = |G_g|$  for all  $g \in \{1, 2, \dots, m\}$ ,

b) Whenever a subject  $i \in S$  with probability of infection  $q_i$  is individually tested under  $\Omega^*$ , subjects with a probability of infection higher than  $q_i$  are also individually tested under  $\Omega^*$ .

c)  $\mathbb{E}[FP(\Omega^*)] \leq \mathbb{E}[FP(\Omega)]$ .

Notice that if  $h(\cdot, k)$  is discrete-concave, then hypothesis 2 holds, as discrete-concavity of  $h(\cdot, k)$  and the fact that  $h(\cdot, k)$  is increasing implies that

$$\frac{k-I-1}{2(k-I)} h(I+1, k) + \frac{k-I+1}{2(k-I)} h(I-1, k) \leq \frac{h(I+1, k)}{2} + \frac{h(I-1, k)}{2} \leq h(I, k).$$

**Corollary 2** Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be an arbitrary partition of  $S$  such that  $|G_g| \geq 2$  for all  $G_g \in \Omega$ . Suppose that  $h(\cdot, k)$  is discrete-concave for every  $k \in \{k' \in \mathbb{N}; k' = |G_g|, \text{ for some } G_g \in \Omega\}$ . Then, there exists an ordered partition  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  such that

a)  $|G_g^*| = |G_g|$  for all  $g \in \{1, 2, \dots, m\}$ ,

b)  $\mathbb{E}[FP(\Omega^*)] \leq \mathbb{E}[FP(\Omega)]$ .

Intuitively, when the dilution effect is relatively small, matching subjects randomly to form the pools is likely to generate many false positives, as each group will have a high chance of having at least one infected subject who will trigger a follow-up test with high probability. If the individual tests have imperfect specificity, those follow-up tests will increase the probability that non-infected subjects are incorrectly classified as infected. Performing ordered pooling, on the other hand, results in less follow-up tests, as this pooling criterion reduces the frequency with which a pooled sample is “contaminated” with an infected specimen that triggers a follow-up test.

Notice that, when we are only considering pools of size  $k = 2$ , we only need to check whether hypotheses 1 or 2 hold for  $I = 1$ . Because  $h(\cdot, k)$  is increasing, both of these hypotheses are clearly satisfied at  $I = 1$ . So a corollary to theorems 2 and 3 is that, if all the pools from a partition have size  $k = 2$ , then an ordered partition in which all pools have size  $k = 2$  will produce a lower expected number of both types of classification errors.

**Corollary 3** *Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be a partition of  $S$  such that  $|G_g| = 2$  for all  $G_g \in \Omega$ . Suppose that  $h(\cdot, 2)$  is increasing. Then, there exists an ordered partition  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  such that*

a)  $|G_g^*| = 2$  for all  $G_g \in \Omega$ ,

b)  $\mathbb{E}[FN(\Omega^*)] \leq \mathbb{E}[FN(\Omega)]$ ,

c)  $\mathbb{E}[FP(\Omega^*)] \leq \mathbb{E}[FP(\Omega)]$ .

Finally, notice that a corollary from theorems 1, 2 and 3 is that, if we are only considering partitions in which all pools are required to have the same size  $k$  that is a dividend of the population size  $n$ , finding the optimal partition when the dilution function is discrete-concave, and hypothesis 2 holds is trivial: it consists of implementing the following ordered partition of  $S$ .

$$\Omega^* = \{\{1, 2, \dots, k\}, \{k+1, k+2, \dots, 2k\}, \dots, \{n-k+1, n-k+2, \dots, n\}\}.$$

**Corollary 4** *Suppose that  $h(\cdot, k)$  is discrete-concave and satisfies hypothesis 2. Let  $\Omega$  be any partition of  $S = \{1, 2, \dots, n\}$  with  $|G_g| = k \forall G_g \in \Omega$ . Then, the following ordered partition of  $S$*

$$\Omega^* = \{\{1, 2, \dots, k\}, \{k+1, k+2, \dots, 2k\}, \dots, \{n-k+1, n-k+2, \dots, n\}\}$$

*is such that*

a)  $\mathbb{E}[T(\Omega^*)] \leq \mathbb{E}[T(\Omega)]$ ,

b)  $\mathbb{E}[FN(\Omega^*)] \leq \mathbb{E}[FN(\Omega)]$ ,

c)  $\mathbb{E}[FP(\Omega^*)] \leq \mathbb{E}[FP(\Omega)]$ .

Corollary 4 has practical implications to situations in which the tester is interested in using the same pool size for all tests, say, because reconfiguring the pool sizes is too costly. But if the costs of changing the pool sizes are not substantial, it may be beneficial to implement a partition with heterogeneous pool sizes. In the next section, we discuss how to find the optimal ordered partition when pool sizes are allowed to be heterogeneous, and how to derive a lower bound to the optimal cost.

## 6 Finding the optimal ordered partition

In general, a tester is interested in implementing a partition that minimizes a convex combination of the expected number of tests, the expected number of false negatives and the expected number of false positives, i.e., they are interested in minimizing the following expression:

$$\mathbb{E}[C(\Omega)] \equiv \lambda_1 \mathbb{E}[FN(\Omega)] + \lambda_2 \mathbb{E}[FP(\Omega)] + \lambda_3 \mathbb{E}[T(\Omega)], \quad (8)$$

where  $\lambda_1, \lambda_2, \lambda_3 \geq 0$ , i.e., the  $\lambda$ 's correspond to the weights assigned to each attribute.

If, for instance  $\lambda_1 > 0$  and  $\lambda_2 = \lambda_3 = 0$ , the tester's sole purpose is minimizing the expected number of false negatives, which can be achieved by individually testing all subjects. Indeed, it can be shown that, as long as the dilution function  $h(\cdot, k)$  is non-decreasing for every  $k$ , individual testing always minimizes the expected number of false negatives, for any realization of  $(q_i)_{i \in S}$ .

**Proposition 1** *If the dilution function satisfies assumption 1 (i.e., if  $h(\cdot, k)$  is increasing for every  $k \in \mathbb{N}$ ), then testing all subjects individually minimizes the expected number of false negatives.*

Individual testing has, however, the obvious disadvantage of requiring too many tests to screen the population, which is costly. So individual testing is usually not optimal when  $\lambda_1 > 0$ , the prevalence of infection is small and the dilution function is not too strong. In those cases, finding the optimal partition can be challenging, as the number of different partitions from even a small batch of 20 subjects can be quite large.<sup>5</sup>

But from the previous section we have seen that, when the dilution function is discrete-concave and satisfies hypothesis 1, then, for any arbitrary partition, there always exists an ordered partition with the same pool size configuration that yields a lower expected number of tests, another (potentially different) ordered partition with same pool size configuration that yields a lower expected number of false negatives, and yet another ordered partition with same pool size configuration that yields a lower expected number of false positives. While these results do not guarantee that an ordered partition will minimize all of these three attributes *simultaneously*, they are indicative that they represent good candidates for an optimal allocation, as deviations from ordered pooling imply that we are able to improve at least one of the attributes.

Though the set of ordered partitions to be considered is still very large,<sup>6</sup> Aprahamian, Bish and Bish [2019] have noticed that one can still find the optimal ordered partition in polynomial time by implementing a *shortest path algorithm*, such as *Dijkstra's* minimum weight path algorithm.

Indeed, if, for example  $S = \{1, 2, 3, 4, 5\}$ , we can form a bijection that maps all the possible ordered partitions of  $S$  into the set of possible paths leading node 1 into node 6 from the directed graph displayed in figure 2.

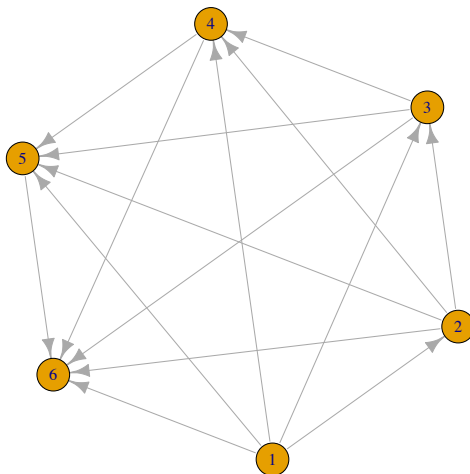


Figure 2: A graph that can be used to represent all the possible ordered partitions of  $S = \{1, 2, 3, 4, 5\}$ .

Table 2 displays examples of ordered partitions of  $S = \{1, 2, 3, 4, 5\}$ , and how they would be mapped into a path from graph 2.

Because we can find a cost associated with each element of the partition, we can map each of these costs into the weights given to each edge of the graph. For example, if we find that the expected cost of pool testing the subgroup  $\{1, 2\}$  is given by \$ 20,00, then this value will represent the weight assigned to the edge that connects node 1 to 3. Once we compute the weights from each edge of the graph, we can then use *Dijkstra's* algorithm, which is  $O(|S|^2)$ , to find the optimum path in polynomial time (e.g., Nemhauser and Wolsey [1998]).

Because we cannot guarantee that the optimal partition will indeed be an ordered partition, we also compute a lower bound to the objective function (8). This can be achieved by computing

$$\lambda_1 \min_{\Omega} (\mathbb{E}[FN(\Omega)]) + \lambda_2 \min_{\Omega} (\mathbb{E}[FP(\Omega)]) + \lambda_3 \min_{\Omega} (\mathbb{E}[T(\Omega)]).$$

<sup>5</sup>The number of different partitions from a set  $S$  of 20 individuals is approximately  $5.17e + 13$ .

<sup>6</sup>Aprahamian, Bish and Bish [2019] shows that the number of different ordered partitions of  $S$  is  $2^{|S|-1}$ .

Table 2: Examples of ordered partitions of  $S = \{1, 2, 3, 4, 5\}$ , and how they would be mapped into paths from the acyclic directed graph depicted in figure 2.

Ordered Partition	Path
$\{\{1\}, \{2, 3, 4, 5\}\}$	$1 \rightarrow 2 \rightarrow 6$
$\{\{1, 2\}, \{3, 4, 5\}\}$	$1 \rightarrow 3 \rightarrow 6$
$\{\{1\}, \{2, 3\}, \{4, 5\}\}$	$1 \rightarrow 2 \rightarrow 4 \rightarrow 6$
$\{\{1\}, \{2, 3, 4\}, \{5\}\}$	$1 \rightarrow 2 \rightarrow 5 \rightarrow 6$
$\{\{1, 2, 3, 4, 5\}\}$	$1 \rightarrow 6$

Indeed, from proposition 1, we have that  $\min_{\Omega}(\mathbb{E}[FN(\Omega)])$  can be trivially obtained by implementing individual testing. To compute  $\min_{\Omega}(\mathbb{E}[FP(\Omega)])$  and  $\min_{\Omega}(\mathbb{E}[T(\Omega)])$ , we only need to implement the *Dijkstra's* algorithm to minimize each of these attributes separately, as theorems 1 and 3 guarantee that we can focus exclusively on ordered partitions to find those minimums when the dilution function is discrete-concave. Also notice that discrete-concavity of  $h(\cdot, k)$  is a very mild assumption: if this hypothesis did not hold, the dilution effect would be too strong, in which case pooled testing would generate too many false negatives, so it would not be a viable alternative to individual testing.

**Remark 1** Notice that

$$\min_{\Omega}(\mathbb{E}[C(\Omega)]) \geq \lambda_1 \min_{\Omega}(\mathbb{E}[FN(\Omega)]) + \lambda_2 \min_{\Omega}(\mathbb{E}[FP(\Omega)]) + (1 - \lambda_1 - \lambda_2) \min_{\Omega}(\mathbb{E}[T(\Omega)]).$$

Moreover, if  $h(\cdot, k)$  is non-decreasing for every  $k \in \mathbb{N}$ , then

$$\min_{\Omega}(\mathbb{E}[FN(\Omega)]) = (1 - S_e) \sum_{i \in S} q_i.$$

If, in addition,  $h(\cdot, k)$  is discrete-concave for every  $k \in \mathbb{N}$ , then  $\min_{\Omega}(\mathbb{E}[FP(\Omega)])$  and  $\min_{\Omega}(\mathbb{E}[T(\Omega)])$  can be obtained in  $\mathcal{O}(|S|^2)$  by implementing *Dijkstra's* algorithm.

In sections 9 and 10 we implement these techniques to find the *optimal ordered partition* in two case studies, as well as an upper bound to the optimal partition.

## 7 Equity

Ideally one would want to implement a testing scheme that is fair, in the sense that it provides equitable expected payoffs to subjects. This is important because, depending on the matching criteria used to form the pools, subjects belonging to certain demographic groups can end up with a disproportionately high probability of being misclassified (either with a false negative or a false positive classification). In this section we show that ordered pooling does not always implement the most equitable allocation when all pool sizes are required to be equal, even if it minimizes the overall expected number of false negatives and false positives. We show, however, that when all pools are of size  $k = 2$  and subjects either only care about false negative classifications or only care about false positive classifications, then ordered pooling is guaranteed to generate the most equitable allocation regardless of the dilution function and the distribution of priors. For pools of size  $k > 2$ , information on the dilution effect and distribution of priors can be used to derive sufficient conditions under which ordered pooling maximizes equity.

Assuming the Dorfman procedure is applied to a partition  $\Omega \equiv \{G_1, G_2, \dots, G_{n/k}\}$  of  $S$ , we denote  $FN_i(\Omega)$  as an indicator variable that equals 1 whenever subject  $i \in S$  is incorrectly classified as healthy, and zero otherwise. Similarly,  $FP_i(\Omega)$  is an indicator variable that equals 1 whenever subject  $i \in S$  is incorrectly classified as infected.

Letting  $\theta \in [0, 1]$  be the weight attributed to false negative classifications, and  $(1 - \theta)$  be the weight attributed to false positive classifications, we have that subject  $i$ 's expected utility from this partition is given by

$$u_i(\Omega) \equiv -\theta \mathbb{E}[FN_i(\Omega)] - (1 - \theta) \mathbb{E}[FP_i(\Omega)].$$

Throughout this section we use the *utilitarian max-min* welfare function to measure equity. As the name suggests, a utilitarian max-min welfare function is a convex combination of the utilitarian and max-min welfare

functions. The utilitarian welfare function, popularized by Harsanyi [1955], is simply the sum of utilities from the agents in the economy, while the max-min welfare function, proposed by Rawls [1971], equals to the lowest utility in the economy. So a utilitarian measure of welfare values *Pareto efficiency*, while a max-min measure values equity. A utilitarian max-min measure puts a positive weight on both of these attributes. More precisely, for a given parameter  $\alpha \in [0, 1]$ , the utilitarian max-min welfare function is given by

$$\pi_\alpha(u_1, u_2, \dots, u_n) \equiv \alpha \min_{i \in S} (u_i) + (1 - \alpha) \sum_{i \in S} u_i. \quad (9)$$

So the parameter  $\alpha$  can be interpreted as the weight put on equity, while  $1 - \alpha$  is the weight put on Pareto efficiency.

The utilitarian max-min welfare function is known to satisfy desirable axioms (e.g., see Schneider and Kim [2020] and Bossert and Kamaga [2020]). But for our purposes, the main reason why we invoke this welfare function is because it satisfies the following property: if an allocation maximizes the utilitarian welfare function *and* the max-min welfare function, then it maximizes the utilitarian max-min welfare function for any parameter  $\alpha$ . This property is not shared by some additive social welfare functions, such as the  $\alpha$ -*fairness* welfare function used in Aprahamian, Bish and Bish [2019],<sup>7</sup> even though the utilitarian and the max-min welfare functions correspond to the extreme opposite instances of the  $\alpha$ -fairness family, with the utilitarian approach putting all weight on Pareto efficiency, and the max-min putting all weight on equity (e.g., Lan et al. [2010] and Bertsimas, Farias and Trichakis [2012]).

The following example shows that ordered pooling does not necessarily maximize social welfare, even when this pooling scheme minimizes both the expected number of false positives and the expected number of false negatives.

**Example 2** Consider the dilution function  $h(I, k) = (1 - S_p) + (S_p + S_e - 1) \left(\frac{I}{k}\right)^{1/10}$ , with  $S_p = .95$  and  $S_e = .9$ . Figure 3 plots this dilution function when  $k = 3$  (i.e., when the pool has 3 subjects).

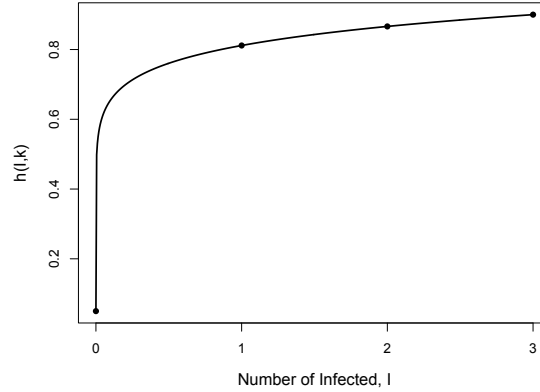


Figure 3: Dilution function  $h(I, k) = (1 - S_p) + (S_p + S_e - 1) \left(\frac{I}{k}\right)^{1/10}$ , with  $S_p = .95$ ,  $S_e = .9$  and  $k = 3$ .

Because  $h(\cdot, k)$  is concave, we have from theorem 1 and corollary 2 that ordered pooling minimizes the expected number tests and expected number of false positives. Moreover one can easily show that this dilution function satisfies hypothesis 1, which, from theorem 2, implies that ordered pooling also minimizes the expected number of false negatives.

Now suppose that the utilitarian max-min welfare function 9 is such that  $\alpha = 1$ , so that the partition that maximizes the social welfare function is the one that maximizes the payoff from the subject with lowest expected utility. Also suppose that  $\theta = 1$ , so that subjects only put weight on the probability that they get a false negative result.

<sup>7</sup>Indeed, consider  $u = (u_1, u_2, u_3, u_4) = (1, 1, 2, 4)$  and  $\bar{u} = (\bar{u}_1, \bar{u}_2, \bar{u}_3, \bar{u}_4) = (36/37, 36/37, 3, 3)$ . Then clearly  $\min u_i > \min \bar{u}_i$  and  $\sum u_i > \sum \bar{u}_i$ . However  $\sum \frac{u_i^{1-\alpha}}{1-\alpha} < \sum \frac{\bar{u}_i^{1-\alpha}}{1-\alpha}$  for  $\alpha = 2$ .

Now consider a population with the following vector of probability of infection:

$$q = (.01, .015, .95, .96, .98, .99).$$

Then, if we use the ordered partition, i.e.,  $\{\{1, 2, 3\}, \{4, 5, 6\}\}$ , the minimum payoff is the one from subject 3 and it is equal to 0.745. If, on the other hand, the partition  $\{\{4, 2, 3\}, \{1, 5, 6\}\}$  is used to group subjects, then the minimum payoff is the one from subject 6 and it is equal to 0.781. So ordered pooling does not maximize social welfare.

Similarly, if  $\theta = 0$ , so that subjects only care about false positive errors, and the vector of probability of infection is given by

$$q = (.001, .01, .02, .03, .98, .99),$$

then the ordered partition, i.e.,  $\{\{1, 2, 3\}, \{4, 5, 6\}\}$ , generates a minimum expected payoff of 0.956, while the partition  $\{\{1, 2, 5\}, \{3, 4, 6\}\}$  generates a higher minimum payoff of 0.957.

We find some instances, however, in which ordered pooling is guaranteed to maximize social welfare. Indeed, when all pools are required to be of size  $k = 2$  and all the weight on individual utilities is put either on false negative or false positive classifications (i.e.,  $\theta = 1$  or  $\theta = 0$ ), then ordered pooling maximizes social welfare.

**Proposition 2** Suppose that  $k = 2$  and  $\theta \in \{0, 1\}$ . Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be a partition of  $S$  such that  $|G_g| = 2$  for all  $G_g \in \Omega$ . Suppose that  $h(\cdot, 2)$  is increasing. Then, there exists an ordered partition  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  such that

- a)  $|G_g^*| = 2$  for all  $G_g \in \Omega$ ,
- b)  $\pi_\alpha(u_1(\Omega^*), u_2(\Omega^*), \dots, u_n(\Omega^*)) \geq \pi_\alpha(u_1(\Omega), u_2(\Omega), \dots, u_n(\Omega))$  for all  $\alpha \in [0, 1]$ .

When  $k > 2$  and  $\theta \in (0, 1)$ , we may use information on the distribution of types to compute an upper bound to the minimum utility.

**Proposition 3** Let  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  be an ordered partition of  $S = \{1, 2, \dots, n\}$  such that  $|G_g^*| = k \forall G_g^* \in \Omega^*$ , and let  $\Omega = \{G_1, G_2, \dots, G_{n/k}\}$  be any other partition such that  $|G_g| = k \forall G_g \in \Omega$ . Then

$$\min_{i \in G_j} u_i(\Omega) \leq -\theta \mathbb{E}[FN_n(\Omega^*)] - (1 - \theta) \mathbb{E}[FP_1(\Omega^*)].$$

Notice that if the dilution function satisfies hypotheses 1 and 2, theorems 2 and 3 imply that ordered pooling minimizes both the expected number of false negatives and the expected number of false positives. So in this case, not only can we bound the max-min component of the welfare function (i.e., the equity component), but also the utilitarian component (i.e., the Pareto efficiency component).

**Corollary 5** Let  $\Omega^* = \{G_1^*, G_2^*, \dots, G_m^*\}$  be an ordered partition of  $S = \{1, 2, \dots, n\}$  such that  $|G_g^*| = k \forall G_g^* \in \Omega^*$ , and let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be any other partition such that  $|G_g| = k \forall G_g \in \Omega$ . Suppose that hypotheses 1 and 2 hold. Then:

$$\pi_\alpha(u_1(\Omega), u_2(\Omega), \dots, u_n(\Omega)) \leq -\alpha (\theta \mathbb{E}[FN_n(\Omega^*)] + (1 - \theta) \mathbb{E}[FP_1(\Omega^*)]) + (1 - \alpha) \sum_{i \in S} u_i(\Omega^*).$$

## 8 Sufficient conditions for the dilution function to be discrete-concave

We saw that, when the dilution function is discrete-concave, then, for any partition, we can find an ordered partition with the same pool size configuration that yields a lower expected number of tests, as well as an ordered partition that yields a lower expected number of false positives (theorems 1 and 3). When the dilution function is not readily available, it can still be estimated by looking at the distribution of a continuous measure used to detect infection. Indeed, most diseases are detected by measuring a continuous biomarker, such as the viral load: if the measure is above a certain threshold, infection is detected. By comparing the distribution of biomarker concentration among infected vs. non infected specimens, and by assuming that the biomarker concentration from a pooled specimen is given by the average concentration of the specimens within the sample, one can then compute the probability that the biomarker concentration from the group is above a certain threshold conditional on the number of infected within the group. In this section, we prove the intuitive result that, as long as the the average biomarker concentration from infected subjects is sufficiently high, the dilution function will not be too strong,

causing it to be discrete-concave, in which case ordered partitions perform well in terms of minimizing the expected number of tests and the expected number of false positives (but not necessarily the expected number of false negatives).

Suppose that the biomarker concentration of an infected subject is given by a random variable  $X_+$ , while the biomarker concentration from a non-infected subject is given by  $X_-$ . Suppose that  $X_+ \sim N(\mu_+, \sigma_+^2)$  and that  $X_- \sim N(\mu_-, \sigma_-^2)$ , where  $\mu_+ > \mu_-$ . Let  $Y_{I,k}$  be the biomarker concentration of a pooled sample of  $k$  subjects, of which exactly  $I \leq k$  are infected. Following Wang, McMahan and Gallagher [2015] and Mokalled et al. [2021], we assume that the biomarker concentration of a pooled sample is given by the average biomarker concentration of subjects within the sample. In this case,  $Y_{I,k} \sim N(\mu_I, \sigma_I^2)$ , where

$$\mu_I = \frac{I\mu_+ + (k-I)\mu_-}{k}$$

and

$$\sigma_I = \sqrt{\frac{I\sigma_+^2 + (k-I)\sigma_-^2}{k^2}}.$$

Suppose that there is an exogenous threshold  $\underline{y} \in [\mu_-, \mu_+]$ , above which infection is detected. Though we can make  $\underline{y}$  a function of the pool size  $k$ , we will assume the threshold to be the same for all pool sizes to avoid clutter notation. In this case, one can show that the dilution function is given by

$$h(I, k) = \int_{\left(\frac{\underline{y}-\mu_I}{\sigma_I}\right)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

Notice that, in order to estimate this dilution function, one only needs to estimate  $\mu_+$ ,  $\mu_-$ ,  $\sigma_+$  and  $\sigma_-$ . These parameters can be easily estimated using data on the viral load or similar biomarker from infected and non-infected subjects, as in Wang, McMahan and Gallagher [2015] and Mokalled et al. [2021]. We now present the main result from this section.

**Proposition 4** *Suppose that  $\sigma_+ \geq \sigma_-$ . If  $\frac{1}{k}\mu_+ + \frac{(k-1)}{k}\mu_- > \underline{y}$ , then the dilution function  $h(\cdot, k)$  is increasing and discrete-concave.*

In words, proposition 4 tells us that, if the average viral load from infected subjects is significantly high, we should observe a small dilution effect, so that the dilution function must be discrete-concave. More precisely, the expression  $\frac{1}{k}\mu_+ + \frac{(k-1)}{k}\mu_-$  corresponds to the expected viral load from a pooled sample in which only one subject is infected. If this average is higher than the threshold point above which infection is detected, then the dilution function should not be “too strong”. Notice that this is a rather mild hypothesis. Indeed, if we had  $\frac{1}{k}\mu_+ + \frac{(k-1)}{k}\mu_- < \underline{y}$ , then pooled tests from samples that contained only one infected specimen would fail to detect infection in more than 50% of the cases (assuming a Gaussian specification). Under small prevalences, this implies that the pooled tests would more often fail than succeed detecting infected samples, in which case the tester should either consider reducing the cutoff point above which infection is detected, or reduce the pool size.

The hypothesis that  $\sigma_+ \geq \sigma_-$  usually holds in practical applications, due to the fact that the biomarker reading from patients with chronic infection often change dramatically over time. Such is the case for HIV infections (e.g., Alizon and Magnus [2012] and Maartens, Celum and Lewin [2014]) and Hepatitis B infections (e.g., Yapali, Talaat and Lok [2014]).

Notice that, when estimating a dilution function through this method, one can find a function  $h(\cdot, k)$  that is not concave for small and non-integer values of  $I$ , though it is discrete-concave, as illustrated by example 3 below.

**Example 3** *If  $\mu_+ = 20$ ,  $\mu_- = 1$ ,  $\sigma_+ = \sigma_- = 1$ ,  $\underline{y} = 4$  and  $k = 5$ , we have that the dilution function is discrete-concave, though it is not concave for values of  $I \in (0, 1)$ , as displayed in figure 4 below.*

## 9 Case Study: Chlamydia Screening in the United States

Chlamydia is the most reported sexually transmitted disease in the US. It affects mostly the young and sexually active, and symptomatic cases are more frequent among women. Though in general it can be cured with antibiotics,

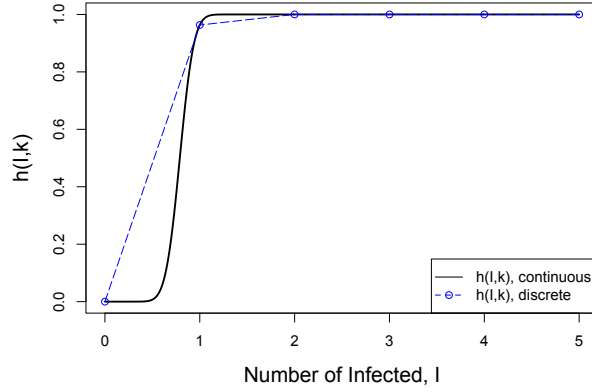


Figure 4: Dilution function when  $\mu_+ = 20$ ,  $\mu_- = 1$ ,  $\sigma_- = \sigma_+ = 1$ ,  $\underline{y} = 4$  and  $k = 5$ . The blue dashed lines depicts the dilution function evaluated at the possible values  $I$  can assume in practice,  $\{0, 1, \dots, 5\}$ , while the solid curve corresponds to the dilution function if the number of infected was not necessarily an integer number.

if left untreated the disease can cause severe sequelae in women, including Pelvic Inflammatory Disease (PID), ectopic pregnancy and infertility (Centers for Disease Control and Prevention [2019a]).

In this section we consider the problem of minimizing the expected cost of screening a batch of samples for Chlamydia, i.e., we consider the problem of minimizing equation (8). For this analysis, we will consider the Ligase Chain Reaction (LCR) chlamydia test, which is commonly used to screen for chlamydia (Arahamian, Bish and Bish [2018]). We compare the performance of the optimal ordered partition (OR) with the following alternative heuristics:

1. **Optimal ordered partition ignoring dilution effects ( $\widehat{\text{OR}}$ ):** The tester implements the optimal ordered partition ignoring the existence of dilution effects. More precisely, if the true dilution function is given by  $h(I, k)$ , with  $h(1, 1) = S_e$  and  $h(0, 1) = 1 - S_p$ , the tester (incorrectly) assumes that the dilution function is given by

$$\widehat{h}(I, k) = \begin{cases} S_e & \text{if } I > 0 \\ 1 - S_p, & \text{if } I = 0, \end{cases}$$

and implements the optimal ordered partition using this dilution function.

2. **Random pooling (R1):** Subjects are randomly pooled into groups of homogeneous size. If the batch size is not a multiple of the pool size, a group of subjects is randomly selected to form the smaller pool. We follow Arahamian, Bish and Bish [2019] and Arahamian, Bish and Bish [2018] by using a static framework, i.e., the pool size is fixed and chosen to minimize the cost function before observing subjects' realized probabilities of infection.
3. **Random pooling with a cutoff (R2):** According to this heuristic, subjects with probability of infection above a certain threshold are individually tested; the remaining subjects are randomly assigned into groups of homogeneous size. If the fraction of subjects who are pooled tested is not a multiple of the pool size, a subgroup of these subjects is randomly selected to form the smaller pool. Like in the previous case, we use a static framework: i.e., the pool size and threshold are fixed and chosen to minimize the cost function before observing subjects' realized probabilities of infection.
4. **Individual testing (IT):** Every subject is individually tested, and there are no followup tests. Notice that, by construction individual testing (IT) cannot yield a lower cost than the optimal ordered partition, as individual testing is, technically speaking, a special type of ordered partition (OR).

Following Aprahamian, Bish and Bish [2019], we use the average between the cost of sequelae for men and women, estimated by Owusu-Edusei Jr et al. [2015], to set the cost of a false negative at \$2,927.<sup>8</sup> Also following Aprahamian, Bish and Bish [2019], we set the screening cost per test at \$55, and the cost of a false positive equal to the cost of an additional screening test (\$55). So the total expected cost of implementing a partition  $\Omega$  is given by

$$\mathbb{E}[C(\Omega)] \equiv 2,927\mathbb{E}[FN(\Omega)] + 55\mathbb{E}[FP(\Omega)] + 55\mathbb{E}[T(\Omega)].$$

The prevalence per demographic group was extracted from the Centers of Disease and Control Prevention (CDC) website for the year 2014.<sup>9</sup> Following Aprahamian, Bish and Bish [2019] and Aprahamian, Bish and Bish [2018], we multiply the prevalence of each group reported by CDC by 3, in order to account for under-reporting.<sup>10</sup> Table 3 reports the prevalence of Chlamydia for each group in the sample.

Table 3: Prevalence of Chlamydia and proportion in population by Gender, Age and race/Ethnicity (Centers for Disease Control and Prevention 2014).

Gender	Race/ Ethnicity	Age Group (years)	Risk (prevalence) (%)	Proportion in general population (%)
Female	Hispanic	15-24	6.54	1.41
		Other	0.65	7.01
	Black	15-24	19.19	1.07
		Other	1.22	5.67
	Other	15-24	4.38	4.29
		Other	0.25	31.31
Male	Hispanic	15-24	1.78	1.53
		Other	0.36	7.16
	Black	15-24	7.45	1.09
		Other	1.05	5.08
	Other	15-24	1.20	4.51
		Other	0.17	29.87

Following Aprahamian, Bish and Bish [2018], we assumed the following functional format for the dilution effect:

$$h(I, k) = (1 - S_p) + (S_e + S_p - 1)(I/k)^\delta. \quad (10)$$

Data from Kacena et al. [1998a] suggest a high sensitivity and high specificity for pools of size of at least 4 subjects, so we set  $S_p = .98$ ,  $S_e = .99$  (details provided in the online Appendix). The parameter  $\delta$  governs how strong the dilution effect is: the higher  $\delta$  is, the stronger the dilution effect. We conservatively set  $\delta \geq 0.1$ , which is relatively high compared to what the dataset indicates, to understand the performance of ordered pooling in a scenario in which dilution effects are non-trivial. Figure 5 plots our calibrated dilution function for  $k = 10$  and  $\delta \in \{0.1, 0.15, 0.2\}$ .

It can be shown that the calibrated dilution function from equation (10) is concave for any  $\delta \in [0, 1]$  and any  $S_p, S_e \in [0, 1]$  such that  $S_e > 1 - S_p$ . Moreover, for any  $\delta \geq 0$ , this dilution function satisfies hypothesis 1. Therefore, if  $\delta \in [0, 1]$ , theorems 1, 2 and 3 indicate the optimal partition is likely an ordered partition. Indeed, if we were not implementing an ordered partition, we would be able to either reduce the expected number of tests, or the expected number of false negatives, or the expected number of false positives (or perhaps all 3 attributes simultaneously). Because our results do not guarantee that the optimal partition is necessarily an ordered partition, we also compute a lower bound to the objective function, by computing a lower bound to each attribute separately (see section 6).

Conducting simulations using the data from table 3, we can confirm that the optimal ordered partition performs significantly better than the 4 alternative methods. The results of these simulations are depicted in table 4. For the simulations, we assumed  $n = 100$ , which corresponds to the approximate number of specimens that

<sup>8</sup>For simplification, this measure does not incorporate health-related reduction in quality of life, nor costs associated with transmissions that a true positive would have averted.

<sup>9</sup><https://wonder.cdc.gov/std-race-age.html>

<sup>10</sup>In part, many end up not being screened because they exhibit no symptoms: approximately 75% of women and 50% of infected men exhibit no symptoms (Centers for Disease Control and Prevention [2000]).

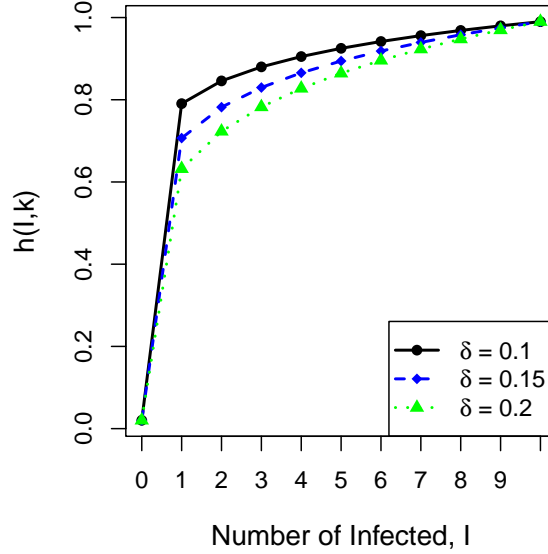


Figure 5: Calibrated dilution function for Chlamydia (eq. (10)) with parameters  $S_p = .98$  and  $S_e = .99$ , and pool size  $k = 10$ .

some state health laboratory tests in a day (Aprahamian, Bish and Bish [2019]). The values displayed in the table correspond to the averages taken from 400 simulations.

Consistent with proposition 1, individual testing ( $IT$ ) generates less false negatives than any other method considered. But because individual testing requires more tests and produces more false positives, it generates the highest expected cost per subject (i.e., a higher  $\mathbb{E}[C]/n$ ). Not surprisingly, as the dilution effect increases, the benefits of taking into account the existence of dilution effects increases (i.e., the difference in total cost when implementing  $OR$  vs  $\widehat{OR}$  increases). Also notice that, as the parameter  $\delta$  increases, the average pool sizes decrease (i.e.,  $\mathbb{E}[k]$  decreases). This happens because, under stronger dilution effects, larger pools are more more likely to generating false negatives, in which case the tester should reduce the average pool size.

Though the results from 4 provide strong evidence that ordered pooling can be significantly less costly than random pooling ( $R1$ ), notice that these results were obtained under the assumption that the tester was allowed to choose heterogenous pool sizes under ordered pooling, while being forced to select the same pool size under random pooling. In some practical applications, however, changing pool sizes may be costly, thus forcing the tester to use the same pool size for every test. So in figures 6a to 6d we display the performance of ordered pooling, random pooling and individual testing in terms of  $\mathbb{E}[T(\Omega)]$ ,  $\mathbb{E}[FN(\Omega)]$ ,  $\mathbb{E}[FP(\Omega)]$  and  $\mathbb{E}[C(\Omega)]$ , respectively, when all pools are required to have the same size  $k$  and the total number of subjects to be tested is given by  $n = 10,000$ ,<sup>11</sup> and  $\delta = 0.15$ . Under ordered pooling, whenever the population size  $n$  was not a multiple of the pool size  $k$ , the subjects with highest probability of infection were grouped into the smaller group. From these figures, we can see that ordered pooling performs better than random pooling in all of the three attributes considered and for all pool sizes, which is consistent with corollary 4, as our calibrated dilution function is discrete-concave and satisfied hypothesis 2. Notice, however, that substantial differences in costs are only noticeable when the pool size is sufficiently large (e.g., when  $k \geq 10$ ), as displayed in figure 6d. In particular, the optimal pool size under ordered pooling (assuming homogeneous pool size) is 13, which yields an expected cost per subject of \$17.01. Meanwhile, the optimal pool size under  $R1$  is 10, which yields an expected cost of \$18.58. Therefore, compared to  $R1$ , ordered pooling reduces costs in approximately 8%.

Regarding equity, for each  $k \geq 2$  we used proposition 3 to compute the upper bound for the minimum utility that can be achieved when pools are of homogeneous size when  $\theta = \lambda_1 / (\lambda_1 + \lambda_2) \approx 0.98$  and  $\alpha \in \{0, 1\}$ , and cross-compared it with the minimum utility obtained under ordered pooling and under random pooling. As displayed in figures 7a and 7b, for all  $k \geq 2$ , the average welfare obtained under ordered pooling was significantly higher than the welfare obtained under random pooling. For  $\alpha = 0$  total welfare under ordered pooling perfectly matches the

<sup>11</sup>Having a large number of subjects being tested helps to smooth out the graphs for instances in which  $n$  is not a multiple of  $k$ .

Table 4: Performance measure of the optimal ordered partition (*OR*) compared to the optimal random partition (*R1*), the optimal *random partition with a cutoff* (*R2*) and the lower bound for each of the attributes considered (*LB*) for the chlamydia case study.

Using the dilution function (10) assuming $\delta = 0.1$							
Model	$\mathbb{E}[FN]$	$\max\{\mathbb{E}[FN_i]\}$	$\mathbb{E}[FP]$	$\mathbb{E}[T]$	$\mathbb{E}[C]/n$	$\mathbb{E}[\max\{k_i\}]$	$\mathbb{E}[k]$
<i>OR</i>	0.1375	0.0073	0.1505	16.0587	12.9397	27.54	11.479
$\widehat{OR}$	0.1682	0.0157	0.147	15.3974	13.4715	24.835	13.5915
% ( $\widehat{OR} - OR$ )	18.2%	53.8%	-2.4%	-4.3%	3.9%	-10.9%	15.5%
<i>R1</i>	0.2063	0.0319	0.1702	19.1972	16.6898	10.0	10.0
% ( <i>R1-OR</i> )	33.3%	-339.7%	11.6%	16.3%	22.5%	-175.4%	-14.8%
<i>R2</i>	0.1053	0.0043	0.2703	20.2703	14.3803	16.0	7.3729
% ( <i>R2-OR</i> )	-30.5%	-40.3%	44.3%	20.8%	10.0%	-72.1%	-55.7%
<i>IT</i>	0.0097	0.0015	2.0006	100.0	56.3843	1	1
% ( <i>IT-OR</i> )	-1317.5%	-372.1%	92.5%	83.9%	77.1%	-2654.0%	-1047.9%
<i>LB</i>	0.0099	-	0.0574	15.1818	8.6712	-	-
% ( <i>OR-LB</i> )	92.8%	-	61.8%	5.5%	33.0%	-	-
Using the dilution function (10) assuming $\delta = 0.15$							
Model	$\mathbb{E}[FN]$	$\max\{\mathbb{E}[FN_i]\}$	$\mathbb{E}[FP]$	$\mathbb{E}[T]$	$\mathbb{E}[C]/n$	$\mathbb{E}[\max\{k_i\}]$	$\mathbb{E}[k]$
<i>OR</i>	0.174	0.0079	0.1528	15.9075	13.9256	29.935	10.8153
$\widehat{OR}$	0.228	0.0205	0.1357	14.6482	14.8047	24.9675	13.6825
% ( $\widehat{OR} - OR$ )	23.7%	61.5%	-12.6%	-8.6%	5.9%	-19.9%	21.0%
<i>R1</i>	0.2851	0.0439	0.1563	18.4273	18.5651	10.0	10.0
% ( <i>R1-OR</i> )	39.0%	-456.0%	2.2%	13.7%	25.0%	-199.3%	-8.2%
<i>R2</i>	0.1494	0.0063	0.2696	19.2129	15.0885	19.0	7.9943
% ( <i>R2-OR</i> )	-16.4%	-20.8%	43.3%	17.2%	7.7%	-57.6%	-35.3%
<i>IT</i>	0.0097	0.0015	2.0006	100.0	56.3843	1	1
% ( <i>IT-OR</i> )	-1693.6%	-428.0%	92.4%	84.1%	75.3%	-2893.5%	-981.5%
<i>LB</i>	0.0097	-	0.0566	14.1985	8.1256	-	-
% ( <i>OR-LB</i> )	94.4%	-	62.9%	10.7%	41.7%	-	-
Using the dilution function (10) assuming $\delta = 0.2$							
Model	$\mathbb{E}[FN]$	$\max\{\mathbb{E}[FN_i]\}$	$\mathbb{E}[FP]$	$\mathbb{E}[T]$	$\mathbb{E}[C]/n$	$\mathbb{E}[\max\{k_i\}]$	$\mathbb{E}[k]$
<i>OR</i>	0.1945	0.0073	0.1701	16.1333	14.6593	32.4475	9.9601
$\widehat{OR}$	0.2796	0.0263	0.1239	13.9905	15.946	24.8475	13.7188
% ( $\widehat{OR} - OR$ )	30.4%	72.1%	-37.2%	-15.3%	8.1%	-30.6%	27.4%
<i>R1</i>	0.3555	0.0535	0.1438	17.7404	20.2408	10.0	10.0
% ( <i>R1-OR</i> )	45.3%	-630.2%	-18.3%	18.0%	27.6%	-224.5%	0.4%
<i>R2</i>	0.1759	0.0071	0.2485	19.1208	15.8003	16.0	7.4578
% ( <i>R2-OR</i> )	-10.6%	-3.3%	31.5%	15.6%	7.2%	-102.8%	-33.6%
<i>IT</i>	0.0097	0.0015	2.0006	100.0	56.3843	1	1
% ( <i>IT-OR</i> )	-1904.9%	-390.5%	91.5%	83.9%	74.0%	-3144.8%	-896.0%
<i>LB</i>	0.0095	-	0.0557	13.2306	7.5851	-	-
% ( <i>OR-LB</i> )	95.1%	-	67.2%	18.0%	48.3%	-	-

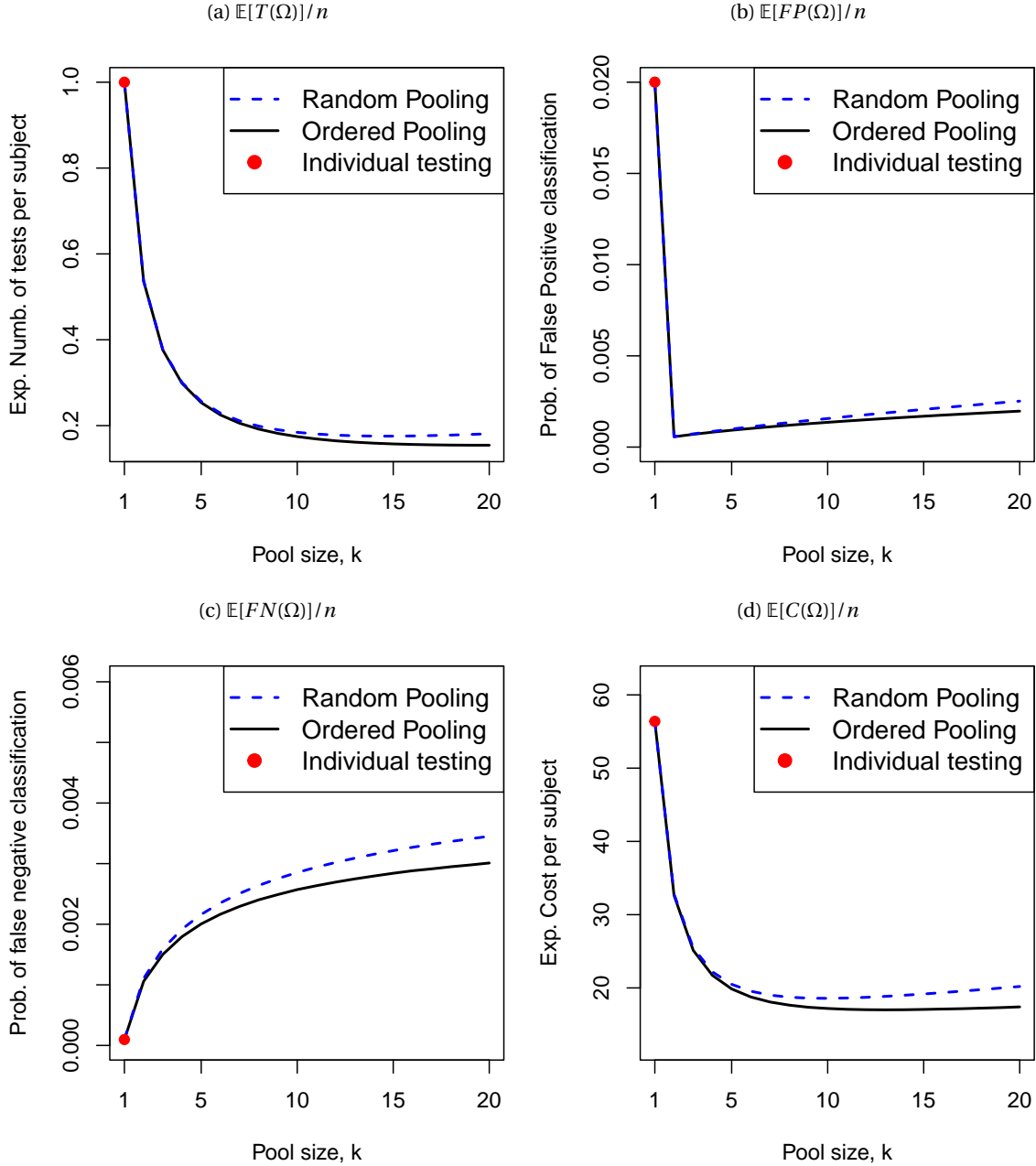


Figure 6: Expected number of tests and classification errors per subject for Chlamydia tests using ordered pooling with homogenous pool sizes and random pooling ( $R1$ ). The red dot corresponds to the case in which subjects are individually tested, i.e.  $k = 1$ . Parameters of the dilution function (10):  $S_e = .99$ ,  $S_p = .98$  and  $\delta = 0.15$ .

upper bound. This is because the calibrated dilution function satisfies hypotheses 1 and 2, which implies, from corollary 4, that ordered pooling maximizes the utilitarian component of the utilitarian max-min welfare function. Though total welfare under ordered pooling does not match perfectly the upper bound when  $\alpha = 1$ , they are very close to one another, and, in addition, ordered pooling performs significantly better than random pooling.

Notice also that individual testing is the most equitable allocation for both  $\alpha = 1$  and  $\alpha = 0$ . Intuitively, when

$\alpha = 1$ , the tester only cares about the minimum utility of subjects being tested. Given our parameters, the subject with lowest utility is usually the one with highest probability of infection. This subject is more affected by the probability of receiving a false negative, as compared to the probability of receiving a false positive. So individually testing this subject maximizes the minimum utility, which is why individual testing performs well in terms of equity. Similarly, when  $\alpha = 0$ , i.e., when the tester only cares about the sum of subjects' utility, individual testing is still optimal. This happens because the cost of a false negative is much higher than the cost of a false positive (\$2,927 vs \$55), and because subjects' welfare function does not internalize the costs of testing subjects. If the testing costs were added to the welfare function, then pooled testing would maximize the utilitarian component of the welfare function, as depicted in figure 6d.

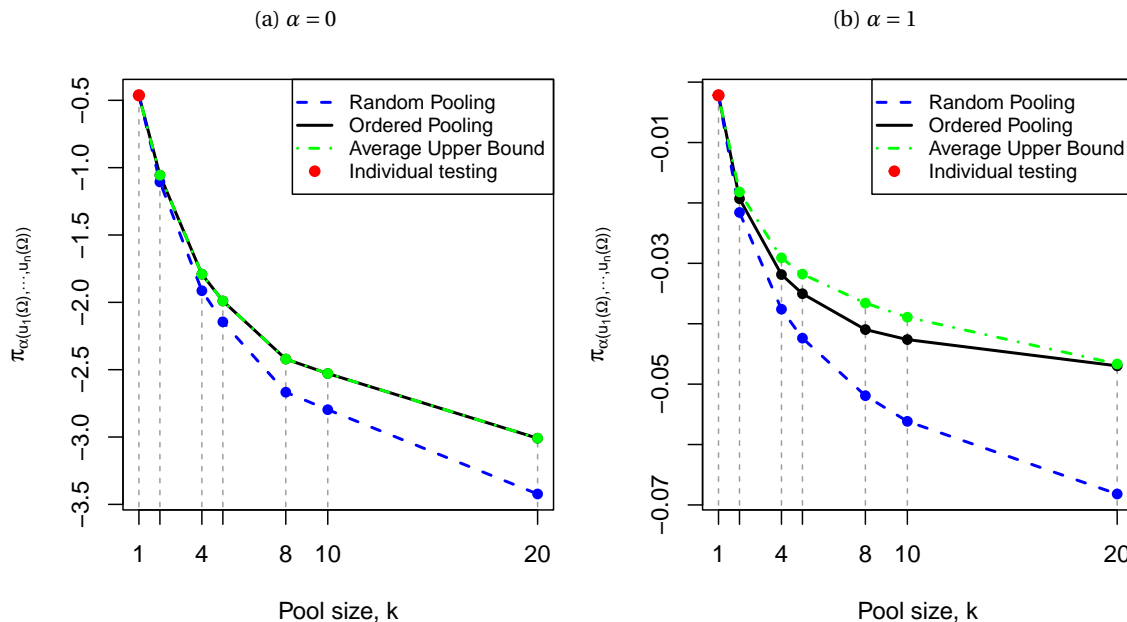


Figure 7: Subjects' welfare under ordered pooling vs random pooling for Chlamydia tests, for values of  $k \leq 20$  that are dividends of  $n = 1000$ . In both simulations we have  $\theta = \lambda_1 / (\lambda_1 + \lambda_2) \approx 0.98$ . Parameters of the dilution function (10):  $S_e = .99$ ,  $S_p = .98$  and  $\delta = 0.15$ .

## 10 Case Study: Hepatitis B Screening in Irish prisons

In this section we estimate the dilution effect of pooled testing for Hepatitis B using surveyed data on Irish prisoners with information on a continuous biomarker reading for individual tests. We then compare the performance of the optimal ordered partition when applied to this dataset.

Hepatitis B is a liver infection disease caused by the hepatitis B virus (HBV). In 2015, 257 million people worldwide were estimated to be living with chronic hepatitis B (World Health Organization [2017]). Left untreated, the disease can lead to deadly sequelae, most commonly cirrhosis and hepatocellular carcinoma (primary liver cancer). In 2015 approximately 890,000 people were estimated to have died due to complications from the disease (World Health Organization [2017]). Because hepatitis B can be transmitted by blood, the WHO recommends that all blood donations be tested for hepatitis B. Due to budget constraints, it is common for blood centers to perform pooled tests on multiple donors to detect HBV (e.g., El-Amine, Bish and Bish [2017]).

In this case study, we assume that the tester's objective is to screen blood donors for HBV, so as to prevent blood recipients from getting infected. So we will set the cost of a false negative equal to the expected cost of infecting a blood recipient with HBV. Because some infected patients may become asymptomatic, while others may exhibit acute symptoms, which may then progress to chronic infection and then possibly death, we use a

simple Markov specification to model how patients transition from each of these states (details are provided in the online Appendix).<sup>12</sup> Using data from Jackson et al. [2003] and Birkmeyer et al. [1993], we estimate the cost of infecting a blood recipient (and therefore, the cost of a false negative) to be equal to  $\lambda_1 = \$1,811.26$ .

Using data from Chahal et al. [2018], we set the cost of each test at \$16.41 (the cost of a core antigen test), and the cost of a false positive equal to the cost of collecting, handling and testing another blood sample to replace the one that was discarded (\$95, according to Birkmeyer et al. [1993]).

So the total expected cost of implementing a partition  $\Omega$  is given by

$$\mathbb{E}[C(\Omega)] \equiv 1811.26\mathbb{E}[FN(\Omega)] + 95\mathbb{E}[FP(\Omega)] + 16.41\mathbb{E}[T(\Omega)].$$

To estimate the dilution effect of pooled samples, we use the dataset from a survey conducted in 1998 on 5 different prisons from Ireland, in which inmates were individually tested for Hepatitis B and other infectious diseases to determine risk factors for infection. The results of this survey have been summarized by Allwright et al. [2000]. As prisons are arguably not the best place to find reliable blood donors, given that blood borne infections, such as HBV, are highly prevalent among inmates (e.g., Smith et al. [2017]), our numerical analysis based on this dataset only provides a conservative estimate of the benefits of implementing pooled testing as opposed to individual testing, as pooled testing is usually more effective when used to screen populations with a low prevalence rate (Kim et al. [2007]). Indeed, under small or no dilution effects, we have that, if the prevalence of a disease is high, then each group is likely to have at least one infected subject. So under a high prevalence rate, most pools end up being retested regardless, in which case it would be more economical to just implement individual testing as opposed to Dorfman testing.

The dataset contains a classification of whether an inmate was infected with Hepatitis B, as well as the Optical Density (OD) readings from a Murex ICE enzyme immunoassay (on oral fluid samples) used to detect the presence of the core antigen for hepatitis B (HBcAg).

In total there were 99 infected, and 1,038 non-infected subjects, which amounts to a prevalence rate of approximately 8.7%. Figure 8 displays the histogram of OD readings from both infected and non-infected subjects. While the distribution of OD readings for the non-infected inmates exhibits a unimodal format, the distribution of OD readings for infected inmates is bimodal. The bimodal distribution of the biomarker among infected subjects is consistent with other reports of the distribution of a continuous biomarker used to detect HBV among infected subjects (e.g., Downs et al. [2020] and Alcalde et al. [2009]).

Let  $X_i$  be the random variable corresponding to the OD reading that subject  $i \in S$  would get under individual testing. The distribution of  $X_i$  is expected to depend on whether the subject is infected or not. Following Wang, McMahan and Gallagher [2015] and Mokalled et al. [2021] we assume that the distribution of the OD reading of a group  $G_g$  is given by the average of the OD readings from each subject in that group, i.e., by  $\sum_{i \in G} X_i / |G_g|$ .

Because the distribution of OD readings of infected subjects is bimodal, we used bootstrap to estimate the cumulative densities of OD readings from a pooled sample. More precisely, for a pool of size  $k$  with  $I \leq k$  infected subjects, we make  $I$  random draws with replacement of OD readings from infected subjects, and  $k - I$  draws with replacement from non-infected subjects, and then take the average of the OD readings from these subjects. We repeat this process 1000 times for each possible combination of  $I$  and  $k$ . We then use these trials to estimate the kernel density of the joint OD readings, and then integrate the estimated density to recover the cumulative distribution of joint OD readings. The estimated cumulative distribution evaluated at a certain cutoff point  $OD^*$  then gives us the probability that infection is *not* detected conditional that the pool has  $k$  subjects, of which exactly  $I \leq k$  are infected. For the sake of comparison, we also estimate the dilution function adopting a parametric approach, by assuming that the OD readings from both infected and non-infected subjects is normally distributed, as described in section 8.

Because the dataset did not include the cutoff values for OD readings above which a subject was classified as infected, we computed our own thresholds as follows: assuming all subjects have the same probability of infection given by  $\mu_r$ , we compute the threshold from each pool size that would minimize the expected cost of testing that group. So the threshold above which infection is detected depends on the pool size. Notice that this approach is optimal under random pooling ( $R1$ ) but it is not necessarily optimal under ordered pooling. This gives us a conservative estimate of the benefits of implementing the optimal ordered partition ( $OR$ ), vs. the optimal random partition ( $R1$ ).

---

<sup>12</sup>For simplification, our approach does not incorporate costs associated with transmissions that a true positive would have averted.

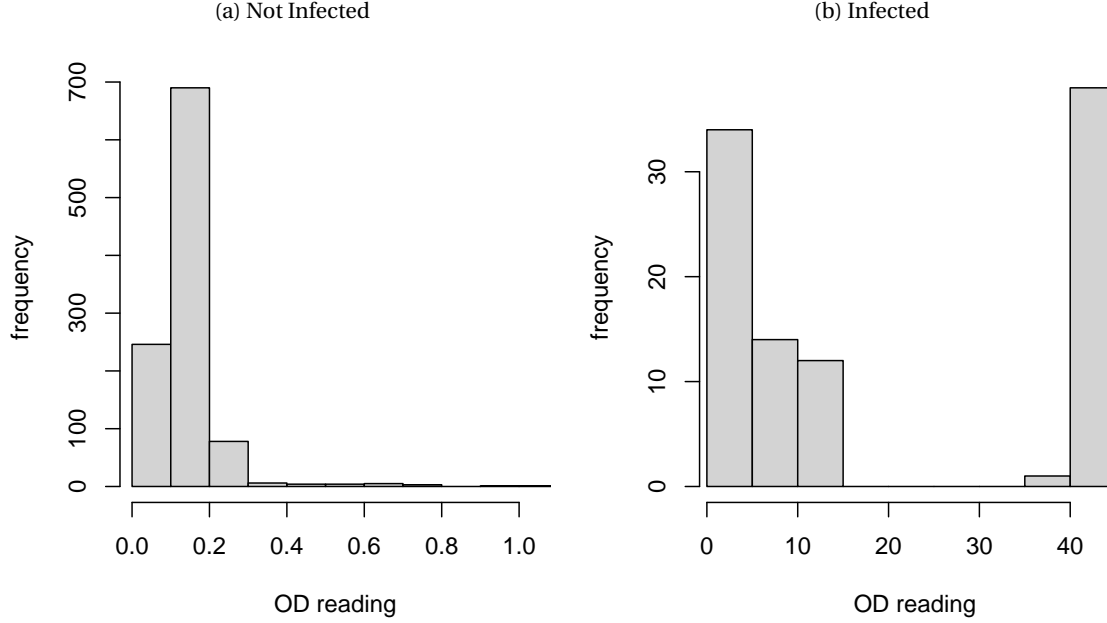


Figure 8: Histogram of OD readings for Hepatitis B of infected and non-infected inmates using the Irish Prisoner dataset collected by Allwright et al. [2000].

For our dataset, both parametric and non-parametric estimates yield a discrete-concave dilution function for all possible pool sizes, which implies that ordered pooling is expected to perform well in terms of minimizing the expected number of tests and the expected number of false positives. This happens because the mean and variance of the OD readings from infected subjects are much higher than the corresponding statistics from non-infected subjects, which implies, from proposition 4, that the dilution function should be discrete-concave. Indeed, our parametric estimation suggests that the distribution of OD readings from infected subjects is given by  $X_+ \sim N(19.87, 339.7)$ , while the distribution of OD readings from non-infected subjects is given by  $X_- \sim N(0.14, 0.007)$ . Hypothesis 1, however, is not always satisfied, so we cannot guarantee, without the aid of numerical exercises, that ordered pooling performs well in terms of minimizing the expected number of false negatives. Figure 9 depicts the estimated dilution function when  $k = 10$  using both the parametric and non-parametric approach. Notice that, under both approaches, the estimated dilution effect is very small, especially when using the non-parametric approach.

As the prevalence of HBV is highly dependent on age (e.g., Centers for Disease Control and Prevention [2019b] and World Health Organization [2017]), we estimate a logistic regression of the probability of infection as a function of age, using the following quadratic specification

$$\text{logit}(\text{Prob}(Y_i = 1 | \text{Age}_i)) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2 + \beta_3 \text{Age}_i^3 + \varepsilon_i,$$

where  $Y_i$  corresponds to the dummy variable that equals 1 if inmate  $i$  is infected, and zero otherwise.  $\text{Age}_i$  corresponds to the inmate's age, and  $(\varepsilon_i)_{i \in S}$  are normally and independently distributed error terms. Figure 10 displays the estimated probability of infection as a function of age.

Using these estimated probabilities, we conducted simulations comparing the performance of the optimal ordered partition (OR) with the 4 heuristics described in section 9:  $\widehat{OR}$ ,  $IT$ ,  $R1$  and  $R2$ . Similar to section 9, we assume the batch size to be equal to  $n = 100$ , and we take the averages from 400 simulations.

The results of these simulations are displayed in table 5. As can be seen from the table, there are very small differences between  $OR$  and  $\widehat{OR}$  under both the parametric and non-parametric specification. This is because our estimations suggest a very small dilution effect, so ignoring it should not have a big impact on the optimal ordered partition. Because the prevalence rate is so high (approximately 8.7%), most subjects should be tested individually, as in this case retests would be too frequent under Dorfman testing. This is why, under the parametric approach,

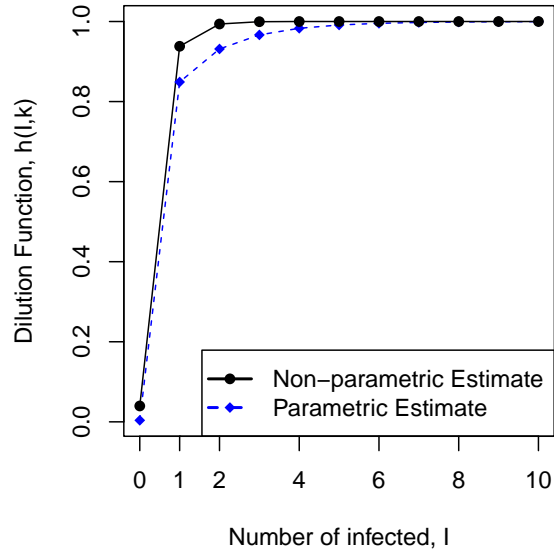


Figure 9: Calibrated dilution function for Hepatitis B when  $k = 10$ .

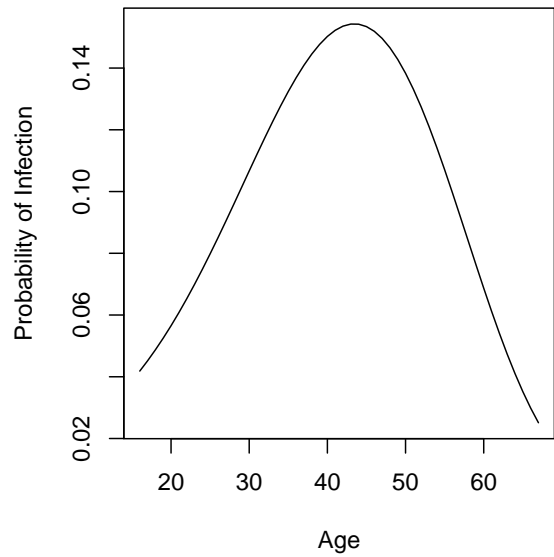


Figure 10: Estimated probability of Hepatitis B infection of inmates as a function of age.

individual testing ( $IT$ ) performs almost as well as the optimal ordered partition ( $OR$ ).

Regardless of how the dilution function is estimated, the optimal pool sizes are usually very small, with many subjects being individually tested. This is consistent with previous studies that suggest that, under a small prevalence rate, the benefits of implementing Dorfman testing as opposed to individual testing are very limited. In the online Appendix we conduct these same simulations after multiplying the estimated probability of infection of each subject by 0.1. In this case, the optimal pool sizes are much larger, so we observe significant gains in implementing  $OR$  as opposed to  $IT$ .

Table 5: Performance measure of the optimal ordered partition (*OR*) compared to the optimal random partition (*R1*), the optimal *random partition with a cutoff* (*R2*) and the lower bound for each of the attributes considered (*LB*) for the Hepatitis B case study.

Non-parametric estimation of $h(I, k)$							
Model	$\mathbb{E}[FN]$	$\max\{\mathbb{E}[FN_i]\}$	$\mathbb{E}[FP]$	$\mathbb{E}[T]$	$\mathbb{E}[C]/n$	$\mathbb{E}[\max\{k\}]$	$\mathbb{E}[k]$
<i>OR</i>	0.31	0.0051	0.3313	59.9333	15.7651	4.004	3.2928
$\widehat{OR}$	0.3246	0.0051	0.3884	58.7828	15.8944	5.0	3.9056
% ( $\widehat{OR} - OR$ )	4.5%	0.0%	14.7%	-2.0%	0.8%	19.9%	15.7%
<i>R1</i>	0.3062	0.0168	0.3205	61.907	16.0102	3.0	2.9412
% ( <i>R1-OR</i> )	-1.2%	230.8%	-3.4%	3.2%	1.5%	-33.5%	-12.0%
<i>R2</i>	0.2848	0.0087	0.386	63.6346	15.9672	3.0	2.5336
% ( <i>R2-OR</i> )	-8.9%	71.3%	14.2%	5.8%	1.3%	-33.5%	-30.0%
<i>IT</i>	0.1466	0.0026	1.4369	100.0	20.4297	1.0	1.0
% ( <i>IT-OR</i> )	-111.5%	-96.3%	76.9%	40.1%	22.8%	-300.4%	-229.3%
<i>LB</i>	0.1459	-	0.1853	58.5312	12.4244	-	-
% ( <i>OR-LB</i> )	52.9%	-	44.1%	2.3%	21.2%	-	-
Parametric estimation of $h(I, k)$							
Model	$\mathbb{E}[FN]$	$\max\{\mathbb{E}[FN_i]\}$	$\mathbb{E}[FP]$	$\mathbb{E}[T]$	$\mathbb{E}[C]/n$	$\mathbb{E}[\max\{k\}]$	$\mathbb{E}[k]$
<i>OR</i>	1.3097	0.0224	0.0457	95.0484	39.3629	7.765	1.0786
$\widehat{OR}$	1.2812	0.0224	0.0478	98.3785	39.3953	3.0975	1.022
% ( $\widehat{OR} - OR$ )	-2.2%	0.0%	4.3%	3.4%	0.1%	-150.7%	-5.5%
<i>R1</i>	1.2678	0.0225	0.0489	100.0	39.4197	1.0	1.0
% ( <i>R1-OR</i> )	-3.3%	0.2%	6.5%	5.0%	0.1%	-676.5%	-7.9%
<i>R2</i>	1.3107	0.031	0.0454	94.6867	39.3205	7.3925	1.0804
% ( <i>R2-OR</i> )	0.1%	38.1%	-0.8%	-0.4%	-0.1%	-5.0%	0.2%
<i>IT</i>	1.2678	0.0224	0.0489	100.0	39.4197	1.0	1.0
% ( <i>IT-OR</i> )	-3.3%	-0.0%	6.5%	5.0%	0.1%	-676.5%	-7.9%
<i>LB</i>	1.2681	-	0.0038	50.8113	31.3102	-	-
% ( <i>OR-LB</i> )	3.2%	-	91.8%	46.5%	20.5%	-	-

## 11 Discussion

We derived sufficient conditions under which ordered pooling minimizes the expected number of tests and both types of classification errors. In order to check whether these conditions are met, one only needs to estimate the dilution function for the pool sizes being considered. Because estimating the dilution function is almost always a requirement before one even considers implementing pool testing schemes in practical applications,<sup>13</sup> information regarding the dilution function is usually readily available from previous studies (e.g., Bateman et al. [2020], Morre et al. [2000] and Kacena et al. [1998b]). We show that, even if this information is not readily available, we can still estimate the dilution effect by comparing the distribution of a continuous biomarker used to detect infection among infected and non-infected individuals.

With information on subjects' prior probability of infection, one can also conduct simulations to evaluate the performance of ordered pooling in terms of equity. Our simulations suggest that, in general, ordered pooling yields more equitable allocations than random pooling.

Because ordered pooling is easy to implement, and does not require knowing subjects' exact probability of infection, only how those probabilities are ordered, these results may prove useful in practical applications.

## Acknowledgements

I am grateful to Dr. Shane Allwright for authorizing the usage of her dataset regarding the prevalence of HBV among Irish prisons' inmates. I would also like to thank all of those who provided invaluable feedback to this research: especial thanks goes to Luis Martins Abreu, Michael Kuhlman, Joshua M. Tebbs, Paulo Saraiva and two anonymous referees.

## References

- Alcalde, Rosana, Fernando Lucas Melo, Anna Nishiya, Suzete Cleusa Ferreira, Mario Dante Langhi Júnior, Simone Sena Fernandes, Luis Augusto Marcondes, Alberto José Silva Duarte, and Jorge Casseb.** 2009. "Distribution of hepatitis B virus genotypes and viral load levels in Brazilian chronically infected patients in São Paulo city." *Rev. Inst. Med. trop. S. Paulo*, 51(5): 269–272.
- Alizon, Samuel, and Carsten Magnus.** 2012. "Modelling the Course of an HIV Infection: Insights from Ecology and Evolution." *Viruses (1999-4915)*, 4(10): 1984 – 2013.
- Allwright, Shane, Fiona Bradley, Jean Long, Joseph Barry, Lelia Thornton, and John V Parry.** 2000. "Prevalence of antibodies to hepatitis B, hepatitis C, and HIV and risk factors in Irish prisoners: results of a national cross sectional survey." *BMJ*, 321(7253): 78–82.
- Aprahamian, Hrayr, Douglas R. Bish, and Erub K. Bish.** 2019. "Optimal Risk-Based Group Testing." *Management Science*, 65(9): 4365–4384.
- Aprahamian, Hrayr, Ebru K. Bish, and Douglas R. Bish.** 2018. "Adaptive risk-based pooling in public health screening." *IISE Transactions*, 50(9): 743–766.
- Aprahamian, Hrayr, Ebru K. Bish, and Douglas R. Bish.** 2020. "Static Risk-Based Group Testing Schemes Under Imperfectly Observable Risk." *Stochastic Systems*, 10(4): 361–390.
- Basso, Leonardo J., Marcel Goic, Marcelo Olivares, Denis Sauré, Charles Thraves, Aldo Carranza, Gabriel Y. Weintraub, Julio Covarrubia, Cristian Escobedo, Natalia Jara, Antonio Moreno, Demian Arancibia, Manuel Fuenzalida, Juan Pablo Uribe, Felipe Zúñiga, Marcela Zúñiga, Miguel ORyan, Emilio Santelices, Juan Pablo Torres, Magdalena Badal, Mirko Bozanic, Sebastián Cancino-Espinoza, Eduardo Lara, and Ignasi Neira.** 2023. "Analytics Saves Lives During the COVID-19 Crisis in Chile." *INFORMS Journal on Applied Analytics*, 53(1): 9–31.
- Basso, Leonardo, Vicente Salinas, Denis Sauré, Charles Thraves, and Natalia Yankovic.** 2022. "The Effect of Correlation and False Negatives in Pool Testing Strategies for COVID-19." *Healthcare Management Science*, 25: 146–165.

---

<sup>13</sup>If the dilution effect is too strong, pool testing may be deemed too imprecise to be considered as a good alternative to individual testing.

- Bateman, Allen C., Shanna Mueller, Kyle Guenther, and Peter Shult.** 2020. "Assessing the dilution effect of specimen pooling on the sensitivity of SARS-CoV-2 PCR tests." *Journal of Medical Virology*.
- Bertsimas, Dimitris, Vivek F. Farias, and Nikolaos Trichakis.** 2012. "On the Efficiency-Fairness Trade-off." *Management Science*, 58(12): 2234–2250.
- Birkmeyer, J.D., L.T. Goodnough, J.P. AuBuchon, P.G. Noordsij, and B. Littenberg.** 1993. "The cost-effectiveness of preoperative autologous blood donation for total hip and knee replacement." *Transfusion*, 33(7): 544–551.
- Bossert, Walter, and Kohei Kamaga.** 2020. "An axiomatization of the mixed utilitarianmaximin social welfare orderings." *Economic Theory*, 69(2): 451–473.
- Burns, Kevin C., and Carl A. Mauro.** 1987. "Group testing with test error as a function of concentration." *Communications in Statistics-theory and Methods*, 16: 2821–2837.
- Centers for Disease Control and Prevention.** 2000. "Tracking the hidden epidemics, trends in STDs in the United States." <https://www.cdc.gov/std/trends2000/trends2000.pdf>.
- Centers for Disease Control and Prevention.** 2019a. "Sexually Transmitted Infections Treatment Guidelines, 2021." <https://www.cdc.gov/std/treatment-guidelines/chlamydia.htm>.
- Centers for Disease Control and Prevention.** 2019b. "Viral Hepatitis Surveillance Report United States 2019." <https://www.cdc.gov/hepatitis/statistics/2019surveillance/pdfs/2019HepSurveillanceRpt.pdf>.
- Chahal, Harinder S, Marion G Peters, Aaron M Harris, Devon McCabe, Paul Volberding, and James G Kahn.** 2018. "Cost-effectiveness of Hepatitis B Virus Infection Screening and Treatment or Vaccination in 6 High-risk Populations in the United States." *Open Forum Infectious Diseases*, 6(1). ofy353.
- Chlebus, Bogdan.** 2001. "Randomized Communication in Radio Networks." 401–456.
- Dorfman, Robert.** 1943. "The Detection of Defective Members of Large Populations." *The Annals of Mathematical Statistics*, 14: 436–440.
- Downs, Louise O., Sabeedah Vawda, Phillip Armand Bester, Katrina A. Lythgoe, Tingyan Wang, David A. Smith Anna L. McNaughton, Tongai Maoponga, Oliver Freeman, Jim Davies Kinga A. Várnai, Kerrie Woods, Christophe Fraser, Eleanor Barnes, Dominique Goedhals, and Philippa C. Matthews.** 2020. "Bimodal distribution and set point HBV DNA viral loads in chronic infection: retrospective analysis of cohorts from the UK and South Africa."
- El-Amine, Hadi, Ebru K. Bish, and Douglas R. Bish.** 2017. "Optimal pooling strategies for nucleic acid testing of donated blood considering viral load growth curves and donor characteristics." *IISE Transactions on Healthcare Systems Engineering*, 7(1): 15–29.
- Fan, Wenxin.** 2020. "Wuhan Tests Nine Million People for Coronavirus in 10 Days." *The Wall Street Journal*.
- Goodrich, Michael, Mikhail Atallah, and Roberto Tamassia.** 2005. "Indexing Information for Data Forensics." Vol. 3531, 206–221.
- Grobe, Nadja, Alhaji Cherif, Xiaoling Wang, Zijun Dong, and Peter Kotanko.** 2020. "Sample pooling: burden or solution?" *Clinical Microbiology and Infection*.
- Harsanyi, John C.** 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy*, 63(4): 309–321.
- Hwang, F. K.** 1975. "A Generalized Binomial Group Testing Problem." *Journal of the American Statistical Association*, 70: 923–926.
- Hwang, F. K.** 1976. "Group Testing with a Dilution Effect." *Biometrika*, 63(3): 671–673.

- Jackson, B.R., M.P. Busch, S.L. Stramer, and J.P. AuBuchon.** 2003. "The cost-effectiveness of NAT for HIV, HCV, and HBV in whole-blood donations." *Transfusion*, 43(6): 721–729.
- Kacena, Katherine A., Sean B. Quinn, René Howell, Guillermo E. Madico, Thomas C. Quin, and Charlotte A. Gaydos.** 1998a. "Pooling Urine Samples for Ligase Chain Reaction Screening for Genital Chlamydia trachomatis Infection in Asymptomatic Women." *Journal of Clinical Microbiology*, 36(2): 481–485.
- Kacena, Katherine A., Sean B. Quinn, Suzanne C. Hartman, Thomas C. Quinn, and Charlotte A. Gaydos.** 1998b. "Pooling of Urine Samples for Screening for Neisseria gonorrhoeae by Ligase Chain Reaction: Accuracy and Application." *Journal of Clinical Microbiology*, 36: 3624–3628.
- Kim, Hae-Young, Michael G. Hudgens, Jonathan M. Dreyfuss, Daniel J. Westreich, and Christopher D. Pilcher.** 2007. "Comparison of Group Testing Algorithms for Case Identification in the Presence of Test Error." *Biometrics*, 63(4): 1152–1163.
- Lan, Tian, David Kao, Mung Chiang, and Ashutosh Sabharwal.** 2010. "An Axiomatic Theory of Fairness in Network Resource Allocation." 1–9.
- Maartens, Gary, Connie Celum, and Sharon R Lewin.** 2014. "HIV infection: epidemiology, pathogenesis, treatment, and prevention." *The Lancet*, 71.
- McMahan, Christopher S., Joshua M. Tebbs, and Christopher R. Bilder.** 2012. "Informative Dorfman Screening." *Biometrics*, 68(1): 287–296.
- Mokalled, Stefani C., Christopher S. McMahan, Joshua M. Tebbs, Derek Andrew Brown, and Christopher R. Bilder.** 2021. "Incorporating the dilution effect in group testing regression." *Statistics in Medicine*, 40(11): 2540–2555.
- Morre, Servaas A., Chris J. L. M. Meijer, Christian Munk, Susanne Kruger-Kjaer, Jeanette F. Winther, Hans O. Jørgensens, and Adriaan J. C. Van Den Brule.** 2000. "Pooling of Urine Specimens for Detection of Asymptomatic Chlamydia trachomatis Infections by PCR in a Low-Prevalence Population: Cost-Saving Strategy for Epidemiological Studies and Screening Programs." *Journal of Clinical Microbiology*, 38: 1679–1680.
- Nemhauser, George, and Laurence Wolsey.** 1998. *Integer and Combinatorial Optimization*. Wiley.
- Nguyen, Ngoc T., Hrayr Aprahamian, Ebru K. Bish, and Douglas R. Bish.** 2019. "A Methodology for Deriving the Sensitivity of Pooled Testing, Based on Viral Load Progression and Pooling Dilution." *Journal of Translational Medicine*, 17(1): 1152–1163.
- Owusu-Edusei Jr, Kwame, Harrell W. Chesson, Thomas L. Gift, Robert C. Brunham, and Gail Bolan.** 2015. "Cost-effectiveness of Chlamydia Vaccination Programs for Young Women." *Emerging infectious diseases*, 21(6): 960–968.
- Rawls, John.** 1971. *A theory of justice*. The Belknap Press of Harvard University Press.
- Schneider, Mark, and Byung-Cheol Kim.** 2020. "The utilitarian maximin social welfare function and anomalies in social choice." *Southern Economic Journal*, 87(2): 629–646.
- Smith, Jacob M., A. Ziggy Uvin, Alexandria Macmadu, and Josiah D. Rich.** 2017. "Epidemiology and Treatment of Hepatitis B in Prisoners." *Current Hepatology Reports*, 16(1).
- Sobel, M., and P.A. Groll.** 1959. "Group testing to eliminate efficiently all defectives in a binomial sample." *The Bell System Technical Journal*, 38(5): 1179–1252.
- Wang, Dewei, Christopher S. McMahan, and Colin M. Gallagher.** 2015. "A general regression framework for group testing data, which incorporates pool dilution effects." *Statistics in Medicine*, 34: 3606–3621.
- Warasi, Md S, Christopher McMahan, J. Tebbs, and Christopher R. Bilder.** 2017. "Group Testing Regression Models with Dilution Submodels." *Statistics in Medicine*, 36 30: 4860–4872.

- Wein, Lawrence M., and Stefanos A. Zenio.** 1996. "Pooled testing for HIV screening: capturing the dilution effect." *Operations Research*, 44(4): 543–569.
- World Health Organization.** 2017. "Global Hepatitis Report 2017." <https://www.who.int/publications/i/item/global-hepatitis-report-2017>.
- Yapali, Suna, Nizar Talaat, and Anna S. Lok.** 2014. "Management of Hepatitis B: Our Practice and How It Relates to the Guidelines." *Clinical Gastroenterology and Hepatology*, 12: 16–26.
- Yelin, Idan, Noga Aharony, Einat Shaer Tamar, Amir Argoetti, Esther Messer, Dina Berenbaum, Einat Shafran, Areen Kuzli, Nagham Gandali, Omer Shkedi, Tamar Hashimshony, Yael Mandel-Gutfreund, Michael Halberthal, Yuval Geffen, Moran Szwarcwort-Cohen, and Roy Kishony.** 2020. "Evaluation of COVID-19 RT-qPCR Test in Multi sample Pools." *Clinical Infectious Diseases*, 71(16): 2073–2078.
- Zou, Shimian, Susan L. Stramer, and Roger Y. Dodd.** 2012. "Donor Testing and Risk: Current Prevalence, Incidence, and Residual Risk of Transfusion-Transmissible Agents in US Allogeneic Donations." *Transfusion Medicine Reviews*, 26(2): 119–128.

# Online Appendix to Pool Testing with Dilution Effects and Heterogeneous Priors

Gustavo Quinderé Saraiva\*

June 23, 2023

## Contents

<b>I When ordered pooling minimizes the expected number of tests</b>	<b>1</b>
I.1 Proof of theorem 1	5
<b>II When ordered pooling minimizes the expected number of false negatives</b>	<b>6</b>
II.1 Proof of theorem 2	10
<b>III When ordered pooling minimizes the expected number of false positives</b>	<b>11</b>
III.1 Proof of theorem 3	16
<b>IV Expected number of tests, Expected number of false negatives and Expected number of false positives for extreme dilution functions</b>	<b>16</b>
<b>V When ordered pooling maximizes social welfare</b>	<b>19</b>
V.1 Proof of proposition 2	19
V.2 Proof of proposition 3	22
V.3 Proof of theorem 5	23
<b>VI Sufficient Conditions for the Dilution Function to be Discrete-Concave</b>	<b>23</b>
VI.1 Proof of proposition 4	23
<b>VII Proof of proposition 1</b>	<b>25</b>
<b>VIII Calibrating the dilution function for Chlamydia</b>	<b>25</b>
<b>IX HBV infection: calibrating the cost of a false negative</b>	<b>26</b>
<b>X Hepatitis B: performance of ordered pooling using parametric estimates of the dilution effect</b>	<b>27</b>
<b>XI Hepatitis B case study using a smaller prevalence rate</b>	<b>27</b>

## I When ordered pooling minimizes the expected number of tests

For any arbitrary group  $G_g \subseteq S$ , we define

$$T_{G_g} \equiv \begin{cases} 1, & \text{if } |G_g| = 1 \\ 1 + |G_g| \sum_{l=0}^{|G_g|} h(l, k) P_{G_g}(l), & \text{if } |G_g| > 1 \end{cases} ,$$

---

\*Business School, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Macul, Santiago, Chile. E-mail address: gsaraiva@uc.cl

which corresponds to the expected number of tests for group  $G_g$ .<sup>1</sup>

**Lemma I.1** *If  $h(\cdot, k)$  is discrete-concave, then for any arbitrary group  $G_g \subseteq S$  such that  $|G_g| = k$ , and any  $l \in G_g$ ,*

$$\sum_{I=0}^{k-1} P_{G_g \setminus \{l\}}(I)[h(I+1, k) - h(I, k)]$$

*is decreasing in the probability of infection from each subject in  $G_g \setminus \{l\}$ .*

**Proof:** Notice that

$$\sum_{I=0}^{k-1} P_{G_g \setminus \{l\}}(I)[h(I+1, k) - h(I, k)]$$

corresponds to a weighted average of  $h(I+1, k) - h(I, k)$ , where the weights are determined by the probability mass function  $P_{G_g \setminus \{l\}}(\cdot)$ . Clearly, increasing the probability of infection from a subject in  $G_g \setminus \{l\}$  causes this average to put more weight on higher values of  $I$  (formally, letting  $Y$  be the random variable associated with the probability mass function  $P_{G_g \setminus \{l\}}(\cdot)$  and  $Y'$  be its transformed version after the probability of infection from a subject in  $G_g \setminus \{l\}$  is increased, we have that  $Y'$  first-order stochastically dominates  $Y$ ). Therefore, because discrete-concavity of  $h(\cdot, k)$  implies that  $h(I+1, k) - h(I, k)$  is decreasing in  $I$ , we have that increasing the probability of infection from a subject in  $G_g \setminus \{l\}$  causes  $\sum_{I=0}^{k-1} P_{G_g \setminus \{l\}}(I)[h(I+1, k) - h(I, k)]$  to decrease. ■

**Lemma I.2** *Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$  such that  $|G_1| = k_1 \geq 2$  and  $|G_2| = k_2 \geq 2$ . Then consider the following ordered partitions of  $G_1 \cup G_2$ :*

$$\{G_1^*, G_2^*\},$$

*and*

$$\{G_2^{**}, G_1^{**}\},$$

*where  $|G_1^*| = |G_1^{**}| = k_1$ ,  $|G_2^*| = |G_2^{**}| = k_2$ ,  $i < j$  for all  $i \in G_1^*$  and all  $j \in G_2^*$  and  $i > j$  for all  $i \in G_1^{**}$  and all  $j \in G_2^{**}$ . If  $h(\cdot, k_1)$  and  $h(\cdot, k_2)$  are both discrete-concave, then*

$$\min\{T_{G_1^*} + T_{G_2^*}, T_{G_1^{**}} + T_{G_2^{**}}\} \leq T_{G_1} + T_{G_2}.$$

**Proof:** If  $\{G_1, G_2\} = \{G_1^*, G_2^*\}$  or  $\{G_1, G_2\} = \{G_1^{**}, G_2^{**}\}$ , the proof is trivial. So suppose that  $\{G_1, G_2\} \neq \{G_1^*, G_2^*\}$  and  $\{G_1, G_2\} \neq \{G_1^{**}, G_2^{**}\}$ , so that

$$\min_{i \in G_1} i < \max_{j \in G_2} j$$

and

$$\min_{j \in G_2} j < \max_{i \in G_1} i.$$

For an arbitrary  $i \in G_1$  and  $j \in G_2$  either one of the following inequalities must hold

$$\sum_{I=0}^{k_1-1} P_{G_1 \setminus \{i\}}(I)[h(I+1, k_1) - h(I, k_1)] \leq \sum_{I=0}^{k_2-1} P_{G_2 \setminus \{j\}}(I)[h(I+1, k_2) - h(I, k_2)] \quad (1)$$

or

$$\sum_{I=0}^{k_2-1} P_{G_2 \setminus \{j\}}(I)[h(I+1, k_2) - h(I, k_2)] \leq \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{i\}}(I)[h(I+1, k_1) - h(I, k_1)]. \quad (2)$$

1. Suppose that inequality 2 holds. Then define

$$\bar{i} \equiv \max_{i \in G_1} i$$

and

$$\underline{j} \equiv \min_{j \in G_2} j.$$

---

<sup>1</sup>The constant 1 corresponds to the pooled test performed for the group, while the second term corresponds to the probability that the pooled test detects infection times the number of subjects in the group (who are each tested individually in the event the pooled test detects infection). If the group is comprised of a single subject, then only one test is performed for that subject.

Because  $q_{\bar{i}} \geq q_i$ , lemma I.1 implies that

$$\sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{i}\}}(I)[h(I+1, k_1) - h(I, k_1)] \leq \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{i}\}}(I)[h(I+1, k_1) - h(I, k_1)]. \quad (3)$$

Analogously,  $q_{\bar{j}} \leq q_j$  and lemma I.1 imply that

$$\sum_{I=0}^{k_2-1} P_{G_2 \setminus \{\bar{j}\}}(I)[h(I+1, k_2) - h(I, k_2)] \leq \sum_{I=0}^{k_2-1} P_{G_2 \setminus \{\bar{j}\}}(I)[h(I+1, k_2) - h(I, k_2)]. \quad (4)$$

Together inequalities 2, 3 and 4 imply that

$$\sum_{I=0}^{k_2-1} P_{G_2 \setminus \{\bar{j}\}}(I)[h(I+1, k_2) - h(I, k_2)] \leq \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{i}\}}(I)[h(I+1, k_1) - h(I, k_1)]. \quad (5)$$

Now let us exchange the position of subjects  $\underline{i}$  and  $\bar{j}$  to create the new groups

$$\tilde{G}_1 = G_1 \cup \{\bar{j}\} \setminus \{\bar{i}\}$$

and

$$\tilde{G}_2 = G_2 \cup \{\bar{i}\} \setminus \{\bar{j}\}.$$

We will show that  $T_{\tilde{G}_1} + T_{\tilde{G}_2} \leq T_{G_1} + T_{G_2}$ . First, notice that

$$\begin{aligned} T_{\tilde{G}_1} + T_{\tilde{G}_2} &\leq T_{G_1} + T_{G_2} \\ \Leftrightarrow \sum_{I=0}^{k_1} h(I, k_1) P_{\tilde{G}_1}(I) + \sum_{I=0}^{k_2} h(I, k_2) P_{\tilde{G}_2}(I) &\leq \sum_{I=0}^{k_1} h(I, k_1) P_{G_1}(I) + \sum_{I=0}^{k_2} h(I, k_2) P_{G_2}(I) \end{aligned} \quad (6)$$

Now notice that

$$\begin{aligned} \sum_{I=0}^{k_1} h(I, k_1) P_{\tilde{G}_1}(I) &= q_{\bar{j}} \sum_{I=0}^{k_1-1} h(I+1, k_1) P_{\tilde{G}_1 \setminus \{\bar{j}\}}(I) + (1 - q_{\bar{j}}) \sum_{I=0}^{k_1-1} h(I, k_1) P_{\tilde{G}_1 \setminus \{\bar{j}\}}(I) \\ &= q_{\bar{j}} \sum_{I=0}^{k_1-1} P_{\tilde{G}_1 \setminus \{\bar{j}\}}(I)[h(I+1, k_1) - h(I, k_1)] + \sum_{I=0}^{k_1-1} h(I, k_1) P_{\tilde{G}_1 \setminus \{\bar{j}\}}(I). \end{aligned}$$

Because  $\tilde{G}_1 \setminus \{\bar{j}\} = G_1 \setminus \{\bar{i}\}$ , the above expression can be rewritten as

$$\sum_{I=0}^{k_1} h(I, k_1) P_{\tilde{G}_1}(I) = q_{\bar{j}} \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{i}\}}(I)[h(I+1, k_1) - h(I, k_1)] + \sum_{I=0}^{k_1-1} h(I, k_1) P_{G_1 \setminus \{\bar{i}\}}(I, k_1).$$

Analogously,

$$\begin{aligned} \sum_{I=0}^{k_2} h(I, k_2) P_{\tilde{G}_2}(I) &= q_{\bar{i}} \sum_{I=0}^{k_2-1} P_{G_2 \setminus \{\bar{i}\}}(I)[h(I+1, k_2) - h(I, k_2)] + \sum_{I=0}^{k_1-1} h(I, k_2) P_{G_2 \setminus \{\bar{i}\}}(I), \\ \sum_{I=0}^{k_1} h(I, k_1) P_{G_1}(I) &= q_{\bar{i}} \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{i}\}}(I)[h(I+1, k_1) - h(I, k_1)] + \sum_{I=0}^{k_1-1} h(I, k_1) P_{G_1 \setminus \{\bar{i}\}}(I), \\ \sum_{I=0}^{k_2} h(I, k_2) P_{G_2}(I) &= q_{\bar{j}} \sum_{I=0}^{k_2-1} P_{G_2 \setminus \{\bar{j}\}}(I)[h(I+1, k_2) - h(I, k_2)] + \sum_{I=0}^{k_2-1} h(I, k_2) P_{G_2 \setminus \{\bar{j}\}}(I). \end{aligned}$$

Replacing these expressions into 6 and rearranging the terms, we get

$$\begin{aligned}
T_{\tilde{G}_1} + T_{\tilde{G}_2} &\leq T_{G_1} + T_{G_2} \tag{7} \\
\iff (q_{\tilde{i}} - q_{\tilde{j}}) \sum_{I=0}^{k_2-1} P_{G_2 \setminus \{\tilde{j}\}}(I)[h(I+1, k_2) - h(I)] &\leq (q_{\tilde{i}} - q_{\tilde{j}}) \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\tilde{i}\}}(I)[h(I+1, k_1) - h(I, k_1)] \\
\iff \sum_{I=0}^{k_2-1} P_{G_2 \setminus \{\tilde{j}\}}(I)[h(I+1, k_2) - h(I, k_2)] &\leq \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\tilde{i}\}}(I)[h(I+1, k_1) - h(I, k_1)], \tag{8}
\end{aligned}$$

which, from inequality 5, holds.

Now, defining

$$\tilde{i}' \equiv \max_{i \in \tilde{G}_1} i$$

and

$$\tilde{j}' \equiv \min_{j \in \tilde{G}_2} j,$$

we have that either one of the following conditions must hold:

- (a)  $q_{\tilde{j}} \geq q_{\tilde{i}'}$  (i.e.,  $j > i$  for all  $i \in \tilde{G}_1$  and all  $j \in \tilde{G}_2$ ), in which case  $\tilde{G}_1 = G_1^*$  and  $\tilde{G}_2 = G_2^*$ , so that  $T_{G_1^*} + T_{G_2^*} \leq T_{\tilde{G}_1} + T_{\tilde{G}_2}$ .
- (b)  $q_{\tilde{j}} < q_{\tilde{i}'}$ . In this case, notice that when we moved from  $G_1$  to  $\tilde{G}_1$  we decreased the probability of infection from one subject in this group while not altering the probability of infection from the remaining subjects within the group. From lemma I.1 this implies that

$$\sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\tilde{i}\}}(I)[h(I+1, k_1) - h(I, k_1)] \leq \sum_{I=0}^{k_1-1} P_{\tilde{G}_1 \setminus \{\tilde{i}'\}}(I)[h(I+1, k_1) - h(I, k_1)]. \tag{9}$$

Analogously, when moving from  $G_2$  to  $\tilde{G}_2$  we increased the probability of infection from one subject in this group while not altering the probability of infection from the remaining subjects in this group. From lemma I.1 this implies that

$$\sum_{I=0}^{k_1-1} P_{\tilde{G}_2 \setminus \{\tilde{j}'\}}(I)[h(I+1, k_1) - h(I, k_1)] \leq \sum_{I=0}^{k_1-1} P_{G_2 \setminus \{\tilde{j}\}}(I)[h(I+1, k_1) - h(I, k_1)]. \tag{10}$$

Together, inequalities 8, 9 and 10 imply that

$$\sum_{I=0}^{k-1} P_{\tilde{G}_2 \setminus \{\tilde{j}'\}}(I)[h(I+1) - h(I)] \leq \sum_{I=0}^{k-1} P_{\tilde{G}_1 \setminus \{\tilde{i}'\}}(I)[h(I+1) - h(I)].$$

So we can redefine  $G_1 = \tilde{G}_1$ ,  $G_2 = \tilde{G}_2$ ,  $\tilde{i} = \tilde{i}'$  and  $\tilde{j} = \tilde{j}'$  and repeat the previous steps iteratively, until the final pair of groups is given by  $G_1^*$  and  $G_2^*$  such that  $|G_1^*| = |G_1| = k_1$ ,  $|G_2^*| = |G_2| = k_2$ ,  $q_i \leq q_j$  for all  $i \in G_1^*$  and all  $j \in G_2^*$ .

Because at each step of this algorithm we reduce the expected number of tests for subjects within these groups, and because at the end of this process we obtain the groups  $G_1^*$  and  $G_2^*$ , we have that

$$T_{G_1^*} + T_{G_2^*} \leq T_{G_1} + T_{G_2},$$

as we wanted to show.

2. If inequality 1 holds, the proof is analogous to case 1, when inequality 2 holds, instead. Indeed, if 1 holds, we define  $\tilde{i} = \min_{i \in G_1} i$  and  $\tilde{j} = \max_{j \in G_m} j$ , and then repeat the steps in case 1 to show that switching subject  $\tilde{i}$  with subject  $\tilde{j}$  weakly reduces the expected number of tests, provided that  $h(\cdot, k_1)$  and  $h(\cdot, k_2)$  are both discrete-concave. Then, we iteratively switch subjects from the new groups in the same fashion until we reach the ordered partition  $\{G_2^{**}, G_1^{**}\}$ . Because at each step of the algorithm the expected number of tests diminishes, we obtain  $T_{G_1^{**}} + T_{G_2^{**}} \leq T_{G_1} + T_{G_2}$ .

■

**Lemma I.3** Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$  such that  $|G_1| = k_1 \geq 1$  and  $|G_2| = 1$ . Let  $\bar{j} = \max(G_1 \cup G_2)$  (i.e.,  $\bar{j}$  is the subject with highest probability of infection in group  $G_1 \cup G_2$ ). Then consider the following ordered partition of  $G_1 \cup G_2$ :

$$\{G_1^*, G_2^*\},$$

where

$$G_1^* \equiv G_1 \cup G_2 \setminus \{\bar{j}\},$$

$$G_2^* \equiv \{\bar{j}\},$$

i.e.,  $\{G_1^*, G_2^*\}$  is the ordered partition that groups the  $k_1$  subjects with lowest probability of infection together, and the subject with highest probability of infection alone.

If  $h(\cdot, k_1)$  is increasing, then

$$T_{G_1^*} + T_{G_2^*} \leq T_{G_1} + T_{G_2}.$$

**Proof:** Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$ . The proof when  $|G_1| = |G_2| = 1$  is trivial. So suppose that  $|G_1| = k_1 \geq 2$  and  $G_2 = \{j\}$ , with  $j < \bar{j} = \max(G_1 \cup G_2)$ . Let

$$G_1^* \equiv G_1 \cup G_2 \setminus \{\bar{j}\},$$

$$G_2^* \equiv \{\bar{j}\}.$$

Because  $|G_2| = |G_2^*| = 1$ , we have that  $T_{G_2} = T_{G_2^*} = 1$ . Therefore,

$$\begin{aligned} & T_{G_1^*} + T_{G_2^*} \leq T_{G_1} + T_{G_2} \\ \Leftrightarrow & \left(1 + k_1 \sum_{I=0}^{k_1} P_{G_1^*}(I) h(I, k_1)\right) + 1 \leq \left(1 + k_1 \sum_{I=0}^{k_1} P_{G_1}(I) h(I, k_1)\right) + 1 \\ \Leftrightarrow & k_1 \left( q_j \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{j}\}}(I) h(I+1, k_1) + (1 - q_j) \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{j}\}}(I) h(I, k_1) \right) \leq \\ & k_1 \left( q_{\bar{j}} \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{j}\}}(I) h(I+1, k_1) + (1 - q_{\bar{j}}) \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{j}\}}(I) h(I, k_1) \right) \\ \Leftrightarrow & 0 \leq (q_{\bar{j}} - q_j) \left( \sum_{I=0}^{k_1-1} P_{G_1 \setminus \{\bar{j}\}}(I) (h(I+1, k_1) - h(I, k_1)) \right) \\ \Leftrightarrow & 0 \leq q_{\bar{j}} - q_j, \end{aligned}$$

which is satisfied, since  $q_{\bar{j}} \geq q_j$  for all  $j < \bar{j}$ . ■

## I.1 Proof of theorem 1

Let  $\{G_1, G_2, \dots, G_m\}$  be an arbitrary partition of  $S = \{1, 2, \dots, n\}$ , where  $m$  is the number of groups from the partition (e.g., if all of the groups from the partition have the same size  $k$ , then  $m = \frac{n}{k}$ ). Also suppose that  $h(\cdot, |G_g|)$  is discrete-concave for every  $G_g \in \{G_1, G_2, \dots, G_m\}$ . Then implement the following algorithm:

### Algorithm I.1

1. Initialize the partition  $\tilde{\Omega} = \{\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_m\}$ , where  $\tilde{G}_g = G_g$  for all  $g \in \{1, 2, \dots, m\}$ .
2. Initialize  $g = 1$ .

3. Set  $w = g + 1$ .

4. Pick groups  $\tilde{G}_g$  and  $\tilde{G}_w$  from  $\tilde{\Omega}$ .

(a) If  $|\tilde{G}_g| \geq 2$  and  $|\tilde{G}_w| \geq 2$ , consider the following ordered partitions of  $\tilde{G}_g \cup \tilde{G}_w$ :

$$\{G_g^*, G_w^*\},$$

and

$$\{G_w^{**}, G_g^{**}\},$$

where  $|G_g^*| = |G_g^{**}| = |\tilde{G}_g|$ ,  $|G_w^*| = |G_w^{**}| = |\tilde{G}_w|$ ,  $i < j$  for all  $i \in G_g^*$  and all  $j \in G_w^*$ , and  $q_i > q_j$  for all  $i \in G_g^{**}$  and all  $j \in G_w^{**}$ .

If  $T_{G_g^*} + T_{G_w^*} \leq T_{G_g^{**}} + T_{G_w^{**}}$ , redefine

$$\tilde{G}_g = G_g^* \quad \text{and} \quad \tilde{G}_w = G_w^*$$

else, redefine

$$\tilde{G}_g = G_w^{**} \quad \text{and} \quad \tilde{G}_w = G_g^{**}.$$

(b) If  $|\tilde{G}_g| = 1$  or  $|\tilde{G}_w| = 1$ , redefine

$$\tilde{G}_g = \tilde{G}_g \cup \tilde{G}_w \setminus \{\max(\tilde{G}_g \cup \tilde{G}_w)\},$$

and

$$\tilde{G}_w = \{\max(\tilde{G}_g \cup \tilde{G}_w)\}.$$

5. If  $w = m$ , proceed to the next step. Else, redefine  $w = w + 1$  and repeat step 4.

6. If  $g = m - 1$ , stop the algorithm, else, redefine  $g = g + 1$  and repeat step 3.

Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be the original partition and  $\tilde{\Omega} = \{\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_m\}$  be the final partition obtained after implementing algorithm I.1. From lemmas I.2 and I.3, at each step of this algorithm the overall expected number of tests weakly diminishes, which implies that  $\mathbb{E}[T(\tilde{\Omega})] \leq \mathbb{E}[T(\Omega)]$ . Because at the end of each step 6 of algorithm I.1 we have that, for each  $w > m$ ,  $q_i \leq q_j$  for all  $i \in \tilde{G}_g$  and all  $j \in \tilde{G}_w$ , we have that  $\tilde{\Omega}$  is an ordered partition of  $S$ . Moreover, because changes made at steps 4a and 4b of algorithm I.1 preserve pool sizes, there exists a permutation  $p$  of the indices  $(1, 2, \dots, m)$  such that  $|G_{p(g)}| = |G_g|$  for all  $g \in \{1, 2, \dots, m\}$ .

As to the second part of the theorem, notice that, at each step 4b of algorithm I.1 we are allocating the subject with highest probability of infection to be tested individually. This implies that if a subject  $i$  is tested individually under  $\tilde{\Omega}$ , then a subject  $j$  with  $q_j > q_i$  is also tested individually under  $\tilde{\Omega}$ .  $\blacksquare$

## II When ordered pooling minimizes the expected number of false negatives

For any arbitrary group  $G_g \subseteq S$ , it can be shown that

$$FN_{G_g} \equiv \begin{cases} (1 - S_e)q_i, & \text{if } G_g = \{i\}, \\ \sum_{I=0}^{|G_g|} P_{G_g}(I)I[1 - h(I, |G_g|)S_e], & \text{if } |G_g| > 1 \end{cases},$$

corresponds to the expected number of false negatives from group  $G_g$ .<sup>2</sup>

<sup>2</sup>That this is true for  $|G_g| = 1$  is trivial. To see this is true for  $|G_g| > 1$ , define  $FN_{G_g, I, d}$  as the expected number of false negatives in group  $G_g$  conditional that the group has exactly  $I$  infected subjects and conditional that infection is detected in the first stage of testing. Similarly, define  $FN_{G_g, I, nd}$  as the expected number of false negatives in group  $G_g$  conditional that the group has exactly  $I$  infected subjects and conditional that infection is *not* detected in the first stage of testing. Then, by the law of iterated expectations we must have

$$FN_{G_g} = \sum_{I=0}^{|G_g|} P_{G_g}(I) \left[ h(I, |G_g|)FN_{G_g, I, d} + (1 - h(I, |G_g|))FN_{G_g, I, nd} \right].$$

Noticing that  $FN_{G_g, I, d} = I(1 - S_e)$  and  $FN_{G_g, I, nd} = I$ , yields the desired result.

For any arbitrary group  $G_g \subseteq S$  such that  $|G_g| \geq 2$  and any  $j \in G_g$  we define

$$A_{G_g, j} \equiv \sum_{I=0}^{|G_g|-1} (I+1)(1-h(I+1, |G_g|)S_e)P_{G_g \setminus j}(I),$$

$$B_{G_g, j} \equiv \sum_{I=0}^{|G_g|-1} I(1-h(I, |G_g|)S_e)P_{G_g \setminus j}(I).$$

From the above expressions,  $A_{G_g, j}$  is the expected number of false negatives in group  $G_g$  conditional that  $j \in G_g$  is infected. Similarly,  $B_{G_g, j}$  is the expected number of false negatives in group  $G_g$  conditional that  $j \in G_g$  is not infected.

Now notice that, for any  $j \in G_g$ ,

$$\begin{aligned} FN_{G_g} &= q_j A_{G_g, j} + (1-q_j) B_{G_g, j} \\ &= q_j (A_{G_g, j} - B_{G_g, j}) + B_{G_g, j}. \end{aligned} \quad (11)$$

**Lemma II.1** *If hypothesis 1 holds for a given  $k \geq 2$ , then for any arbitrary group  $G_g \subseteq S$  such that  $|G_g| = k$ , and any  $l \in G_g$ ,*

$$A_{G_g, l} - B_{G_g, l}$$

*is decreasing in the probability of infection from each subject in  $G_g \setminus \{l\}$ .*

**Proof:** Notice that

$$A_{G_g, l} - B_{G_g, l} = \sum_{I=0}^{k-1} P_{G_g \setminus \{l\}}(I) [(I+1)(1-S_e h(I+1, k)) - I(1-S_e h(I, k))],$$

which corresponds to a weighted average of  $(I+1)(1-S_e h(I+1, k)) - I(1-S_e h(I, k))$ , where the weights are determined by the probability mass function  $P_{G_g \setminus \{l\}}(\cdot)$ . Clearly, increasing the probability of infection from a patient in  $G_g \setminus \{l\}$  causes this average to put more weight on higher values of  $I$  (formally, letting  $Y$  be the random variable associated with the probability mass function  $P_{G_g \setminus \{l\}}(\cdot)$  and  $Y'$  be its transformed version after the probability of infection from a patient in  $G_g \setminus \{l\}$  is increased, we have that  $Y'$  first-order stochastically dominates  $Y$ ). So it suffices to show that  $(I+1)(1-S_e h(I+1, k)) - I(1-S_e h(I, k))$  is decreasing in  $I$ , a property that is satisfied if, for any  $I \in \{1, 2, \dots, k-1\}$ ,

$$\begin{aligned} &(I+1)(1-S_e h(I+1, k)) - I(1-S_e h(I, k)) - \\ &- [I(1-S_e h(I, k)) - (I-1)(1-S_e h(I-1, k))] \leq 0 \\ \iff &\frac{I+1}{2I} h(I+1, k) + \frac{I-1}{2I} h(I-1, k) \geq h(I, k). \end{aligned}$$

■

**Lemma II.2** *Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$  such that  $|G_1| = k_1 \geq 2$  and  $|G_2| = k_2 \geq 2$ . Then consider the following ordered partitions of  $G_1 \cup G_2$ :*

$$\{G_1^*, G_2^*\},$$

*and*

$$\{G_2^{**}, G_1^{**}\},$$

*where  $|G_1^*| = |G_1^{**}| = k_1$ ,  $|G_2^*| = |G_2^{**}| = k_2$ ,  $i < j$  for all  $i \in G_1^*$  and all  $j \in G_2^*$  and  $i > j$  for all  $i \in G_1^{**}$  and all  $j \in G_2^{**}$ . If  $h(\cdot, k_1)$  and  $h(\cdot, k_2)$  satisfy hypothesis 1, then*

$$\min\{FN_{G_1^*} + FN_{G_2^*}, FN_{G_1^{**}} + FN_{G_2^{**}}\} \leq FN_{G_1} + FN_{G_2}.$$

**Proof:** If  $\{G_1, G_2\} = \{G_1^*, G_2^*\}$  or  $\{G_1, G_2\} = \{G_1^{**}, G_2^{**}\}$ , the proof is trivial. So suppose that  $\{G_1, G_2\} \neq \{G_1^*, G_2^*\}$  and  $\{G_1, G_2\} \neq \{G_1^{**}, G_2^{**}\}$ , so that

$$\min_{i \in G_1} i < \max_{j \in G_2} j$$

and

$$\min_{j \in G_2} j < \max_{i \in G_1} i.$$

For an arbitrary  $i \in G_1$  and  $j \in G_2$  either one of the following inequalities must hold

$$A_{G_1,i} - B_{G_1,i} \leq A_{G_2,j} - B_{G_2,j} \quad (12)$$

or

$$A_{G_2,j} - B_{G_2,j} \leq A_{G_1,i} - B_{G_1,i} \quad (13)$$

1. Assume that inequality 13 holds, and define

$$\bar{i} \equiv \max_{i \in G_1} i$$

and

$$\underline{j} \equiv \min_{j \in G_2} j.$$

Because  $q_{\bar{i}} \geq q_i$ , lemma II.1 implies that

$$A_{G_1,i} - B_{G_1,i} \leq A_{G_1,\bar{i}} - B_{G_1,\bar{i}}. \quad (14)$$

Analogously,  $q_{\underline{j}} \leq q_j$  and lemma II.1 imply that

$$A_{G_2,\underline{j}} - B_{G_1,\underline{j}} \leq A_{G_1,j} - B_{G_1,j}. \quad (15)$$

Together inequalities 13, 14 and 15 imply that

$$A_{G_2,\underline{j}} - B_{G_2,\underline{j}} \leq A_{G_1,\bar{i}} - B_{G_1,\bar{i}}. \quad (16)$$

Now let us exchange the position of subjects  $\underline{j}$  and  $\bar{i}$  to create the new groups

$$\tilde{G}_1 = G_1 \cup \{\underline{j}\} \setminus \{\bar{i}\}$$

and

$$\tilde{G}_2 = G_2 \cup \{\bar{i}\} \setminus \{\underline{j}\}.$$

Then we must have  $FN_{\tilde{G}_1} + FN_{\tilde{G}_2} \leq FN_{G_1} + FN_{G_2}$ . Indeed, from expression 11, we have that

$$\begin{aligned} FN_{G_1} &= q_{\bar{i}} A_{G_1,\bar{i}} + (1 - q_{\bar{i}}) B_{G_1,\bar{i}}, \\ FN_{G_2} &= q_{\underline{j}} A_{G_2,\underline{j}} + (1 - q_{\underline{j}}) B_{G_2,\underline{j}}. \end{aligned}$$

Moreover, because

$$\begin{aligned} A_{\tilde{G}_1,\underline{j}} &= A_{G_1,\bar{i}}, \\ B_{\tilde{G}_1,\underline{j}} &= B_{G_1,\bar{i}}, \\ A_{\tilde{G}_2,\bar{i}} &= A_{G_2,\underline{j}}, \\ B_{\tilde{G}_2,\bar{i}} &= B_{G_2,\underline{j}}, \end{aligned}$$

we also have that

$$\begin{aligned} FN_{\tilde{G}_1} &= q_{\underline{j}} A_{G_1,\bar{i}} + (1 - q_{\underline{j}}) B_{G_1,\bar{i}}, \\ FN_{\tilde{G}_2} &= q_{\bar{i}} A_{G_2,\underline{j}} + (1 - q_{\bar{i}}) B_{G_2,\underline{j}}. \end{aligned}$$

Therefore,

$$\begin{aligned}
& FN_{\tilde{G}_1} + FN_{\tilde{G}_2} \leq FN_{G_1} + FN_{G_2} \\
& \iff q_{\underline{j}}(A_{G_1, \bar{i}} - B_{G_1, \bar{i}}) + q_{\bar{i}}(A_{G_2, \underline{j}} - B_{G_2, \underline{j}}) \leq q_{\bar{i}}(A_{G_1, \bar{i}} - B_{G_1, \bar{i}}) + q_{\underline{j}}(A_{G_2, \underline{j}} - B_{G_2, \underline{j}}) \\
& \iff (q_{\bar{i}} - q_{\underline{j}})(A_{G_2, \underline{j}} - B_{G_2, \underline{j}}) \leq (q_{\bar{i}} - q_{\underline{j}})(A_{G_1, \bar{i}} - B_{G_1, \bar{i}}) \\
& \iff (A_{G_2, \underline{j}} - B_{G_2, \underline{j}}) \leq (A_{G_1, \bar{i}} - B_{G_1, \bar{i}}),
\end{aligned}$$

which, from inequality 16, holds.

Now, defining

$$\bar{i}' \equiv \max_{i \in \tilde{G}_1} i$$

and

$$\underline{j}' \equiv \min_{j \in \tilde{G}_2} j,$$

we have that either one of the following conditions must hold:

- (a)  $q_{\underline{j}} \geq q_{\bar{i}'}$ , in which case  $\tilde{G}_1 = G_1^*$  and  $\tilde{G}_2 = G_2^*$ , so that  $FN_{G_1^*} + FN_{G_2^*} \leq FN_{G_1} + FN_{G_2}$ .
- (b)  $q_{\underline{j}} < q_{\bar{i}'}$ . In this case, notice that when we moved from  $G_1$  to  $\tilde{G}_1$  we decreased the probability of infection from one subject in this group while not altering the probability of infection from the remaining subjects in the group. From lemma II.1 this implies that

$$A_{G_1, \bar{i}} - B_{G_1, \bar{i}} \leq A_{\tilde{G}_1, \bar{i}'} - B_{\tilde{G}_1, \bar{i}'}. \quad (17)$$

Analogously, when moving from  $G_2$  to  $\tilde{G}_2$  we increased the probability of infection from one subject in this group while not altering the probability of infection from the remaining subjects in the group. From lemma II.1 this implies that

$$A_{\tilde{G}_2, \underline{j}'} - B_{\tilde{G}_2, \underline{j}'} \leq A_{G_2, \underline{j}} - B_{G_2, \underline{j}}. \quad (18)$$

Together inequalities 16, 17 and 18 imply that

$$A_{\tilde{G}_2, \underline{j}'} - B_{\tilde{G}_2, \underline{j}'} \leq A_{\tilde{G}_1, \bar{i}'} - B_{\tilde{G}_1, \bar{i}'}$$

So we can redefine  $G_1 = \tilde{G}_1$ ,  $G_2 = \tilde{G}_2$ ,  $\bar{i} = \bar{i}'$  and  $\underline{j} = \underline{j}'$  and repeat the previous steps iteratively, until the final pair of groups is given by  $G_1^*$  and  $G_2^*$ .

Because at each step of this algorithm we reduce the expected number of false negatives for subjects in these groups, and because at the end of this process we obtain the groups  $G_1^*$  and  $G_2^*$ , we have that

$$FN_{G_1^*} + FN_{G_2^*} \leq FN_{G_1} + FN_{G_2},$$

as we wanted to show.

2. If inequality 12 holds, the proof is analogous to case in which inequality 13 holds (case 1). Indeed, when 12 holds, we define  $\underline{i} = \min_{i \in G_1} i$  and  $\bar{j} = \max_{j \in G_m} j$ , and then repeat the same steps in case 13 to show that switching the positions of subjects  $\underline{i}$  and  $\bar{j}$  weakly reduces the expected number of tests, provided that  $h(\cdot, k_1)$  and  $h(\cdot, k_2)$  both satisfy hypothesis 1. Then, we iteratively switch subjects from the new groups in the same fashion until we reach the ordered partition  $\{G_2^{**}, G_1^{**}\}$ . Because at each step of the algorithm the expected number of false negatives diminishes, we obtain  $FN_{G_1^{**}} + FN_{G_2^{**}} \leq FN_{G_1} + FN_{G_2}$ . ■

**Lemma II.3** Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$  such that  $|G_1| = k \geq 1$  and  $|G_2| = 1$ . Let  $\bar{j} \equiv \max(G_1 \cup G_2)$  and  $\underline{j} \equiv \min(G_1 \cup G_2)$ . Then consider the following ordered partitions of  $G_1 \cup G_2$ :

$$\{G_1^*, G_2^*\},$$

and

$$\{G_2^{**}, G_1^{**}\},$$

where

$$\begin{aligned} G_1^* &\equiv G_1 \cup G_2 \setminus \{\bar{j}\} \\ G_2^* &\equiv \{\bar{j}\} \\ G_1^{**} &\equiv \{\underline{j}\} \\ G_2^{**} &\equiv G_1 \cup G_2 \setminus \{\underline{j}\}, \end{aligned}$$

i.e.,  $\{G_1^*, G_2^*\}$  is an ordered partition of  $G_1 \cup G_2$  that individually tests the subject with highest probability of infection, while  $\{G_1^{**}, G_2^{**}\}$  is an ordered partition of  $G_1 \cup G_2$  that individually tests the subject with lowest probability of infection.

If  $h(\cdot, k)$  satisfies hypothesis 1, then

$$\min\{FN_{G_1^*} + FN_{G_2^*}, FN_{G_1^{**}} + FN_{G_2^{**}}\} \leq FN_{G_1} + FN_{G_2}.$$

**Proof:** If  $k = 1$ , the proof is trivial, as all partitions are the same. So we will assume that  $k > 1$ . As the proof is also trivial for the cases in which  $G_2 = \{\bar{j}\}$  or  $G_2 = \{\underline{j}\}$ , let us assume that  $G_2 = \{i\}$ , with  $\underline{j} < i < \bar{j}$ . Then we have that

$$FN_{G_1} + FN_{G_2} = q_{\bar{j}}(A_{G_1, \bar{j}} - B_{G_1, \bar{j}}) + B_{G_1, \bar{j}} + q_i(1 - S_e)$$

and

$$FN_{G_1^*} + FN_{G_2^*} = q_i(A_{G_1, \bar{j}} - B_{G_1, \bar{j}}) + B_{G_1, \bar{j}} + q_{\bar{j}}(1 - S_e).$$

Now notice that one of the following inequalities must hold:

$$FN_{G_1^*} + FN_{G_2^*} \leq FN_{G_1} + FN_{G_2} \tag{19}$$

or

$$FN_{G_1^*} + FN_{G_2^*} > FN_{G_1} + FN_{G_2} \tag{20}$$

If inequality (19) holds, the Lemma is proven. So suppose that inequality (20) holds. Then we must have

$$\begin{aligned} q_{\bar{j}}(A_{G_1, \bar{j}} - B_{G_1, \bar{j}}) + B_{G_1, \bar{j}} + q_i(1 - S_e) &< q_i(A_{G_1, \bar{j}} - B_{G_1, \bar{j}}) + B_{G_1, \bar{j}} + q_{\bar{j}}(1 - S_e) \\ \iff (A_{G_1, \bar{j}} - B_{G_1, \bar{j}}) &< (1 - S_e). \end{aligned} \tag{21}$$

We want to show that

$$\begin{aligned} FN_{G_1^{**}} + FN_{G_2^{**}} &\leq FN_{G_1} + FN_{G_2} \\ \iff q_i(A_{G_1, \underline{j}} - B_{G_1, \underline{j}}) + B_{G_1, \underline{j}} + q_{\underline{j}}(1 - S_e) &\leq q_{\underline{j}}(A_{G_1, \underline{j}} - B_{G_1, \underline{j}}) + B_{G_1, \underline{j}} + q_i(1 - S_e) \\ \iff (A_{G_1, \underline{j}} - B_{G_1, \underline{j}}) &\leq (1 - S_e). \end{aligned} \tag{22}$$

From lemma II.1, we have that  $(A_{G_1, \underline{j}} - B_{G_1, \underline{j}}) \leq (A_{G_1, \bar{j}} - B_{G_1, \bar{j}})$ . Therefore, from inequality 21, we have that  $(A_{G_1, \underline{j}} - B_{G_1, \underline{j}}) \leq (1 - S_e)$ , as we wanted to show.  $\blacksquare$

## II.1 Proof of theorem 2

Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be an arbitrary partition of  $S = \{1, 2, \dots, n\}$ , where  $m$  is the number of groups from the partition (e.g., if all of the groups from the partition have the same size  $k$ , then  $m = \frac{n}{k}$ ). Also suppose that hypothesis 1 holds for every  $k \in \{|G_1|, |G_2|, \dots, |G_m|\}$ . Then implement the following algorithm:

### Algorithm II.1

1. Initialize the partition  $\tilde{\Omega} = \{\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_m\}$ , where  $\tilde{G}_g = G_g$  for all  $g \in \{1, 2, \dots, m\}$ .
2. Initialize  $g = 1$ .
3. Set  $w = g + 1$ .
4. Pick groups  $\tilde{G}_g$  and  $\tilde{G}_w$  from  $\tilde{\Omega}$ .
5. Then consider the following ordered partitions of  $\tilde{G}_g \cup \tilde{G}_w$ :

$$\{G_g^*, G_w^*\},$$

and

$$\{G_w^{**}, G_g^{**}\},$$

where  $|G_g^*| = |G_w^{**}| = |\tilde{G}_g|$ ,  $|G_w^*| = |G_g^{**}| = |\tilde{G}_w|$ ,  $i < j$  for all  $i \in G_g^*$  and all  $j \in G_w^*$ , and  $q_i > q_j$  for all  $i \in G_g^{**}$  and all  $j \in G_w^{**}$ .

If  $T_{G_g^*} + T_{G_w^*} \leq T_{G_g^{**}} + T_{G_w^{**}}$ , redefine

$$\tilde{G}_g = G_g^* \quad \text{and} \quad \tilde{G}_w = G_w^*$$

else, redefine

$$\tilde{G}_g = G_w^{**} \quad \text{and} \quad \tilde{G}_w = G_g^{**}.$$

6. If  $w = m$ , proceed to the next step. Else, redefine  $w = w + 1$  and repeat step 4.
7. If  $g = m - 1$ , stop the algorithm, else, redefine  $g = g + 1$  and repeat step 3.

Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be the original partition and  $\tilde{\Omega} = \{\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_m\}$  be the final partition obtained after implementing algorithm II.1. From lemmas II.2 and II.3, at each step of this algorithm the overall expected number of false negatives weakly diminishes, which implies that  $\mathbb{E}[FN(\tilde{\Omega})] \leq \mathbb{E}[FN(\Omega)]$ . Because at the end of each step 7 of algorithm II.1 we have that, for each  $w > m$ ,  $q_i \leq q_j$  for all  $i \in \tilde{G}_g$  and all  $j \in \tilde{G}_w$ , we have that  $\tilde{\Omega}$  is an ordered partition of  $S$ . Moreover, because changes made at each step 5 of algorithm II.1 preserve pool sizes, there exists a permutation  $p$  of the indices  $(1, 2, \dots, m)$  such that  $|G_{p(g)}| = |G_g|$  for all  $g \in \{1, 2, \dots, m\}$ . ■

**Proposition II.1** For a given partition  $\Omega = \{G_1, G_2, \dots, G_m\}$  of  $S$ , if

$$I \frac{\partial^2 h(I, |G_g|)}{\partial I^2} + 2 \frac{\partial h(I, |G_g|)}{\partial I} \geq 0, \quad \forall I \in [0, |G_g|], \quad \text{and} \quad \forall G_g \in \Omega \quad (23)$$

then

$$\frac{I+1}{2I} h(I+1, |G_g|) + \frac{I-1}{2I} h(I-1, |G_g|) \geq h(I, |G_g|), \quad \forall I \in \{1, 2, \dots, |G_g|\}, \quad \text{and} \quad \forall G_g \in \Omega, \quad (24)$$

but the converse is not necessarily true.

**Proof:** The condition

$$I \frac{\partial^2 h(I, k)}{\partial I^2} + 2 \frac{\partial h(I, k)}{\partial I} \geq 0, \quad \forall I \in [0, k],$$

holds if and only if

$$\frac{\partial^2 I h(I, k)}{\partial I^2} \geq 0, \quad \forall I \in [0, k],$$

i.e., if and only if the function  $f : [0, k] \rightarrow \mathbb{R}$  such that

$$f(I) = Ih(I, k) \quad \forall I \in [0, k]$$

is convex. Clearly, if  $f(\cdot)$  is convex, hypothesis 1 holds.

But we can find instances in which equation 23 does not hold for some  $I \in [0, k]$ , and yet equation 24 holds for every  $I \in \{1, 2, \dots, k-1\}$ . Indeed, suppose that  $k = 3$  and

$$h(I, k) = 1/[1 + \exp(-20(I/k - 3/4))],$$

then hypothesis 1 holds (for  $I = 1$  and  $I = 2$ ), even though condition 8 does not hold for non-integer values of  $I$  greater than 2 and sufficiently close to  $k = 3$ .  $\blacksquare$

### III When ordered pooling minimizes the expected number of false positives

For any arbitrary group  $G_g \subseteq S$  s.t.  $|G_g| \geq 2$ , it can be shown that

$$FP_{G_g} \equiv \begin{cases} (1 - S_p)(1 - q_i), & \text{if } G_g = \{i\}, \\ \sum_{I=0}^{|G_g|} P_{G_g}(I)h(I, |G_g|)(|G_g| - I)(1 - S_p) & \text{if } |G_g| \geq 2 \end{cases}$$

corresponds to the expected number of false positives from group  $G_g$ .<sup>3</sup>

For any arbitrary group  $G_g \subseteq S$  such that  $|G_g| \geq 2$  and any  $j \in G_g$  we define

$$C_{G_g, j} \equiv \sum_{I=0}^{|G_g|-1} P_{G_g \setminus j}(I)h(I+1, |G_g|)(|G_g| - (I+1))(1 - S_p),$$

$$D_{G_g, j} \equiv \sum_{I=0}^{|G_g|-1} P_{G_g \setminus j}(I)h(I, |G_g|)(|G_g| - I)(1 - S_p).$$

From the above expressions,  $C_{G_g, j}$  is the expected number of false positives in group  $G_g$  conditional that  $j \in G_g$  is infected. Similarly,  $D_{G_g, j}$  is the expected number of false positives in group  $G_g$  conditional that  $j \in G_g$  is not infected.

Now notice that, for any  $j \in G_g$ ,

$$\begin{aligned} FP_{G_g} &= q_j C_{G_g, j} + (1 - q_j) D_{G_g, j} \\ &= q_j (C_{G_g, j} - D_{G_g, j}) + D_{G_g, j}. \end{aligned} \tag{25}$$

**Lemma III.1** *If hypothesis 2 holds for a given  $k \geq 2$ , then for any arbitrary group  $G_g \subseteq S$  such that  $|G_g| = k$ , and any  $l \in G_g$ ,*

$$C_{G_g, l} - D_{G_g, l}$$

*is decreasing in the probability of infection from each subject in  $G_g \setminus \{l\}$ .*

**Proof:** Notice that

$$C_{G_g, l} - D_{G_g, l} = \sum_{I=0}^{k-1} P_{G_g \setminus \{l\}}(I)(1 - S_p)[(k - I - 1)h(I + 1, k) - (k - I)h(I, k)],$$

---

<sup>3</sup>That this is true for  $|G_g| = 1$  is trivial. To see this is true for  $|G_g| > 1$ , define  $FP_{G_g, I, d}$  as the expected number of false positives in group  $G_g$  conditional that the group has exactly  $I$  infected subjects and conditional that infection is detected in the first stage of testing. Similarly, define  $FP_{G_g, I, nd}$  as the expected number of false positives in group  $G_g$  conditional that the group has exactly  $I$  infected subjects and conditional that infection is *not* detected in the first stage of testing. Then, by the law of iterated expectations we must have

$$FP_{G_g} = \sum_{I=0}^{|G_g|} P_{G_g}(I) \left[ h(I, |G_g|) FP_{G_g, I, d} + (1 - h(I, |G_g|)) FP_{G_g, I, nd} \right].$$

Noticing that  $FP_{G_g, I, d} = (k - I)(1 - S_p)$  and  $FP_{G_g, I, nd} = 0$ , yields the desired result.

which corresponds to a weighted average of  $(k - I - 1)(1 - S_p)h(I + 1, k) - (k - I)(1 - S_p)h(I, k)$ , where the weights are determined by the probability mass function  $P_{G_g \setminus \{l\}}(\cdot)$ . Clearly, increasing the probability of infection from a subject in  $G_g \setminus \{l\}$  causes this average to put more weight on higher values of  $I$  (formally, letting  $Y$  be the random variable associated with the probability mass function  $P_{G_g \setminus \{l\}}(\cdot)$  and  $Y'$  be its transformed version after the probability of infection from a subject in  $G_g \setminus \{l\}$  is increased, we have that  $Y'$  first-order stochastically dominates  $Y$ ). So it suffices to show that  $(k - I - 1)(1 - S_p)h(I + 1, k) - (k - I)(1 - S_p)h(I, k)$  is decreasing in  $I$ , a property that is satisfied if, for any  $I \in \{1, 2, \dots, k - 1\}$ ,

$$\begin{aligned} & (k - I - 1)h(I + 1, k) - (k - I)h(I, k) - \\ & - [(k - I)h(I, k) - (k - I + 1)h(I - 1, k)] \leq 0 \\ \iff & \frac{k - I - 1}{2(k - I)}h(I + 1, k) + \frac{k - I + 1}{2(k - I)}h(I - 1, k) \leq h(I, k). \end{aligned}$$

■

**Lemma III.2** *Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$  such that  $|G_1| = k_1 \geq 2$  and  $|G_2| = k_2 \geq 2$ . Then consider the following ordered partitions of  $G_1 \cup G_2$ :*

$$\{G_1^*, G_2^*\},$$

and

$$\{G_2^{**}, G_1^{**}\},$$

where  $|G_1^*| = |G_1^{**}| = k_1$ ,  $|G_2^*| = |G_2^{**}| = k_2$ ,  $i < j$  for all  $i \in G_1^*$  and all  $j \in G_2^*$  and  $i > j$  for all  $i \in G_1^{**}$  and all  $j \in G_2^{**}$ . If  $h(\cdot, k_1)$  and  $h(\cdot, k_2)$  satisfy hypothesis 2, then

$$\min\{FP_{G_1^*} + FP_{G_2^*}, FP_{G_1^{**}} + FP_{G_2^{**}}\} \leq FP_{G_1} + FP_{G_2}.$$

**Proof:** If  $\{G_1, G_2\} = \{G_1^*, G_2^*\}$  or  $\{G_1, G_2\} = \{G_1^{**}, G_2^{**}\}$ , the proof is trivial. So suppose that  $\{G_1, G_2\} \neq \{G_1^*, G_2^*\}$  and  $\{G_1, G_2\} \neq \{G_1^{**}, G_2^{**}\}$ , so that

$$\min_{i \in G_1} i < \max_{j \in G_2} j$$

and

$$\min_{j \in G_2} j < \max_{i \in G_1} i.$$

For an arbitrary  $i \in G_1$  and  $j \in G_2$  either one of the following inequalities must hold

$$C_{G_1, i} - D_{G_1, i} \leq C_{G_2, j} - D_{G_2, j} \tag{26}$$

or

$$C_{G_2, j} - D_{G_2, j} \leq C_{G_1, i} - D_{G_1, i} \tag{27}$$

1. Assume that inequality 27 holds, and define

$$\bar{i} \equiv \max_{i \in G_1} i$$

and

$$\underline{j} \equiv \min_{j \in G_2} j.$$

Because  $q_{\bar{i}} \geq q_i$ , lemma III.1 implies that

$$C_{G_1, i} - D_{G_1, i} \leq C_{G_1, \bar{i}} - D_{G_1, \bar{i}}. \tag{28}$$

Analogously,  $q_{\underline{j}} \leq q_j$  and lemma II.1 imply that

$$C_{G_2, \underline{j}} - D_{G_2, \underline{j}} \leq C_{G_1, j} - D_{G_1, j}. \tag{29}$$

Together inequalities 27, 28 and 29 imply that

$$C_{G_2, \underline{j}} - D_{G_2, \underline{j}} \leq C_{G_1, \bar{i}} - D_{G_1, \bar{i}}. \tag{30}$$

Now let us exchange the position of subjects  $\underline{i}$  and  $\bar{j}$  to create the new groups

$$\tilde{G}_1 = G_1 \cup \{\bar{j}\} \setminus \{\underline{i}\}$$

and

$$\tilde{G}_2 = G_2 \cup \{\underline{i}\} \setminus \{\bar{j}\}.$$

Then we must have  $FP_{\tilde{G}_1} + FP_{\tilde{G}_2} \leq FP_{G_1} + FP_{G_2}$ . Indeed, from expression 11, we have that

$$\begin{aligned} FP_{G_1} &= q_{\bar{i}} C_{G_1, \bar{i}} + (1 - q_{\bar{i}}) D_{G_1, \bar{i}}, \\ FP_{G_2} &= q_{\underline{j}} C_{G_2, \underline{j}} + (1 - q_{\underline{j}}) D_{G_2, \underline{j}}. \end{aligned}$$

Moreover, because

$$\begin{aligned} C_{\tilde{G}_1, \underline{j}} &= C_{G_1, \bar{i}}, \\ D_{\tilde{G}_1, \underline{j}} &= D_{G_1, \bar{i}}, \\ C_{\tilde{G}_2, \bar{i}} &= C_{G_2, \underline{j}}, \\ D_{\tilde{G}_2, \bar{i}} &= D_{G_2, \underline{j}}, \end{aligned}$$

we also have that

$$\begin{aligned} FP_{\tilde{G}_1} &= q_{\underline{j}} C_{G_1, \bar{i}} + (1 - q_{\underline{j}}) D_{G_1, \bar{i}}, \\ FP_{\tilde{G}_2} &= q_{\bar{i}} C_{G_2, \underline{j}} + (1 - q_{\bar{i}}) D_{G_2, \underline{j}}. \end{aligned}$$

Therefore,

$$\begin{aligned} &FP_{\tilde{G}_1} + FP_{\tilde{G}_2} \leq FP_{G_1} + FP_{G_2} \\ \iff &q_{\underline{j}}(C_{G_1, \bar{i}} - D_{G_1, \bar{i}}) + q_{\bar{i}}(C_{G_2, \underline{j}} - D_{G_2, \underline{j}}) \leq q_{\bar{i}}(C_{G_1, \bar{i}} - D_{G_1, \bar{i}}) + q_{\underline{j}}(C_{G_2, \underline{j}} - D_{G_2, \underline{j}}) \\ \iff &(q_{\bar{i}} - q_{\underline{j}})(C_{G_2, \underline{j}} - D_{G_2, \underline{j}}) \leq (q_{\bar{i}} - q_{\underline{j}})(C_{G_1, \bar{i}} - D_{G_1, \bar{i}}) \\ \iff &(C_{G_2, \underline{j}} - D_{G_2, \underline{j}}) \leq (C_{G_1, \bar{i}} - D_{G_1, \bar{i}}), \end{aligned}$$

which, from inequality 30, holds.

Now, defining

$$\bar{i}' \equiv \max_{i \in \tilde{G}_1} i$$

and

$$\underline{j}' \equiv \min_{j \in \tilde{G}_2} j,$$

we have that either one of the following conditions must hold:

- (a)  $q_{\underline{j}} \geq q'_{\bar{i}}$ , in which case  $\tilde{G}_1 = G_1^*$  and  $\tilde{G}_2 = G_2^*$ , so that  $FP_{G_1^*} + FP_{G_2^*} \leq FP_{G_1} + FP_{G_2}$ .
- (b)  $q_{\underline{j}} < q'_{\bar{i}}$ . In this case, notice that when we moved from  $G_1$  to  $\tilde{G}_1$  we decreased the probability of infection from one subject in this group while not altering the probability of infection from the remaining subjects in the group. From lemma III.1 this implies that

$$C_{G_1, \bar{i}} - D_{G_1, \bar{i}} \leq C_{\tilde{G}_1, \bar{i}'} - D_{\tilde{G}_1, \bar{i}'}. \quad (31)$$

Analogously, when moving from  $G_2$  to  $\tilde{G}_2$  we increased the probability of infection from one subject in this group while not altering the probability of infection from the remaining subjects in the group. From lemma III.1 this implies that

$$C_{\tilde{G}_2, \underline{j}'} - D_{\tilde{G}_2, \underline{j}'} \leq C_{G_2, \underline{j}} - D_{G_2, \underline{j}}. \quad (32)$$

Together inequalities 30, 31 and 32 imply that

$$C_{\tilde{G}_2, \underline{j}'} - D_{\tilde{G}_2, \underline{j}'} \leq C_{\tilde{G}_1, \bar{i}'} - D_{\tilde{G}_1, \bar{i}'}$$

So we can redefine  $G_1 = \tilde{G}_1$ ,  $G_2 = \tilde{G}_2$ ,  $\bar{i} = \bar{i}'$  and  $\underline{j} = \underline{j}'$  and repeat the previous steps iteratively, until the final pair of groups is given by  $G_1^*$  and  $G_2^*$ .

Because at each step of this algorithm we reduce the expected number of false negatives for subjects in these groups, and because at the end of this process we obtain the groups  $G_1^*$  and  $G_2^*$ , we have that

$$FP_{G_1^*} + FP_{G_2^*} \leq FP_{G_1} + FP_{G_2},$$

as we wanted to show.

2. If inequality 26 holds, the proof is analogous to case the case in which inequality 27 holds (case 1). Indeed, when 26 holds, we define  $\underline{i} = \min_{i \in G_1} i$  and  $\bar{j} = \max_{j \in G_2} j$ , and then repeat the same steps in case 27 to show that switching the positions of subjects  $\underline{i}$  and  $\bar{j}$  weakly reduces the expected number of false positives, provided that  $h(\cdot, k_1)$  and  $h(\cdot, k_2)$  both satisfy hypothesis 2. Then, we iteratively switch subjects from the new groups in the same fashion until we reach the ordered partition  $\{G_2^{**}, G_1^{**}\}$ . Because at each step of the algorithm the expected number of false positives diminishes, we obtain  $FP_{G_1^{**}} + FP_{G_2^{**}} \leq FP_{G_1} + FP_{G_2}$ . ■

**Lemma III.3** *Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$  such that  $|G_1| = k \geq 1$  and  $|G_2| = 1$ . Let  $\bar{j} = \max(G_1 \cup G_2)$  (i.e.,  $\bar{j}$  is the subject with highest probability of infection in group  $G_1 \cup G_2$ ). Then consider the following ordered partition of  $G_1 \cup G_2$ :*

$$\{G_1^*, G_2^*\},$$

where

$$G_1^* \equiv G_1 \cup G_2 \setminus \{\bar{j}\},$$

$$G_2^* \equiv \{\bar{j}\},$$

i.e.,  $\{G_1^*, G_2^*\}$  is the ordered partition that groups the  $k$  subjects with lowest probability of infection together, and the subject with highest probability of infection alone.

If  $h(\cdot, k)$  is increasing, then

$$FP_{G_1^*} + FP_{G_2^*} \leq FP_{G_1} + FP_{G_2}.$$

**Proof:** Let  $G_1$  and  $G_2$  be two disjoint subsets of  $S$ . The proof when  $|G_1| = |G_2| = 1$  is trivial. So suppose that  $|G_1| = k_1 \geq 2$  and  $G_2 = \{j\}$ , with  $j < \bar{j} = \max(G_1 \cup G_2)$ . Let

$$G_1^* \equiv G_1 \cup G_2 \setminus \{j\},$$

$$G_2^* \equiv \{j\}.$$

Because  $G_2 = \{j\}$  and  $G_2^* = \{\bar{j}\}$ , we have that  $FP_{G_2} = (1 - S_p)(1 - q_j)$  and  $FP_{G_2^*} = (1 - S_p)(1 - q_{\bar{j}})$ . Therefore,

$$\begin{aligned}
& FP_{G_1^*} + FP_{G_2^*} \leq FP_{G_1} + FP_{G_2} \\
\iff & (1 - S_p) \sum_{I=0}^k P_{G_1^*}(I)(k - I)h(I, k) + (1 - S_p)(1 - q_{\bar{j}}) \leq (1 - S_p) \sum_{I=0}^k P_{G_1}(I)(k - I)h(I, k) + (1 - S_p)(1 - q_j) \\
\iff & (1 - S_p) \left( q_j \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I)(k - I - 1)h(I + 1, k) + (1 - q_j) \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I)h(I, k)(k - I) \right) \leq \\
& (1 - S_p) \left( q_{\bar{j}} \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I)(k - I - 1)h(I + 1, k) + (1 - q_{\bar{j}}) \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I)h(I, k)(k - I) \right) + (1 - S_p)(q_{\bar{j}} - q_j) \\
\iff & (q_{\bar{j}} - q_j) \left( \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I) [(k - I)h(I, k) - (k - I - 1)h(I + 1, k)] \right) \leq (q_{\bar{j}} - q_j) \\
\iff & \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I) [(k - I)h(I, k) - (k - I - 1)h(I + 1, k)] \leq 1, \tag{33}
\end{aligned}$$

Now notice that, since  $h(\cdot, k)$  is increasing and since  $h(I, k) \leq (1 - S_p) \leq 1$ , we have that

$$\begin{aligned}
\sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I) [(k - I)h(I, k) - (k - I - 1)h(I + 1, k)] & \leq \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I) [(k - I)h(I + 1, k) - (k - I - 1)h(I + 1, k)] \\
& = \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I)h(I + 1, k) \\
& \leq \sum_{I=0}^{k-1} P_{G_1 \setminus \{\bar{j}\}}(I) = 1,
\end{aligned}$$

which implies that inequality 33 is satisfied. ■

### III.1 Proof of theorem 3

Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be an arbitrary partition of  $S = \{1, 2, \dots, n\}$  such that  $|G_g| \geq 2$  for all  $G_g \in \Omega$ , where  $m$  is the number of groups from the partition (e.g., if all of the groups from the partition have the same size  $k$ , then  $m = \frac{n}{k}$ ). Also suppose that hypothesis 2 holds for every  $k \in \{|G_1|, |G_2|, \dots, |G_m|\}$ . Then implement the following algorithm:

#### Algorithm III.1

1. Initialize the partition  $\tilde{\Omega} = \{\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_m\}$ , where  $\tilde{G}_g = G_g$  for all  $g \in \{1, 2, \dots, m\}$ .
2. Initialize  $g = 1$ .
3. Set  $w = g + 1$ .
4. Pick groups  $\tilde{G}_g$  and  $\tilde{G}_w$  from  $\tilde{\Omega}$ .
  - (a) If  $|G_g| \geq 2$  and  $|G_w| \geq 2$ , consider the following ordered partitions of  $\tilde{G}_g \cup \tilde{G}_w$ :

$$\{G_g^*, G_w^*\},$$

and

$$\{G_w^{**}, G_g^{**}\},$$

where  $|G_g^*| = |G_w^{**}| = |\tilde{G}_g|$ ,  $|G_w^*| = |G_g^{**}| = |\tilde{G}_w|$ ,  $i < j$  for all  $i \in G_g^*$  and all  $j \in G_w^*$ , and  $q_i > q_j$  for all  $i \in G_g^{**}$  and all  $j \in G_w^{**}$ .

If  $FP_{G_g^*} + FP_{G_w^*} \leq FP_{G_g^{**}} + FP_{G_w^{**}}$ , redefine

$$\tilde{G}_g = G_g^* \quad \text{and} \quad \tilde{G}_w = G_w^*$$

else, redefine

$$\tilde{G}_g = G_w^{**} \quad \text{and} \quad \tilde{G}_w = G_g^{**}.$$

(b) If  $|\tilde{G}_g| = 1$  or  $|\tilde{G}_w| = 1$ , redefine

$$\tilde{G}_g = \tilde{G}_g \cup \tilde{G}_w \setminus \{\max(\tilde{G}_g \cup \tilde{G}_w)\},$$

and

$$\tilde{G}_w = \{\max(\tilde{G}_g \cup \tilde{G}_w)\}.$$

5. If  $w = m$ , proceed to the next step. Else, redefine  $w = w + 1$  and repeat step 4.

6. If  $g = m - 1$ , stop the algorithm, else, redefine  $g = g + 1$  and repeat step 3.

Let  $\Omega = \{G_1, G_2, \dots, G_m\}$  be the original partition and  $\tilde{\Omega} = \{\tilde{G}_1, \tilde{G}_2, \dots, \tilde{G}_m\}$  be the final partition obtained after implementing algorithm III.1. From lemmas III.2 and III.3, at each step of this algorithm the overall expected number of false positives weakly diminishes, which implies that  $\mathbb{E}[FP(\tilde{\Omega})] \leq \mathbb{E}[FP(\Omega)]$ . Because at the end of each step 6 of algorithm III.1 we have that, for each  $w > m$ ,  $q_i \leq q_j$  for all  $i \in \tilde{G}_g$  and all  $j \in \tilde{G}_w$ , we have that  $\tilde{\Omega}$  is an ordered partition of  $S$ . Moreover, because changes made at steps 4a and 4b of algorithm III.1 preserve pool sizes, there exists a permutation  $p$  of the indices  $(1, 2, \dots, m)$  such that  $|G_{p(g)}| = |G_g|$  for all  $g \in \{1, 2, \dots, m\}$ .

As to the second part of the theorem, notice that, at each step 4b of algorithm III.1 we are allocating the subject with highest probability of infection to be tested individually. This implies that if a subject  $i$  is tested individually under  $\tilde{\Omega}$ , then a subject  $j$  with  $q_j > q_i$  is also tested individually under  $\tilde{\Omega}$ . ■

## IV Expected number of tests, Expected number of false negatives and Expected number of false positives for extreme dilution functions

It is a well known result that, in the absence of dilution effects, ordered pooling is more desirable than other pooling criterion in terms of minimizing expected number of tests, the expected number of false positives and the expected number of false negatives. This result is stated formally in proposition IV.1.

**Proposition IV.1** (Aprahamian, Bish and Bish (2019)) *Suppose that the dilution function is given by:*

$$h(I, k) = \begin{cases} S_e, & \text{if } I > 0 \\ 1 - S_p, & \text{if } I = 0. \end{cases}$$

Then,

- For any partition  $\Omega$  of  $S$ , there exists an ordered partition  $\Omega^*$  such that  $\mathbb{E}[T(\Omega^*)] \leq \mathbb{E}[T(\Omega)]$ .
- The probability that an infected subject is incorrectly diagnosed as healthy is the same across any partition  $\Omega$  of  $S$  such that  $|G_g| \geq 2$  for all  $G_g \in \Omega$ , and is given by  $(1 - S_e^2)$ .
- For any partition  $\Omega$  of  $S$  such that  $|G_g| \geq 2$  for all  $G_g \in \Omega$ , there exists an ordered partition  $\Omega^*$  such that  $\mathbb{E}[FP(\Omega^*)] \leq \mathbb{E}[FP(\Omega)]$ .

**Proof:** Suppose that the dilution function is given by

$$h(I, k) = \begin{cases} S_e, & \text{if } I > 0 \\ 1 - S_p, & \text{if } I = 0. \end{cases}$$

Then,

- Ordered pooling minimizes the expected number of tests. Indeed this follows directly from theorem 1 and the fact that  $h(\cdot, k)$  is concave.
- Any matching criteria produces the same expected number of false negatives. Indeed, consider any arbitrary infected patient from the population. Then regardless of whom he is matched with in the pooled sample, the probability that his pooled sample tests positive for infection is  $(1 - S_e)$ . If the pooled test does accuse infection, which happens with probability  $S_e$ , the infected subject is individually tested and diagnosed as healthy with probability  $(1 - S_e)$ . So using simple probability theory, the overall probability that an arbitrary infected subject is incorrectly diagnosed as healthy is given by

$$(1 - S_e) + S_e(1 - S_e) = 1 - S_e^2,$$

which does not depend on whom the subject is matched with in the pool.

- c) Ordered pooling minimizes the expected number of false positives. Indeed, this follows directly from theorem 3 and the fact that this dilution function satisfies hypothesis 2. ■

So from proposition IV.1, we can see that, in the absence of dilution effects, ordered partitions seem like the optimal choice, as they generate the same  $\mathbb{E}[FN(\Omega)]$  as any other matching criterion, and they may also potentially minimize  $\mathbb{E}[T(\Omega)]$  and  $\mathbb{E}[FP(\Omega)]$  simultaneously (notice that the proposition does not guarantee that the ordered partition that minimizes  $\mathbb{E}[T(\Omega)]$  is the same as the ordered partition that minimizes  $\mathbb{E}[FP(\Omega)]$ ).

Next, we consider the case in which  $h(\cdot, k)$  is affine, i.e., when  $h(I, k) = a + bI$ . It can be shown that, in this case, an ordered partition also performs weakly better than any partition in which all pool sizes are equal in all of the three dimensions considered, namely,  $\mathbb{E}[T(\Omega)]$ ,  $\mathbb{E}[FN(\Omega)]$  and  $\mathbb{E}[FP(\Omega)]$ .

**Proposition IV.2** *Suppose that the dilution function is given by  $h(I, k) = a + bI$ . Then,*

- a) *For any pair of partitions  $\Omega$  and  $\Omega'$  of  $S$  such that  $|G_g| = k$  for all  $G_g \in \Omega \cup \Omega'$ , we must have  $\mathbb{E}[T(\Omega)] = \mathbb{E}[T(\Omega')]$ .*  
b) *If  $\Omega$  is a partition of  $S$  such that  $|G_g| = k$  for all  $G_g \in \Omega$ , and  $\Omega^*$  is an ordered partition of  $S$  such that  $|G_g^*| = k$  for all  $G_g^* \in \Omega^*$ , then  $\mathbb{E}[FN(\Omega^*)] \leq \mathbb{E}[FN(\Omega)]$  and  $\mathbb{E}[FP(\Omega^*)] \leq \mathbb{E}[FP(\Omega)]$ .*

**Proof:** Suppose that the dilution function is given by  $h(I, k) = a + bI$ , where  $b \geq 0$ .

- a) First, we show that the expected number of tests is not affected by the matching criteria used to form the pools.

Let  $X_{i,j} = 1$  if individual  $i$  from group  $j$  is infected. Moreover, assume that  $Prob(X_{i,j} = 1) = q_{i,j}$ . Then, the expected number of tests is given by

$$T \equiv \frac{n}{k} + \sum_{i=1}^{n/k} \sum_{l=0}^k h(l, k) Prob\left(\sum_{j=1}^n X_{i,j} = l\right).$$

Assuming that  $h(I, k) = a + bI$ , we have that

$$\begin{aligned} T &= \frac{n}{k} + \sum_{i=1}^{n/k} \sum_{l=0}^k \left\{ [a + bI] Prob\left(\sum_{j=1}^n X_{i,j} = l\right) \right\} \\ &= \frac{n}{k} + \sum_{i=1}^{n/k} \left[ a + b \sum_{l=0}^k l Prob\left(\sum_{j=1}^n X_{i,j} = l\right) \right]. \end{aligned} \quad (34)$$

Because  $\sum_{l=0}^k l Prob(\sum_{j=1}^n X_{i,j} = l)$  equals to the expected number of infected in group  $i$ , and because the number of infected in each group follows a Poisson-binomial distribution with parameters  $(k, \{q_{i,j}\}_{j=1}^k)$ , this expectation is given by  $\sum_{j=1}^k q_{i,j}$ . Replacing this into expression 34, we get

$$\begin{aligned} T &= \frac{n}{k} + \sum_{i=1}^{n/k} \left[ a + b \sum_{j=1}^k q_{i,j} \right] \\ &= \frac{n}{k} + \frac{n}{k} a + b \sum_{i=1}^{n/k} \sum_{j=1}^k q_{i,j}, \end{aligned}$$

which clearly does not depend on how the  $q_{i,j}$ 's are grouped.

- b) As to the proof that ordered pooling achieves a minimum in the expected number of false negatives, it follows directly from theorem 2 and the fact that  $h(I, k) = a + bI$  (with  $b \geq 0$ ) satisfies hypothesis 1. Finally, notice that hypothesis 2 is clearly satisfied when  $h(I, k)$  is affine, so that ordered pooling also minimizes the expected number of false positives.

Finally, we consider the extreme case in which the dilution effect is so strong that the probability of detecting infection from a pool comprised of at least one healthy subject is given by  $1 - S_p$ , the same probability of detecting infection from a healthy subject through an individual test. In this case we show that, conditional that all pools have the same size, ordered partitions generates the *highest* expected number of tests, though it minimizes the expected number of false negatives and generates the same expected number of false positives as any other pooling criteria. ■

**Proposition IV.3** *Suppose that the dilution function is given by*

$$h(l, k) = \begin{cases} 1 - S_p, & \text{if } l < k \\ S_e, & \text{if } l = k. \end{cases}$$

Then,

- If  $\Omega^*$  is an ordered partition of  $S$  such that  $|G_g^*| = k$  for all  $G_g^* \in \Omega^*$ , and  $\Omega$  is any partition of  $S$  such that  $|G_g| = k$  for all  $G_g \in \Omega$ , then  $\mathbb{E}[T(\Omega^*)] \geq \mathbb{E}[T(\Omega)]$ .
- If  $\Omega$  is a partition of  $S$  such that  $|G_g| = k$  for all  $G_g \in \Omega$ , and  $\Omega^*$  is an ordered partition of  $S$  such that  $|G_g^*| = k$  for all  $G_g^* \in \Omega^*$ , then  $\mathbb{E}[FN(\Omega^*)] \leq \mathbb{E}[FN(\Omega)]$ .
- The probability that a healthy subject is incorrectly diagnosed as infected is the same across any partition  $\Omega$  of  $S$  such that  $|G_g| \geq 2$  for all  $G_g \in \Omega$ , and is given by  $(1 - S_p)^2$ .

**Proof:** Suppose that the dilution function is given by

$$h(l, k) = \begin{cases} 1 - S_p, & \text{if } l < k \\ S_e, & \text{if } l = k. \end{cases}$$

- First, we will show that ordered pooling *maximizes* the expected number of tests.

Consider a partition  $\Omega \equiv \{G_1, G_2, \dots, G_{n/k}\}$  of  $S = \{1, 2, \dots, n\}$ , where  $|G_g| = k$  for all  $G_g \in \Omega$  (i.e., each group has the same size). The probability that group  $G_g$  has exactly  $k$  infected subjects is given by  $P_{G_g}(k) = \prod_{j \in G_g} q_j$ , and the expected number of tests from this partition is given by

$$\mathbb{E}[T(\Omega)] = \frac{n}{k} + n(1 - S_p) + k(S_p + S_e - 1) \sum_{g \in \Omega} P_{G_g}(k).$$

Because  $k(S_p + S_e - 1)$  is a positive constant (recall that we assume that  $S_e > 1 - S_p$ ), we have that the partition that minimizes the expected number of tests is the one that *minimizes*  $\sum_{G_g \in \Omega} P_{G_g}(k)$ . We now show that ordered pooling *maximizes*  $\sum_{G_g \in \Omega} P_{G_g}(k)$ .

Suppose by way of contradiction that  $\Omega = \{G_1, G_2, \dots, G_{n/k}\}$  maximizes  $\sum_{G_g \in \Omega} P_{G_g}(k)$  among all partitions that have all pool sizes equal to  $k$ , and suppose that  $\Omega$  is not an ordered partition. Then there are  $G_l, G_w \in \Omega$  such that  $P_{G_l}(k) < P_{G_w}(k)$  and there is a  $i \in G_l$  and a  $j \in G_w$  such that  $(1 - q_i) > (1 - q_j)$ . Then, denoting  $A \equiv \prod_{\tau \in G_l \setminus \{i\}} (1 - q_\tau)$  and  $B \equiv \prod_{\tau \in G_w \setminus \{j\}} (1 - q_\tau)$ , we must have that  $A < B$ . Moreover, if  $\Omega$  maximizes  $\sum_{G_g \in \Omega} P_{G_g}(k)$  we must have

$$\begin{aligned} (1 - q_i)A + (1 - q_j)B &\geq (1 - q_j)A + (1 - q_i)B \\ \iff (1 - q_j)(B - A) &\geq (1 - q_i)(B - A) \\ \iff (1 - q_j) &\geq (1 - q_i), \end{aligned}$$

a contradiction with  $(1 - q_i) > (1 - q_j)$ .  $\rightarrow \leftarrow$

- As to the proof that ordered pooling minimizes the expected number of false negatives, it follows directly from theorem 2 and the fact that this dilution function clearly satisfies hypothesis 1.

c) Finally, under this extreme dilution effect, if a healthy subject is pooled tested, the test will detect infection with probability  $1 - S_p$  regardless of who the healthy patient is matched with. In this contingency, the subject is tested again as is declared infected with probability  $1 - S_p$ . Because those events are independent, the overall probability that a healthy subject is diagnosed as infected is  $(1 - S_p)^2$ . Because this probability does not depend on who the subject is matched with, this implies that any matching mechanism generates the same expected number of false positives. ■

While it is very unlikely that pooled testing schemes are ever considered as a good alternative to individual testing under such an extreme dilution effect, proposition IV.3 helps to illustrate that a partition used to form the pools may perform well in one dimension (e.g., by minimizing the expected number of false negatives), but poorly in others (e.g., by generating an excessively high expected number of tests).

## V When ordered pooling maximizes social welfare

### V.1 Proof of proposition 2

It suffices to show that ordered pooling maximizes the minimum utility and the sum of utilities. That ordered pooling maximizes the sum of utilities follows immediately from corollary 3. So it only remains to show that ordered pooling maximizes the minimum utility.

Without loss of generality, we can assume that there are only two groups, and that the probability of infection from subjects is given by

$$q_1 \leq q_2 \leq q_3 \leq q_4.$$

Then, the ordered partition is given by

$$\Omega^* = \{\{1, 2\}, \{3, 4\}\}.$$

The only possible alternative partitions are

$$\Omega' = \{\{1, 3\}, \{2, 4\}\}$$

and

$$\Omega'' = \{\{1, 4\}, \{2, 3\}\}.$$

1. Suppose that  $\lambda = 1$ .

For any  $q_i \in [0, 1]$ , let

$$A(q_i) \equiv q_i[1 - S_e h(2, 2)] + (1 - q_i)[1 - S_e h(1, 2)],$$

i.e.,  $A(q_i)$  is the probability that a subject is *not* detected as infected given that this subject is infected and the person this subject is matched with has probability of infection of  $q_i$ .

In this case, if the ordered partition  $\Omega^*$  is implemented, subjects' utilities are given by

$$u_1(\Omega^*) = 1 - q_1 A(q_2), \quad u_2(\Omega^*) = 1 - q_2 A(q_1), \quad u_3(\Omega^*) = 1 - q_3 A(q_4), \quad u_4(\Omega^*) = 1 - q_4 A(q_3).$$

The utilities from the other 2 other possible partitions,  $\Omega'$  and  $\Omega''$ , are given by

$$u_1(\Omega') = 1 - q_1 A(q_3), \quad u_2(\Omega') = 1 - q_2 A(q_4), \quad u_3(\Omega') = 1 - q_3 A(q_1), \quad u_4(\Omega') = 1 - q_4 A(q_2),$$

and

$$u_1(\Omega'') = 1 - q_1 A(q_4), \quad u_2(\Omega'') = 1 - q_2 A(q_3), \quad u_3(\Omega'') = 1 - q_3 A(q_2), \quad u_4(\Omega'') = 1 - q_4 A(q_1).$$

Because  $A(q_i) \geq 0$ , we have that a subject's utility is decreasing in the subject's probability of infection. Moreover, because  $h(\cdot, 2)$  is increasing (see assumption 1), we have that  $[1 - S_e h(2, 2)] \leq [1 - S_e h(1, 2)]$ , which implies that  $A(q_i)$  is decreasing in  $q_i$ . This implies that a subject's utility is increasing in the probability of infection of the subject he is matched with. This implies that the subject with lowest utility from each group is the subject who has the highest probability of infection from that group.

So suppose that, under ordered pooling, subject 2 is the one with the lowest utility, i.e., suppose that

$$u_2(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*).$$

In this case, we have that

$$u_3(\Omega') = 1 - q_3 A(q_1) \leq 1 - q_2 A(q_1) = u_2(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

and

$$u_4(\Omega'') = 1 - q_4 A(q_1) \leq 1 - q_2 A(q_1) = u_2(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

so that ordered pooling maximizes the minimum utility.

Suppose, instead, that subject 4 is the one with lowest utility under ordered pooling, i.e., suppose that

$$u_4(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*).$$

In this case, we have that

$$u_4(\Omega') = 1 - q_4 A(q_2) \leq 1 - q_4 A(q_3) = u_4(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

and

$$u_4(\Omega'') = 1 - q_4 A(q_1) \leq 1 - q_4 A(q_3) = u_4(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

so that again, ordered pooling maximizes the minimum utility.

## 2. Suppose that $\lambda = 0$ .

For any  $q_i \in [0, 1]$ , let

$$B(q_i) \equiv q_i(1 - S_p)h(1, 2) + (1 - q_i)(1 - S_p)h(0, 2),$$

i.e.,  $B(q_i)$  is the probability that a subject is detected as infected given that this subject is *not* infected and the person this subject is matched with has probability of infection of  $q_i$ .

In this case, if the ordered partition  $\Omega^*$  is implemented, subjects' utilities are given by

$$u_1(\Omega^*) = 1 - (1 - q_1)B(q_2), \quad u_2(\Omega^*) = 1 - (1 - q_2)B(q_1), \quad u_3(\Omega^*) = 1 - (1 - q_3)B(q_4), \quad u_4(\Omega^*) = 1 - (1 - q_4)B(q_3).$$

The utilities from the other 2 other possible partitions,  $\Omega'$  and  $\Omega''$ , are given by

$$u_1(\Omega') = 1 - (1 - q_1)B(q_3), \quad u_2(\Omega') = 1 - (1 - q_2)B(q_4), \quad u_3(\Omega') = 1 - (1 - q_3)B(q_1), \quad u_4(\Omega') = 1 - (1 - q_4)B(q_2),$$

and

$$u_1(\Omega'') = 1 - (1 - q_1)B(q_4), \quad u_2(\Omega'') = 1 - (1 - q_2)B(q_3), \quad u_3(\Omega'') = 1 - (1 - q_3)B(q_2), \quad u_4(\Omega'') = 1 - (1 - q_4)B(q_1).$$

Clearly, a subject's utility is increasing in its own probability of infection. Moreover, because  $h(\cdot, 2)$  is increasing (see assumption 1), we have that  $(1 - S_p)h(1, 2) \leq (1 - S_p)h(2, 2)$ , which implies that  $B(q_i)$  is increasing in  $q_i$ . Because the term  $B(q_i)$  enters negatively in the utility function from each agent, we have that a subject's utility is decreasing in the probability of infection of the subject he is matched with. This implies that the subject with lowest utility from each group is the subject who has the *lowest* probability of infection from that group.

So suppose that, under ordered pooling, subject 1 is the one with the lowest utility, i.e., suppose that

$$u_1(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*).$$

In this case, we have that

$$u_1(\Omega') = 1 - (1 - q_1)B(q_3) \leq 1 - (1 - q_1)B(q_2) = u_1(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

and

$$u_1(\Omega'') = 1 - (1 - q_1)B(q_4) \leq 1 - (1 - q_1)B(q_2) = u_1(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

so that ordered pooling maximizes the minimum utility.

Suppose, instead, that subject 3 is the one with lowest utility under ordered pooling, i.e., suppose that

$$u_3(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*).$$

In this case, we have that

$$u_2(\Omega') = 1 - (1 - q_2)B(q_4) \leq 1 - (1 - q_3)B(q_4) = u_3(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

and

$$u_1(\Omega'') = 1 - (1 - q_1)B(q_4) \leq 1 - (1 - q_3)B(q_4) = u_3(\Omega^*) = \min_{i \in \{1,2,3,4\}} u_i(\Omega^*),$$

so that again, ordered pooling maximizes the minimum utility. ■

## V.2 Proof of proposition 3

For a given pool size of  $k \in \mathbb{N}$ , we can assume, without loss of generality, that there are only two pools (i.e., that  $m = 2$ ).<sup>4</sup> Consider the following ordered partition of  $S = \{1, 2, \dots, n\}$

$$\Omega^* = \{G_1^*, G_2^*\},$$

where

$$G_1^* = \{1, 2, \dots, k\}$$

and

$$G_2^* = \{k+1, k+2, \dots, n\},$$

and  $|G_1^*| = |G_2^*| = k$ .

- D) Suppose that  $\theta = 1$ . Consider any partition  $\Omega = \{G_1, G_2\}$ , such that  $|G_1| = |G_2| = k$ . For each group  $\tilde{G}_g \subseteq \Omega$ , define

$$A_{i, G_g} \equiv \sum_{I=0}^{k-1} (1 - S_e h(I+1, k)) P_{G \setminus \{i\}(I)},$$

i.e.,  $A_{i, G_g}$  is the probability that a subject in group  $G_g$  is *not* detected as infected given that this subject is infected.

Then, the utility from each subject  $i \in \{1, 2, \dots, n\}$  under partition  $\Omega$  is given by

$$u_i(\Omega) = \begin{cases} 1 - q_i A_{i, G_1}, & \text{if } i \in G_1 \\ 1 - q_i A_{i, G_2}, & \text{if } i \in G_2 \end{cases}$$

Clearly, the utility from each subject is decreasing in their own probability of infection. Because the dilution function  $h(\cdot, k)$  is increasing, we also have that, for any group  $G_g$  and any  $j \in G_g \setminus \{i\}$ ,  $A_{i, G_g}$  is decreasing in  $q_j$ . This implies that the utility from any subject is *increasing* in the probability of infection of the subjects he is matched with. Therefore, the subject with lowest utility from each group is the subject who has the highest probability of infection in that group.

Because under ordered pooling subject  $n$  (i.e., the one with highest probability of infection) is matched with the the  $k-1$  subjects with highest probability of infection excluding subject  $n$ , we have that subject  $n$ 's utility is maximized under ordered pooling. Therefore,

$$u_n(\Omega^*) \geq u_n(\Omega) \geq \min_{i \in S} u_i(\Omega),$$

as we wanted to show.

---

<sup>4</sup>To extend the result for  $m > 2$ , one only needs to follow an algorithm similar to the one presented in the proofs of theorems 1, 2 and 3.

II) Suppose that  $\theta = 0$ . Consider any partition  $\Omega = \{G_1, G_2\}$ , such that  $|G_1| = |G_2| = k$ . For each group  $G_g \subseteq \Omega$ , define

$$B_{i,G_g} \equiv \sum_{I=0}^{k-1} (1 - S_p) h(I, k) P_{G_g \setminus \{i\}}(I),$$

i.e.,  $B_{i,G_g}$  is the probability that a subject in group  $G_g$  is detected as infected given that this subject is *not* infected.

Then, the utility from each subject  $i \in \{1, 2, \dots, n\}$  under partition  $\Omega$  is given by

$$u_i(\Omega) = \begin{cases} 1 - (1 - q_i) A_{i,G_1}, & \text{if } i \in G_1 \\ 1 - (1 - q_i) A_{i,G_2}, & \text{if } i \in G_2 \end{cases}$$

Clearly, the utility from each subject is *increasing* in their own probability of infection. Because the dilution function  $h(\cdot, k)$  is increasing, we also have that, for any group  $G_g$  and any  $j \in G_g \setminus \{i\}$ ,  $B_{i,G_g}$  is *increasing* in  $q_j$ . This implies that the utility from any subject is *decreasing* in the probability of infection of the subjects he is matched with. Therefore, the subject with lowest utility from each group is the subject who has the *lowest* probability of infection in that group.

Because under ordered pooling subject 1 (i.e., the one with lowest probability of infection) is matched with the the  $k - 1$  subjects with lowest probability of infection excluding subject 1, we have that subject 1's utility is maximized under ordered pooling. Therefore,

$$u_n(\Omega^*) \geq u_n(\Omega) \geq \min_{i \in S} u_i(\Omega),$$

as we wanted to show. ■

### V.3 Proof of theorem 5

Theorem 5 follows directly from proposition 3 and theorems 2 and 3, and the fact that, if a partition maximizes the sum of utilities and minimum utility, it also maximizes the utilitarian max-min welfare function for any parameter  $\delta \in [0, 1]$ . ■

## VI Sufficient Conditions for the Dilution Function to be Discrete-Concave

### VI.1 Proof of proposition 4

To avoid clutter notation, define  $h(I) \equiv h(I, k)$ .

1. First, let us show that  $h(\cdot)$  is discrete-concave. A sufficient condition for this to be true is that  $h''(I) \leq 0$  for  $I \geq 1$  and that

$$h(2) + h(0) \leq 2h(1).$$

- I) First, let us show that  $h''(I) \leq 0$  for  $I \geq 1$ . First, notice that

$$h(I) = \int_{\left(\frac{y-\mu_I}{\sigma_I}\right)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy,$$

where

$$\mu_I = \frac{I}{k} \mu_+ + \frac{k-I}{k} \mu_-,$$

and

$$\sigma_I = \sqrt{\frac{I}{k^2} \sigma_+^2 + \frac{(k-I)}{k^2} \sigma_-^2}.$$

Taking the first derivative of this expression, we get

$$h'(I) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\frac{y-\mu_I}{\sigma_I}\right)^2}{2}} \left[ \frac{(\mu_+ - \mu_-)}{\sigma_I} - \frac{(\mu_I - y)(\sigma_+^2 - \sigma_-^2)}{2\sigma_I^3 k^2} \right].$$

Defining

$$D_I \equiv \frac{(\mu_+ - \mu_-)}{\sigma_I} - \frac{(\mu_I - y)(\sigma_+^2 - \sigma_-^2)}{2\sigma_I^3 k^2},$$

we have that  $D_I$  is greater than zero for all  $I \geq 1$ . Indeed, since  $\mu_I \leq \mu_+$  and  $\mu_- < \underline{y}$ , we have that

$$\mu_+ - \mu_- \geq \mu_I - \underline{y}.$$

Therefore, a sufficient condition for  $D_I \geq 0$  is that

$$\begin{aligned} \frac{1}{\sigma_I} &\geq \frac{\sigma_+^2 - \sigma_-^2}{2k^2\sigma_I^3} \\ \Leftrightarrow \sigma_I^2 &\geq \frac{\sigma_+^2 - \sigma_-^2}{2k^2} \\ \Leftrightarrow \frac{I(\sigma_+^2 - \sigma_-^2) + k\sigma_-^2}{k^2} &\geq \frac{\sigma_+^2 - \sigma_-^2}{2k^2} \\ \Leftrightarrow k\sigma_-^2 &\geq \left(\frac{1}{2} - I\right)(\sigma_+^2 - \sigma_-^2), \end{aligned}$$

a condition that always holds for  $I \geq 1$ , since  $(\frac{1}{2} - I)(\sigma_+^2 - \sigma_-^2) < 0$  for all  $I \geq 1$  and  $k\sigma_-^2 > 0$ .

Now, taking the second derivative of  $h(\cdot)$ , we get

$$\begin{aligned} h''(I) &= -\frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\frac{y-\mu_I}{\sigma_I}\right)^2}{2}} \left(\frac{\mu_I - y}{\sigma_I}\right) D_I^2 \\ &+ \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\frac{y-\mu_I}{\sigma_I}\right)^2}{2}} \left[ \frac{3(\mu_I - y)(\sigma_+^2 - \sigma_-^2)^2}{4\sigma_I^5 k^4} - \frac{(\mu_+ - \mu_-)(\sigma_+^2 - \sigma_-^2)}{\sigma_I^3 k^2} \right] \end{aligned} \quad (35)$$

Clearly, the term

$$-\frac{1}{\sqrt{2\pi}} e^{-\frac{\left(\frac{y-\mu_I}{\sigma_I}\right)^2}{2}} \left(\frac{\mu_I - y}{\sigma_I}\right) D_I^2$$

from equation (35) is negative. So it suffices to show that

$$\left[ \frac{3(\mu_I - y)(\sigma_+^2 - \sigma_-^2)^2}{4\sigma_I^5 k^4} - \frac{(\mu_+ - \mu_-)(\sigma_+^2 - \sigma_-^2)}{\sigma_I^3 k^2} \right] \leq 0 \quad (36)$$

Because  $\mu_+ - \mu_- \geq \mu_I - \underline{y}$ , a sufficient condition for inequality (36) to hold is that

$$\begin{aligned} \frac{(\sigma_+^2 - \sigma_-^2)}{\sigma_I^3 k^2} &\geq \frac{3(\sigma_+^2 - \sigma_-^2)^2}{4\sigma_I^5 k^4} \\ \Leftrightarrow \sigma_I^2 &\geq \frac{3(\sigma_+^2 - \sigma_-^2)}{4k^2} \\ \Leftrightarrow \frac{I(\sigma_+^2 - \sigma_-^2) + k\sigma_-^2}{k^2} &\geq \frac{3(\sigma_+^2 - \sigma_-^2)}{4k^2} \\ \Leftrightarrow k\sigma_-^2 &\geq \left(\frac{3}{4} - I\right)(\sigma_+^2 - \sigma_-^2), \end{aligned}$$

a condition that always holds for  $I \geq 1$ , since  $\sigma_+^2 \geq \sigma_-^2$ .

II) Now, let us show that

$$h(2) + h(0) \leq 2h(1) \iff h(2) - h(1) \leq h(1) - h(0).$$

This condition holds if and only if

$$\int_{k\left(\frac{y-\mu_2}{\sigma_2}\right)}^{k\left(\frac{y-\mu_1}{\sigma_1}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \leq \int_{k\left(\frac{y-\mu_1}{\sigma_1}\right)}^{k\left(\frac{y-\mu_0}{\sigma_0}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \quad (37)$$

Because  $\frac{y-\mu_0}{\sigma} > 0$ , and  $\frac{y-\mu_1}{\sigma} < 0$  and  $\frac{y-\mu_2}{\sigma} < 0$ , and because the standard normal distribution has a greater mass around 0, we have that a sufficient condition for inequality (37) to hold is that

$$\begin{aligned} \frac{y-\mu_1}{\sigma_1} - \frac{y-\mu_2}{\sigma_2} &\leq \frac{y-\mu_0}{\sigma_0} - \frac{y-\mu_1}{\sigma_1} \\ \iff \frac{2(y-\mu_1)}{\sigma_1} &\leq \frac{2(y-\mu_0)}{\sigma_0} + \frac{2(y-\mu_2)}{\sigma_2} \\ \iff \frac{2(y-\mu_1)}{\sigma_1} &\leq \frac{2(y-\mu_0)}{\sqrt{\sigma_1^2 - \varepsilon}} + \frac{2(y-\mu_2)}{\sqrt{\sigma_1^2 + \varepsilon}}, \end{aligned} \quad (38)$$

where

$$\varepsilon \equiv \frac{\sigma_+^2 - \sigma_-^2}{k^2}.$$

One can easily show that this inequality holds for  $\varepsilon = 0$ , i.e., when  $\sigma_+^2 = \sigma_-^2$ . Moreover, since  $y - \mu_0 > 0$  and  $y - \mu_1 < 0$ , we have that the right hand side of inequality (38) is increasing in  $\varepsilon$ . Therefore, the inequality holds for any  $\sigma_+^2 \geq \sigma_-^2$ .

2. Now, let us show that  $h(I)$  is increasing. We have already seen that  $h'(I) \geq 0$  for  $I \geq 0$ . So it only remains to show that

$$\begin{aligned} h(1) &\geq h(0) \\ \iff \int_{\left(\frac{y-\mu_1}{\sigma_1}\right)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy &\geq \int_{\left(\frac{y-\mu_0}{\sigma_0}\right)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy. \end{aligned}$$

Since  $\frac{y-\mu_1}{\sigma_1} < 0$  and  $\frac{y-\mu_0}{\sigma_0} > 0$ , this condition holds. ■

## VII Proof of proposition 1

For a given pool of size  $k$ , and for a given patient  $i$  belonging to this pool, let  $P_{-i}(I)$  denote the probability that exactly  $I \leq k - 1$  of the other members in the pool are infected. Then the probability that patient  $i$  is incorrectly diagnosed as not infected is given by

$$q_i \sum_{I=0}^{k-1} [F(I+1)(1 - S_e) + (1 - F(I+1))] P_{-i}(I),$$

which is greater than or equal to

$$q_i \sum_{I=0}^{k-1} [F(I+1)(1 - S_e) + (1 - F(I+1))(1 - S_e)] P_{-i}(I) = q_i(1 - S_e),$$

the probability the subject is incorrectly diagnosed as not infected if he is tested individually. ■

## VIII Calibrating the dilution function for Chlamydia

Following Aprahamian, Bish and Bish (2018), we assume that the dilution function for pooled testing for Chlamydia follows the format

$$h(I, k) = (1 - S_p) + (S_e + S_p - 1)(I/k)^\delta. \quad (39)$$

In their experiments, Kacena et al. (1998) found that pools of size  $k = 4$  exhibited perfect sensitivity and specificity of

$$1 - h(0, 4) = 97/98 = .98.$$

Because pooled testing with pools of size  $k = 4$  exhibited perfect sensitivity, it would be natural to assume that individual testing should also exhibit perfect sensitivity. But because previous studies seem to suggest otherwise (e.g., Schachter et al. (1994)), it is possible that the perfect sensitivity observed in Kacena et al. (1998) was the result of sampling variation. So we conservatively set  $S_e = 0.99$ .

Assuming that specificity is constant among all pool sizes, we must have  $h(0, 1) = h(0, 4)$ , so we set  $S_p = 1 - h(0, 4) = 97/98$ .

In Kacena et al. (1998), the prevalence of Chlamydia for pools of size  $k = 10$  was given by  $\mu_r = 62/520$ , and the proportion of times infection was detected in infected pooled samples of size  $k = 10$  was given by  $37/38$ . So, if we were to follow Aprahamian, Bish and Bish (2018)'s approach, we would set  $\delta$  so that

$$37/38 = \frac{1}{1 - (1 - \mu_r)^{10}} \sum_{I=1}^{10} h(I, 10) \binom{10}{I} \mu_r^I (1 - \mu_r)^{10-I},$$

i.e., we would choose  $\delta$  so that the sensitivity for pools of size 10 obtained in Kacena et al. (1998) is consistent with the dilution function  $\hat{h}$ . This would procedure  $\delta = 0.0089$ , so that we would end up with very small dilution effects. But given that it is possible that the results from Kacena et al. (1998) are subject to sampling variation, we conservatively choose much higher values for  $\delta$  in our numerical exercises (between 0.1 and 0.2), thus allowing for the existence of non-negligible dilution effects.

## IX HBV infection: calibrating the cost of a false negative

Because we assume subjects are being screened for HBV infection for blood donation, we assume that the cost of a false negative is equal to the expected cost of infecting someone with HBV. Because some infected patients may become asymptomatic, while others may exhibit acute symptoms, which may then progress to chronic infection and then possibly death, we use a simple Markov specification to model how patients transition from each of these states.

Following Jackson et al. (2003) and Birkmeyer et al. (1993), we assume that, after receiving a blood transfusion infected with *HBV*, the patient may either be asymptomatic for the rest of his life or exhibit symptoms of acute infection following the transfusion. If the patient has acute infection, he may either recover and be cured, or progress to chronic infection, which is not curable. Chronic infection not only considerably reduces the patient's quality of life, but also increases the probability that the patient dies at the end of every period.

Costs and transition probabilities from this model were extracted from Jackson et al. (2003) and Birkmeyer et al. (1993) and are summarized in table 1. Following the literature, we set the monetary value of 1 quality-adjusted life year (QALY) to be equal to \$50,000 (i.e., one year of perfect healthy is worth \$50,000). We also assume a yearly discount factor of 3%. These parameters result in a total expected cost of a false negative to be equal to 1811.264.

## X Hepatitis B: performance of ordered pooling using parametric estimates of the dilution effect

In the main text we used Monte Carlo simulations to estimate the cumulative distribution of OD readings of pooled samples. In this section we use a parametric approach instead, by assuming that the distribution of OD readings from infected and non-infected subjects follow a normal distribution.

Suppose that the *OD* reading of a healthy and infected subject are governed by the random variables  $X_-$  and  $X_+$ , respectively. Applying the method of moments to our dataset assuming that both  $X_-$  and  $X_+$  are normally distributed yields the following estimate for the distribution of  $X_-$  and  $X_+$ :  $X_- \sim N(0.1448, 0.007)$  and  $X_+ \sim N(20, 340)$ .

Table 1: Costs and transition probabilities

<b>Costs</b>	<b>Estimate</b>
Acute viral hepatitis	One-time cost of \$800 for medical evaluation, plus 2.5% probability of one-time hospitalization at \$5,000 <sup>1</sup>
Chronic hepatitis B	\$150 per year, plus a one-time cost of \$1,200 (liver biopsy) <sup>1</sup>
<b>Quality-of-life adjustments</b>	<b>Estimate</b>
Acute hepatitis (symptomatic)	2 weeks (one-time subtraction) <sup>1</sup>
Hospitalization for acute hepatitis	1 week (one-time subtraction) <sup>1</sup>
Symptomatic chronic hepatitis	0.9 <sup>1</sup>
<b>Transition Probabilities</b>	<b>Estimate</b>
Acute Infection (just after receiving infected blood)	25% <sup>1</sup>
Chronic Infection (just after an acute infection)	10% <sup>2</sup>
Symptoms, given chronic hepatitis	14% <sup>1</sup>
Yearly increase in probability of death, given chronic hepatitis	0.35% <sup>1</sup>

<sup>1</sup> Jackson et al. (2003);

<sup>2</sup> Birkmeyer et al. (1993)

## XI Hepatitis B case study using a smaller prevalence rate

Prisons are arguably not the best place to find reliable blood donors, given that blood borne infections, such as HBV and HIV are highly prevalent among inmates (e.g., Smith et al. (2017)). So our numerical analysis based on this dataset only provides a conservative estimate of the benefits of implementing pooled testing as opposed to individual testing, as pooled testing is usually more effective when used to screen populations with a low prevalence rate (Kim et al. (2007)). So in this section we divide the estimated probability of infection from each subject by 10, in order to mimic situations in which the tester is interested in screening blood infection from non-inmates, who are much less likely to be infected.

The results of our simulations are depicted in table 2 below. As it can be seen from the table, under small prevalence levels the benefits of implementing *OR* as opposed to *IT* are much higher: under the parametric approach, *OR* generates costs that are 62.7% lower compared with *IT*. Under the non-parametric approach, this difference is even greater (74.5%).

## References

- Aprahamian, Hrayr, Douglas R. Bish, and Erub K. Bish.** 2019. "Optimal Risk-Based Group Testing." *Management Science*, 65(9): 4365–4384.
- Aprahamian, Hrayr, Ebru K. Bish, and Douglas R. Bish.** 2018. "Adaptive risk-based pooling in public health screening." *IIE Transactions*, 50(9): 743–766.
- Birkmeyer, J.D., L.T. Goodnough, J.P. AuBuchon, P.G. Noordsij, and B. Littenberg.** 1993. "The cost-effectiveness of preoperative autologous blood donation for total hip and knee replacement." *Transfusion*, 33(7): 544–551.
- Jackson, B.R., M.P. Busch, S.L. Stramer, and J.P. AuBuchon.** 2003. "The cost-effectiveness of NAT for HIV, HCV, and HBV in whole-blood donations." *Transfusion*, 43(6): 721–729.
- Kacena, Katherine A., Sean B. Quinn, René Howell, Guillermo E. Madico, Thomas C. Quin, and Charlotte A. Gaydos.** 1998. "Pooling Urine Samples for Ligase Chain Reaction Screening for Genital Chlamydia trachomatis Infection in Asymptomatic Women." *Journal of Clinical Microbiology*, 36(2): 481–485.
- Kim, Hae-Young, Michael G. Hudgens, Jonathan M. Dreyfuss, Daniel J. Westreich, and Christopher D. Pilcher.** 2007. "Comparison of Group Testing Algorithms for Case Identification in the Presence of Test Error." *Biometrics*, 63(4): 1152–1163.
- Schachter, J, WE Stamm, T C Quinn, WW Andrews, JD Burczak, and H H Lee.** 1994. "Ligase chain reaction to detect Chlamydia trachomatis infection of the cervix." *Journal of Clinical Microbiology*, 32(10): 2540–2543.

Table 2: Performance measure of the optimal ordered partition (*OR*) compared to the optimal random partition (*R1*), the optimal *random partition with a cutoff* (*R2*) and the lower bound for each of the attributes considered (*LB*) for the Hepatitis B case study, when estimated probabilities of infection are multiplied by 0.1.

Non-Parametric estimation of $h(I, k)$ multiplying prevalences by 0.1							
Model	$\mathbb{E}[FN]$	$\max\{\mathbb{E}[FN_i]\}$	$\mathbb{E}[FP]$	$\mathbb{E}[T]$	$\mathbb{E}[C]/n$	$\mathbb{E}[\max\{k\}]$	$\mathbb{E}[k]$
<i>OR</i>	0.0654	0.001	0.1752	19.9697	4.6286	16.304	12.4722
$\widehat{OR}$	0.0628	0.001	0.1675	20.4681	4.656	14.31	11.1089
% ( $\widehat{OR} - OR$ )	-4.1%	-0.7%	-4.6%	2.4%	0.6%	-13.9%	-12.3%
R1	0.0672	0.0012	0.1809	20.3112	4.7216	13.0	12.5
% ( <i>R1-OR</i> )	2.6%	21.9%	3.2%	1.7%	2.0%	-25.4%	0.2%
R2	0.0672	0.0012	0.181	20.3163	4.7234	13.0	12.5
% ( <i>R2-OR</i> )	2.7%	23.0%	3.2%	1.7%	2.0%	-25.4%	0.2%
IT	0.0147	0.0003	1.5603	100.0	18.1577	1	1
% ( <i>IT-OR</i> )	-1147.2%	-281.3%	88.8%	80.0%	74.5%	-1530.4%	-1147.2%
LB	0.0147	-	0.089	19.6671	3.5774	-	-
% ( <i>OR-LB</i> )	77.6%	-	49.2%	1.5%	22.7%	-	-
Parametric estimation of $h(I, k)$ multiplying prevalences by 0.1							
Model	$\mathbb{E}[FN]$	$\max\{\mathbb{E}[FN_i]\}$	$\mathbb{E}[FP]$	$\mathbb{E}[T]$	$\mathbb{E}[C]/n$	$\mathbb{E}[\max\{k\}]$	$\mathbb{E}[k]$
<i>OR</i>	0.2335	0.0041	0.0043	16.8352	6.9957	15.964	12.525
$\widehat{OR}$	0.2335	0.0041	0.0043	16.8326	6.996	16.0	12.4944
% ( $\widehat{OR} - OR$ )	0.0%	0.1%	-0.3%	-0.0%	0.0%	0.2%	-0.2%
R1	0.234	0.0042	0.0045	17.2129	7.0675	13.0	12.5
% ( <i>R1-OR</i> )	0.2%	1.8%	4.3%	2.2%	1.0%	-22.8%	-0.2%
R2	0.2337	0.0041	0.0045	17.2035	7.0606	13.0	12.5
% ( <i>R2-OR</i> )	0.1%	1.4%	4.2%	2.1%	0.9%	-22.8%	-0.2%
IT	0.1268	0.0022	0.0531	100.0	18.7568	1	1
% ( <i>IT-OR</i> )	-1152.5%	-82.4%	91.8%	83.2%	62.7%	-1496.4%	-1152.5%
LB	0.1267	-	0.0006	16.8319	5.0566	-	-
% ( <i>OR-LB</i> )	45.8%	-	87.2%	0.0%	27.7%	-	-

**Smith, Jacob M., A. Ziggy Uvin, Alexandria Macmadu, and Josiah D. Rich.** 2017. "Epidemiology and Treatment of Hepatitis B in Prisoners." *Current Hepatology Reports*, 16(1).