Pooling designs

Benoît R. Kloeckner [*][†]

November 18, 2020

Assume we are given N takings from patients to be analyzed for presence of a virus by RT-PCR. While in some cases the best course of action is to individually test every taking, pooling has been proposed as a way to save time and reagents in limited supplies and to overcome a limited number of PCR machines; the important feature of RT-PCR in this respect is the low sensitivity to dilution: up to some extent and up to the usual limits in sensitivity and specificity, a pool where several takings are mixed will be positive if and only if at least one of the patients has the disease. The simplest pooling method is to gather (fractions of) takings in pools of k, and analyze each pool. If the result is negative, all takings are negative, and if the result is positive then takings are reanalyzed individually; one is left with optimizing k using an estimation of the prevalence in the patients. More sophisticated methods have been considered, notably by Mutesa et al.[**?**]. We propose to explore which combinatorial structures have best properties for pooling RT-PCR tests. This document is a draft, outlining a few direction of research. It is oriented toward the involved mathematics, but keeping an eye on practical application to testing.

The mathematical object which best fits the situation to be modeled is called a hypergraph.

Definition 1. A *hypergraph* is a pair (V, E) where V is a set (whose elements are called vertices) and E is a set of non-empty subsets of V (whose elements are called edges). Given a vertex $x \in V$, let x^* be the set of edges containing x. Given a subset $X \subset V$ of vertices, let $X^* = \{e \in E \mid \exists x \in X, x \in e\}$ be the set of all edges incident to some element of X.

Let us define a *pooling design* as a hypergraph (V, E) satisfying the following property:

$$\forall x \in V, \forall X \subset V, \quad x^* = X^* \implies X = \{x\}$$

The interpretation is as follows: each vertex corresponds to a patient (or rather their taking), and each edge to a pool where fractions of takings will be mixed to be analyzed together. The condition that $x^* = X^*$ should only occurs when $X = \{x\}$ ensures that,

[*] Univ Paris Est Creteil, CNRS, LAMA, F-94010 Creteil, France

[†] Univ Gustave Eiffel, LAMA, F-77447 Marne-la-Vallée, France

whenever there is at most one positive taking, its uniqueness is guaranteed by the tests and it can be identified.

Given a pooling design (V, E), there are several numbers of importance to analyze its practical interest in pooled testing:

- its *order* $v = |V|$, i.e. the number of patients considered in one application of the design (to analyze N takings, we need $v \geq N$ up to filling some vertices with empty takings, or to divide N into several groups of at most v patients),
- its *size* $e = |E|$, i.e. the number of RT-PCR involved in one application of the design,
- its *compression rate* $r = \frac{e}{v}$, i.e. the factor applied to the number of RT-PCR involved compared to the individual testing – we aim for $r < 1$ (but it can in certain circumstances be worth accepting $r \geq 1$ in exchange for redundancy, when one expect too many false negatives or false positives which are non-reproduced in different tests (in particular, independent from the quality of the taking); this will not be addressed here),
- its *max degree* $\Delta = \max\{|x^*| \colon x \in V\}$, i.e. the maximum number of fractions some taking must be divided into (can be limited by the amount of RNA obtained in each taking),
- its *max edge size* $S = \max\{|e| \colon e \in E\}$, i.e. the maximal number of sub-takings to be pooled for a run of RT-PCR (can be limited by sensibility requirements: volume limitation of PCR machines implies that excessive dilution can reduce the RT-PCR sensibility),
- its *detection capacity*, i.e. the maximal number of positive taking that can be guaranteed and identified. Formally, letting $\mathcal{P}_{\leq n}(V)$ be the set of subsets of V with at most n elements, we set
$$c = \max\Big\{n \colon \forall X, Y \in \mathcal{P}_{\leq n}(V), X^* = Y^* \implies X = Y\Big\}.$$
The definition of a pooling design ensures $c \geq 1$, but larger is better. A refined analysis would involve, for each outcome with $n > c$, a second round of testing to count and identify all positives.

Example 2. The individual testing consist in taking V the set of all N takings, and $E = \Big\{\{x\} \colon x \in V\Big\}$: each edge is a single vertex. Then we have

$$v = e = N \qquad\qquad r = 1 \qquad\qquad \Delta = S = 1 \qquad\qquad c = N$$

Example 3. The hypercube design of [?] with dimension $D \geq 2$ consist in taking $V = \{1, 2, 3\}^D$ and E the set of coordinate slices, i.e.

$$E = \bigcup_{k=1}^{D} \Big\{\{1, 2, 3\}^{k-1} \times \{i\} \times \{1, 2, 3\}^{D-k} \colon i \in \{1, 2, 3\}\Big\}.$$

It has

$$v = 3^D \qquad e = 3D \qquad r = \frac{D}{3^{D-1}} \qquad \Delta = D \qquad S = 3^{D-1} \qquad c = 1$$

The detection capacity has the good property to not depend on a prior assumption on the number of positives (if there are more than one, the design detects this event), and while for large D the detection capacity is low compared to the order, subsequent rounds of testing can be used efficiently. The main question we want to raise is whether it is possible to improve on this state-of-the-art pooling design.

Question 4. Given upper bounds on Δ and S, which values of v, r, c or v, r, γ are realized by a pooling design?

Proposition 5. *Let (V, E) be a pooling design of order v, size e and detection capacity c. Then*

$$e \geq vH\Big(\frac{c}{v}\Big) - \frac{1}{2}\log_2(c) - \frac{3}{2}$$

where $H(x) := -x\log_2(x) - (1-x)\log_2(1-x)$ is the binary entropy function. In particular, the compression rate satisfies

$$r \geq H\Big(\frac{c}{v}\Big) - o_{v\to\infty}(1)$$

In particular, since the prevalence at which the pooling design can be used is mostly driven by c/v, we have a quantitative estimate of how much a large prevalence prevents a good compression rate.

Proof. There are 2^e possible sets of results of the PCR on all edges, which must suffice to distinguish a number of cases at least (neglecting the case where there are more than c positive takings) $\sum_{k=0}^{c}\binom{v}{k}$. The inequality follows from the estimate

$$\sum_{k=0}^{c}\binom{v}{k} \geq \frac{2^{vH(c/v)}}{\sqrt{8c(1-c/v)}}$$

by neglecting the first term 1, taking the $\log_2$ of both sides and removing a positive term on the right. □

Table 1 compares $H(c/v)$ and the actual compression rate for the hypercube design with various values of D. Some room for improvement seems available, but only by a factor less than 2: these pooling designs are not too far from optimal in their prevalence regime.

Let us give some more examples issued from the field of *incidence geometry*.

Example 6. The complete quadrilateral can be described with $V = \{1,2,3,4,5,6\}$ and $E = \big\{\{1,2,3\},\{3,4,5\},\{5,6,2\},\{1,4,6\}\big\}$. It has

$$v = 6 \qquad e = 4 \qquad r = \frac{2}{3} \qquad \Delta = 2 \qquad S = 3 \qquad c = 1$$

D	2	3	4	5	6	$\to\infty$
$r \simeq$	0.67	0.33	0.15	0.062	0.025	
$H(c/v) \simeq$	0.50	0.23	0.096	0.039	0.015	$0.53\,r$

Table 1: Comparison of the compression rate r of the hypercube design with its (asymptotic) lower bound.

For comparison, we note that $H(c/v) \simeq 0.65$, very close to the compression rate: this pooling design is close to optimal in its prevalence regime (although the small v makes it more likely that more takings than expected turn out positive; concretely, the probability to have more than 1 positive taking drops below 5% at prevalence $\leq 6.28\%$).

Example 7. The dual of the Hesse configuration (dual meaning that vertices are represented by lines, and edges by points) has (if I computed correctly)

$$v = 12 \qquad e = 9 \qquad r = \frac{3}{4} \qquad \Delta = 3 \qquad S = 4 \qquad c = 2$$

Again $H(c/v) \simeq 0.65$ but the compression rate is somewhat higher at 0.75. Compared to the complete quadrilateral, the larger v makes it less likely that the number of positive takings exceeds the average value (precisely, the probability to have more than 2 positive taking drops below 5% at prevalence $\leq 7.18\%$), giving this pooling design slightly more cases of application.

Question 8. In view of the previous examples, does there exist pooling designs with $v \gg 1$, $c/v \simeq 1/6$ and compression rate $\simeq 2/3$?

Such pooling designs could operate at prevalence up to $\sim 16\%$ while being optimal at these rates. Currently, only pooled test with many rounds can operate at such high prevalences, using pooling design would make testing highly parallel.

Example 9. The Schläfli double six has

$$v = 30 \qquad e = 12 \qquad r = \frac{2}{5} \qquad \Delta = 2 \qquad S = 5 \qquad c = 1$$

Here $H(c/v) \simeq 0.21$ while the compression rate is twice larger; there should be some room for improvement. This pooling design can be used up to prevalence 1.19% with less than 5% probability to exceed the detection capacity.

Last, we propose an alternative version of the hypercube designed to increase the detection ratio.

Example 10. Let p be a prime number (or a primitive number) and $\mathbb{F}_p$ be the Field with p elements. Choose a dimension $D \geq 2$ and a parameter $k \geq D$. We set $V = \mathbb{F}_p^D$ (for $p = 3$, we thus have the same vertex set than in the hypercube design). Let $(\phi_1, \dots, \phi_k)$ be linear forms such that any D of them are linearly independent. Without loss of generality, we can assume $(\phi_1, \dots, \phi_D)$ is the canonical dual basis (i.e. $\phi_i(x_1, \dots, x_D) = x_i$). Last, we let E be the set of all levels of all the ϕ_i:

$$E = \Big\{ \phi_i^{-1}(y) \colon i \in \{1, \dots, k\}, y \in \mathbb{F}_p \Big\}.$$

The pooling design (V, E) is called the *generalized hybercube design* of parameters (p, D, k). It has

$$v = p^D \qquad e = kp \qquad r = \frac{k}{p^{D-1}} \qquad \Delta = k \qquad S = p^{D-1}$$

and the remaining question is how large can be c.

Question 11. What is the value of c in the previous example? Given v_0, what choice of p, D, k such that $v \simeq v_0$ minimizes $\frac{r}{H(c/v)}$? Given a prevalence, what is the best value of r that can be achieved with a generalized hypercube for which detection capacity is exceeded with probability less than 5%?

Question 12. Given one of the above most efficient pooling design, find an algorithm to manage the case when the detection capacity is exceeded, with as few new rounds as possible, each having as few tests as possible.