Le dépistage massif est un élément essentiel de la lutte contre la propagation du coronavirus. Mais comment parer à une éventuelle pénurie de réactifs et de matériel ? En opérant les tests sur des mélanges de prélèvements et en faisant appel aux mathématiques.

Christian Genest
Université McGill
Christiane Rousseau
Université de Montréal

Le dépistage par gr

Avec la crainte anticipée d'une deuxième vague de COVID-19, de nombreux experts s'accordent à dire que l'instauration d'un plan de dépistage à grande échelle est nécessaire pour enrayer la propagation du coronavirus. Réalisés auprès d'échantillons représentatifs de la population, les tests immunologiques, sérologiques ou antigéniques permettraient aussi d'estimer la prévalence de la maladie, de juger du degré d'immunité collective et d'adapter les moyens de gestion de la pandémie.

Pour qu'elle soit couronnée de succès, la mise en œuvre d'une stratégie de dépistage massif présuppose l'accès à des ressources conséquentes en personnel et en matériel. Avec l'accroissement de la demande à l'échelle planétaire, une pénurie des produits réactifs nécessaires aux analyses en laboratoire s'est rapidement profilée à l'horizon et demeure une préoccupation des autorités de santé publique au Canada et dans le reste du monde.

Sachant que, pour la plupart, les tests s'avèrent (fort heureusement) négatifs, peut-on mettre les mathématiques à profit pour faire mieux ? Il s'avère que oui, notamment en réalisant des tests de groupe sur des mélanges de prélèvements judicieusement construits.

Le dépistage par groupe

Imaginons qu'un laboratoire ait reçu 100 prélèvements à des fins de dépistage. Il les divise au hasard en 5 groupes de taille 20. Puis, groupe par groupe, il utilise la moitié de chacun des 20 prélèvements pour constituer un mélange auquel il applique le test.

Si un test effectué sur un mélange est négatif, on peut tout de suite conclure qu'aucun membre du groupe concerné n'est infecté. Si le test est positif, alors on procède à des tests individuels sur la deuxième moitié de chacun des 20 prélèvements.

Si les 100 prélèvements d'origine proviennent de personnes saines, cette procédure permet de s'en assurer en faisant 5 tests plutôt que 100. Si un seul individu est infecté, il suffit de 5 + 20 = 25 tests pour le repérer. Si deux personnes sont infectées, on peut aussi les identifier avec 25 tests si elles sont dans le même groupe mais il en faut 5 + 20 + 20 = 45 si elles appartiennent à des groupes différents. Et ainsi de suite s'il y a trois personnes infectées ou plus.

Comme on peut le constater, le dépistage par groupe permet donc de réaliser d'importantes économies, pourvu bien sûr que la sensibilité et la spécificité du test ne soient pas affectées par le mélange, comme on l'a supposé ici puisque c'est très souvent le cas en pratique.

Le laboratoire aurait aussi pu appliquer la même stratégie de dépistage à 10 groupes de 10 prélèvements. Si un seul individu est infecté, on n'aurait alors eu besoin que de 20 tests pour l'identifier. En revanche, il aurait fallu 10 tests pour conclure que personne n'est infecté.

Laquelle de ces deux stratégies est la meilleure ? Et en existe-t-il d'autres qui leur soient préférables ? La réponse dépend de la *prévalence* de la maladie, c'est-à-dire de la proportion de la population qui est infectée.

Dans une note parue en 1943 dans *The Annals of Mathematical Statistics*, l'Américain Robert Dorfman rapporte que, dans sa forme la plus élémentaire, le dépistage par groupe avait déjà été utilisé pendant la Deuxième Guerre mondiale pour détecter les cas de syphilis parmi les conscrits. L'approche s'est imposée et il en existe aujourd'hui bien des variantes qui sont notamment employées partout en Amérique du Nord pour tester la présence du VIH, de la grippe ou du Virus du Nil occidental.

Optimiser l'algorithme

Dorfman a montré comment déterminer la taille optimale d'un groupe en fonction de la prévalence $p \in [0, 1]$ de la maladie. Dénotons par $n \geq 2$ la taille du groupe et supposons que ses membres constituent un échantillon aléatoire représentatif de la population.

Si X dénote le nombre inconnu de personnes infectées dans le groupe, cette variable obéit alors à une loi binomiale[1] de paramètres n et p, d'où

$$\Pr(X=0) = (1-p)^n,$$

puisque chaque individu a une probabilité $1-p$ d'être sain, et

$$\Pr(X>0) = 1 - \Pr(X=0) = 1-(1-p)^n.$$

Si $X=0$, on ne fera alors que $N=1$ test. Cependant si $X>0$, on fera $N=n+1$ tests. En moyenne, le nombre de tests qu'on effectuera, appelé espérance de N et noté $\mathrm{E}(N)$, est égal à

$$\begin{aligned}\mathrm{E}(N) &= 1\times\Pr(X=0) + (n+1)\times\Pr(X>0)\\ &= n+1-n(1-p)^n.\end{aligned}$$

Cette fonction est croissante en p. Si $p=0$, on a $\mathrm{E}(N)=1$, ce qui est évident puisque personne n'a la maladie et donc un seul test suffit pour le confirmer. Si $p=1$, on a $\mathrm{E}(N)=n+1$ parce que le premier test sera forcément positif.

Pour toute valeur de $p \in [0, 1]$, il est possible de déterminer le ***coût relatif*** lié à l'emploi du dépistage par groupe en étudiant le comportement du ratio

$$\mathrm{E}(N)/n = 1 + 1/n - (1-p)^n.$$

en fonction de n. Plus $\mathrm{E}(N)/n$ est petit, plus il est payant d'avoir recours aux tests par lot, à condition bien sûr que le ratio soit inférieur à 1. Quand $p=0$, on trouve $\mathrm{E}(N)/n = 1/n$, de sorte qu'on a intérêt à prendre n aussi grand que possible. Quand $p=1$, on a toujours

$$\mathrm{E}(N)/n = 1 + 1/n > 1$$

car le test de groupe est toujours positif et ne fait donc qu'ajouter au fardeau.

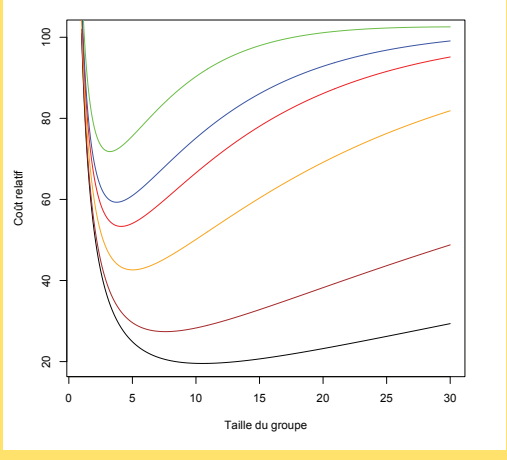

Figure 1
Tracé de la courbe $100\times E(N)/n$ en fonction de n pour différentes valeurs de p : 1 % (noir), 2 % (marron), 5 % (orange), 8 % (rouge), 10 % (bleu) et 15 % (vert).

1. *Il s'agit ici d'une approximation qui se justifie dans la mesure où la population est très grande par rapport à la taille des groupes.*

p (%)	n	Coût relatif (%)
1	11	20
2	8	27
5	5	43
8	4	53
10	4	59
15	3	72

Pour une valeur de p fixée, la fonction $100 \times E(N)/n$ représente le pourcentage moyen de tests effectués en fonction de la taille, n, du groupe. La figure montre le graphe de cette fonction pour différentes valeurs de p, correspondant à une prévalence de 1 % (noir), 2 % (marron), 5 % (orange), 8 % (rouge), 10 % (bleu) et 15 % (vert). Comme on peut le constater, la taille optimale des mélanges, qui correspond au minimum de la courbe, varie en fonction de la proportion, p, d'individus infectés dans la population. Le tableau ci-contre, rapporté par Dorfman, donne le choix optimal de n pour quelques valeurs de p.

Généralisations

Le protocole de test décrit ci-dessus est un exemple d'algorithme adaptatif à deux rondes. On le dit *adaptatif* parce que le choix (et donc le nombre) de tests à réaliser à la deuxième étape dépend du résultat du test réalisé au premier tour. Il existe plusieurs moyens d'améliorer la performance de ce type d'algorithme. La procédure peut notamment être étendue en augmentant le nombre de rondes.

Voici un algorithme classique, que nous appellerons *algorithme de division binaire*, et qui possède certaines propriétés d'optimalité (voir figure 2).

a) On prend un entier n de la forme $n = 2^s$ et on effectue k rondes de tests, où $k \leq s + 1$.

b) À la première ronde, on teste un mélange des prélèvements de tout le groupe.

c) Si le test s'avère positif, on divise alors le groupe en deux sous-groupes de 2^{s-1} prélèvements et on teste un mélange de chacun d'eux.

d) On poursuit de même jusqu'à la k^e ronde, où on teste individuellement les membres d'un sous-groupe déclaré positif à la ronde précédente. Dans le cas particulier $k = s + 1$, ce sous-groupe n'a plus que deux éléments.

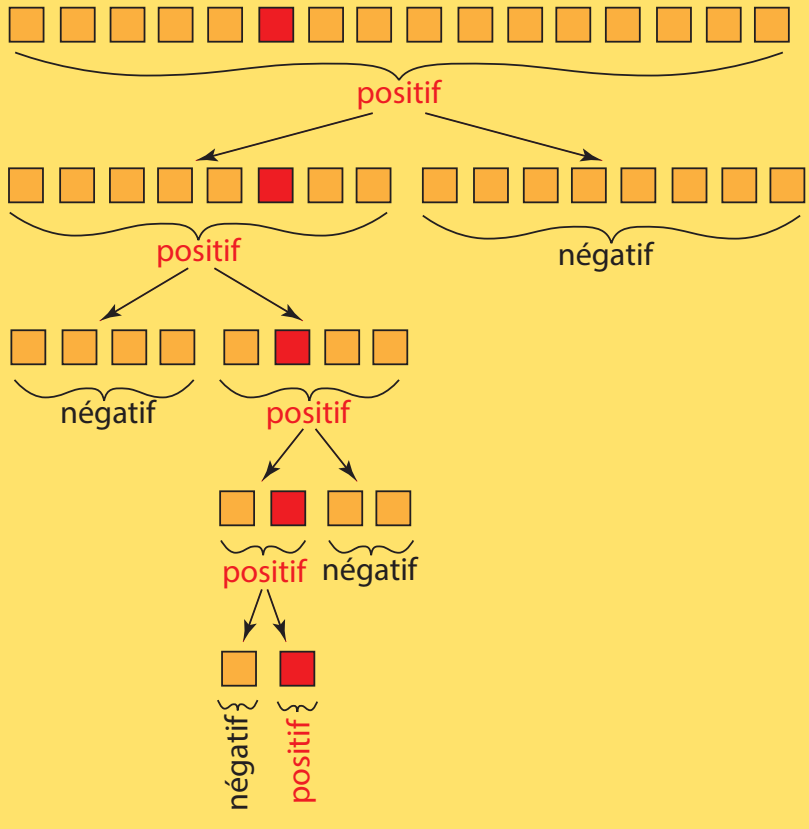

Figure 2.
Représentation graphique d'un algorithme adaptatif de division binaire à 5 rondes appliqué à un échantillon initial de n = 2^4 = 16 prélèvements, dont un seul est infecté.

Si le groupe contient un seul individu infecté, cet algorithme permettra de l'identifier en exactement $s+1=\log_2(n)+1$ rondes. En règle générale, plus le nombre de rondes est élevé, meilleures sont les économies engendrées par cette approche. Mais s'il faut compter 24 à 48 heures pour qu'un test s'avère concluant, les délais de livraison des résultats d'analyse risquent d'être contre-productifs. Notons aussi que cette amélioration requiert des prélèvements biologiques plus importants. Ceci n'est pas vraiment considéré comme un problème et se retrouvera dans tous les algorithmes présentés ci-dessous.

Un algorithme non-adaptatif

Afin de mieux contrôler le temps de réponse, on peut aussi envisager l'emploi de méthodes *non-adaptatives* de dépistage par groupe. Ces protocoles ne comportent qu'une seule ronde, ce qui permet de réaliser tous les tests simultanément. Ils s'avèrent en outre très efficaces pour le dépistage des cas si l'on dispose d'une estimation fiable de la prévalence de la maladie.

Expliquons le concept au moyen de l'exemple suivant, développé par une équipe de chercheurs canado-rwandaise dans le cadre de l'actuelle lutte à la COVID-19. On forme d'abord un échantillon aléatoire de taille $n=3^m$. On établit ensuite une correspondance entre les 3^m individus et les points d'un hypercube discret $\{0, 1, 2\}^m$. Voir la figure 3 pour une illustration dans le cas $m=3$.

L'approche proposée consiste alors à réaliser simultanément $3m$ tests sur des mélanges d'échantillons comportant chacun 3^{m-1} individus. Les mélanges sont toutefois constitués selon des modalités bien précises, soit des tranches de l'hypercube. En effet, si $x_1,\ldots,x_m$ dénotent les axes de coordonnées de l'hypercube, chacun des mélanges correspond alors aux individus situés dans l'hyperplan $x_i=t$, où $i\in\{1,\ldots,m\}$ et $t\in\{0,1,2\}$ est une tranche de 3^{m-1} individus.

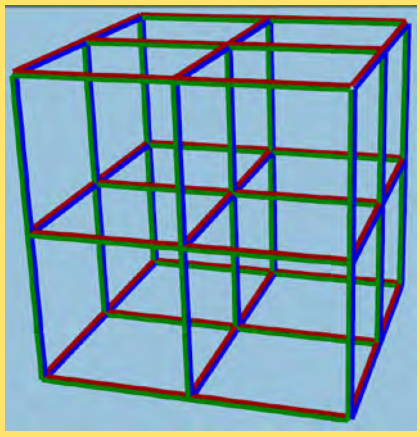

Figure 3.
Hypercube discret $\{0,1,2\}^3$. Chaque point du treillis correspond à un individu d'un échantillon aléatoire de taille $n=3^3=27$. Les $3^2=9$ individus de chacun des 9 mélanges sont situés le long d'une tranche rouge, bleue ou verte.

Lorsque $m=3$, comme dans la figure 3, on effectue alors $3\times3=9$ tests sur des groupes de $3^2=9$ individus. Quand $m=4$, ce qui est la valeur retenue pour l'implantation au Rwanda, on effectue plutôt 12 tests à partir d'un échantillon aléatoire de $n=81$ individus. C'est donc dire que chaque prélèvement est divisé en quatre portions égales et contribue à quatre tests différents. De plus, chaque test est un mélange de 27 échantillons.

Cette approche s'appuie sur une technique de construction de codes correcteurs d'erreurs décrite dans l'encadré[2]. Un de ses grands avantages, c'est que la composition des mélanges est telle que s'il n'y a qu'un seul individu infecté dans l'échantillon, il peut être identifié avec certitude. En revanche, si plus d'une personne est infectée, il faut alors procéder à une seconde ronde de tests.

2. *Voir aussi http://accromath.uqam.ca/accro/wp-content/uploads/2020/02/Codes.pdf pour une introduction aux codes correcteurs d'erreurs.*

Examinons l'exemple canado-rwandais dans le cas $n = 81 = 3^4$. Sachant que le nombre X d'individus infectés dans l'échantillon obéit à une loi binomiale de paramètres $n = 81$ et p, on a

$$\Pr(X \leq 1) = (1-p)^{81} + 81p(1-p)^{80}.$$

Cette stratégie est intéressante si la probabilité que $X \leq 1$ est grande. Mais pour que $\Pr(X \leq 1) \geq 0{,}95$, par exemple, il faut avoir $p \leq 0{,}44\,\%$; autrement dit, la prévalence de la maladie doit être faible. Déjà, à 1 % de prévalence on a $\Pr(X \leq 1) = 0{,}806$ et dans presque 20 % des cas il faudra recourir à une deuxième ronde. Toutefois, toujours pour 1 % de prévalence, on a

$$\Pr(X = 2) = \binom{81}{2} p^2(1-p)^{79} = 0{,}146,$$

d'où $\Pr(X \leq 2) = 0{,}952$.

On peut facilement contrôler les coûts si on se montre astucieux dans la façon de mener la deuxième ronde lorsqu'il y a plus d'un individu infecté. Par exemple, tous les cas correspondant à $X = 2$ satisfont à la propriété suivante (se référer à la figure 4 où on montre, dans le cas $m = 3$, les tranches dont le test est positif) :

(P) Pour toute valeur de $i \in \{1, \ldots, m\}$, il y a au plus deux tranches de la forme $x_i = s$ et $x_i = t$ menant à un test positif, et il existe au moins une valeur de i pour laquelle on a exactement deux tests positifs de cette forme.

Si k est le nombre de valeurs de $i \in \{1, \ldots, m\}$ pour lesquelles on a deux tests positifs de la forme spécifiée en (P), le nombre de tests additionnels requis pour identifier tous les individus infectés est donné dans le tableau ci-contre, dont la construction est décrite dans *La covid en 19 questions*.

k	Nombre de tests additionnels
1	0
2	4
3	8
4	16

Si la propriété (P) n'est pas vérifiée, on doit alors tester tous les individus des tranches ayant conduit à un test positif, ce qui se traduit par au plus 81 tests supplémentaires. Au final, le coût total relatif est donc inférieur ou égal à $100 \times 13{,}06/81 \approx 16{,}1\,\%$.

Perspectives

La recherche d'algorithmes adaptés à la lutte au coronavirus bat son plein. Par exemple, un algorithme a récemment été élaboré et implanté en laboratoire par une équipe de recherche israélienne. Il consiste à réaliser 48 tests sur un même échantillon aléatoire de taille $8 \times 48 = 384$. Chaque prélèvement individuel est divisé en six parts égales. Chacun des tests est fondé sur le mélange de 48 de ces parts, une par individu. Chaque sujet est donc présent dans six tests différents. En laboratoire, un robot a été programmé pour préparer les 48 mélanges. L'algorithme permet d'identifier jusqu'à quatre personnes

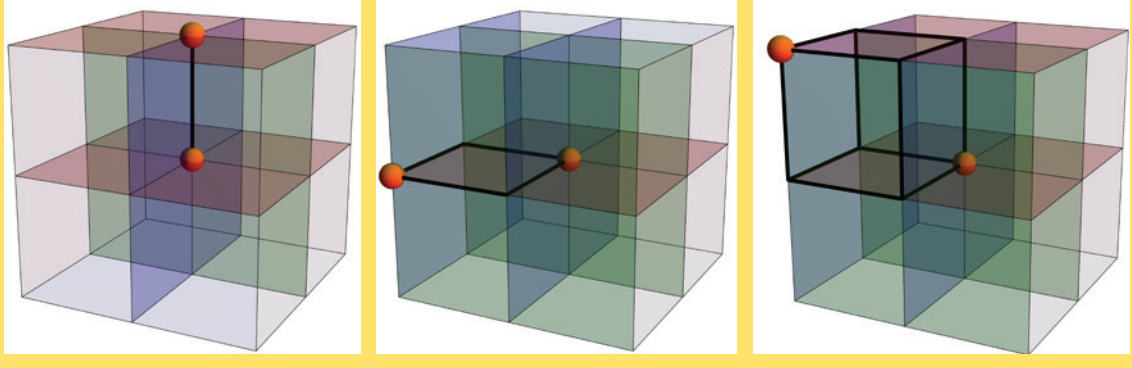

Figure 4.
L'hypercube discret $\{0, 1, 2\}^3$ et les trois possibilités pour deux individus infectés : sur une même droite parallèle à un axe (gauche), aux sommets opposés d'un carré dans un plan (centre) ou aux sommets opposés d'un cube (droite). Dans chaque cas, les tranches menant à un test positif sont identifiées en rouge, vert ou bleu avec les mêmes conventions de couleurs qu'à la figure 3.

infectées. On a ainsi besoin de huit fois moins de tests que d'individus. Encore une fois, plus le pourcentage d'individus infectés est petit, meilleure est la performance de l'algorithme.

Bien sûr, d'autres considérations entrent en jeu dans l'élaboration d'une procédure de test statistique. Pour simplifier, nous avons implicitement supposé ici que le test employé était infaillible. En pratique, même les meilleures procédures peuvent donner de faux positifs ou de faux négatifs. La sensibilité et la spécificité des tests sont des éléments importants à prendre en compte au moment d'en recommander l'implantation, de même que leur faisabilité en termes de temps, de coûts et de complexité des manipulations.

Construire un algorithme non-adaptatif

On veut construire un algorithme non-adaptatif pour tester un groupe de personnes. On va représenter cet algorithme par un tableau à T lignes et n colonnes, ou de manière équivalente par une matrice de dimension $T \times n$ dont toutes les entrées sont des 0 ou des 1. La i^{e} ligne de la matrice représente le i^{e} test. Les entrées égales à 1 de la j^{e} colonne représentent les tests auxquels participe le j^{e} individu. Ainsi, $m_{ij} = 1$ si le j^{e} individu contribue au i^{e} test et 0 sinon.

Construisons un vecteur X de longueur n représentant le groupe qu'on va tester : sa i^{e} coordonnée, x_i, vaut 1 si le i^{e} individu est infecté et 0 sinon. On décide de traiter X comme un mot de longueur n et ses quelques coordonnées égales à 1 comme des *erreurs* dans un mot initial X_0 dont toutes les coordonnées auraient été nulles. Des algorithmes de ***codes correcteurs d'erreurs*** permettent de corriger les erreurs qui se seraient produites dans la transmission de X_0, ce qui revient à identifier quelles sont les coordonnées x_i de X qui valent 1, soit exactement le but recherché. Tel qu'illustré dans l'exemple, un algorithme de correction d'erreurs ne peut corriger qu'un nombre maximum d'erreurs, k, choisi au préalable dans sa construction.

La matrice M est la *matrice du code*. Une propriété suffisante pour que le code puisse corriger k erreurs est que la matrice soit *k-disjointe*. Définissons cette notion. On ne veut pas que si des individus $j_1, \ldots, j_k$ (pas nécessairement distincts) sont infectés, alors un $j_{k+1}{}^{\text{e}}$ individu infecté passe inaperçu. Se donner la colonne j (soit l'individu j), c'est la même chose que se donner le sous-ensemble A_j de $\{1, \ldots, T\}$ de ses entrées égales à 1 (soit l'ensemble des tests auxquels contribue cet individu). Si des individus $j_1, \ldots, j_k$ sont infectés, alors tous les tests correspondant à la réunion $A_{j_1} \cup \cdots \cup A_{j_k}$ seront positifs. Pour que le $j_{k+1}{}^{\text{e}}$ individu infecté ne passe pas inaperçu, il ne faut pas que $A_{j_{k+1}}$ soit inclus dans $A_{j_1} \cup \cdots \cup A_{j_k}$. La matrice M est *k-disjointe* si ceci est vérifié pour tous $j_1, \ldots, j_k$ et tout j_{k+1} distinct de $j_1, \ldots, j_k$. La matrice 12×81 correspondant à l'exemple ci-dessus utilisé au Rwanda serait une matrice 1-disjointe.

Il existe deux grands types de méthodes pour construire des matrices *k-disjointes*. La première est probabiliste : on génère au hasard des matrices dont les entrées sont 1 avec probabilité q et 0 sinon. Pour n, T et k bien choisis, la probabilité que la matrice soit *k-disjointe* est non nulle ; par tâtonnements, on finit donc par générer une matrice *k-disjointe*.

La deuxième méthode, algébrique, est empruntée à la théorie des codes correcteurs d'erreurs de Reed-Solomon. Elle permet de construire des matrices M dont toutes les lignes ont un même nombre m d'entrées non nulles et toutes les colonnes ont un même nombre c d'entrées non nulles. Ainsi, tous les tests se font sur des sous-groupes de taille m et chaque prélèvement individuel est divisé en c parties pour être inclus dans c tests distincts.