Christian Genest; Christiane Rousseau
Skupinový screening

Pokroky matematiky, fyziky a astronomie, Vol. 66 (2021), No. 2, 73–80

Persistent URL: http://dml.cz/dmlcz/148977

Skupinový screening

Christian Genest, Christiane Rousseau

Abstrakt. Skupinový screening je nezbytnou součástí boje proti šíření koronaviru. Jak ale čelit možnému nedostatku činidel a materiálu? Tím, že budeme provádět testy na bázi promíchání vzorků a využijeme přitom matematiku.

V obavách z šíření COVID-19 souhlasí mnoho vědců s tím, že zavedení plánu testování ve velkém měřítku je nezbytné pro zastavení šíření koronaviru. Testy imunologické, sérologické nebo k detekci antigenu vykonané na reprezentativních vzorcích populace by také umožnily odhadnout prevalenci nemoci, posoudit stupeň kolektivní imunity a přizpůsobit opatření k řešení pandemie.

Přístup k vhodným materiálním a personálním zdrojům je předpokladem k tomu, aby bylo zavedení strategie hromadného testování úspěšné. S rostoucí celosvětovou poptávkou je nedostatek činidel nezbytných k vykonání laboratorních testů na obzoru a vzbuzuje obavy orgánů ochrany veřejného zdraví v České republice i ve světě.

Jelikož víme, že většina testů je (naštěstí) negativní, mohlo by využití matematiky přispět ke zlepšení aktuální situace? Ukazuje se, že ano, zejména tím, že budeme provádět skupinové testy na pečlivě sestavených směsích vzorků.

1. Skupinový screening

Představme si, že daná laboratoř obdržela 100 vzorků za účelem testování. Náhodně je rozdělí do pěti skupin po dvaceti. Poté v každé skupině použije polovinu každého z 20 vzorků na vytvoření směsi, kterou následně otestuje.

Jestliže je test vykonaný na dané směsi negativní, můžeme ihned vyvodit, že žádný člen dané skupiny není infikovaný. Jestliže je test pozitivní, provedou se následně jednotlivé testy na druhé polovině každého z 20 vzorků.

Pokud 100 původních vzorků pochází od zdravých jedinců, tento postup to potvrdí vykonáním 5 testů namísto 100. Jestliže je nakažen pouze jeden jedinec, postačí $5 + 20 = 25$ testů k jeho identifikaci. V případě, že jsou nakažené 2 osoby, můžeme je také identifikovat 25 testy, pokud jsou ve stejné skupině; je ale třeba $5 + 20 + 20 = 45$ testů, pokud patří do různých skupin. Stejným způsobem můžeme pokračovat v případě, že jsou nakažené tři nebo více osob.

Článek je převzat z časopisu Accromath se svolením Institut des sciences mathématiques du Québec. Z francouzského originálu *Le dépistage par groupe*, Accromath 15 (2) (2020), 30–35, přeložily Lucie Krejčí a Johanna G. Nešlehová.

Christian Genest, PhD, FRSC, Department of Mathematics and Statistics, McGill University, 805, rue Sherbrooke Ouest, Montréal (Québec), Canada H3A 0B9, e-mail: christian.genest@mcgill.ca, Christiane Rousseau, PhD, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal (Québec), Canada H3C 3J7, e-mail: christiane.rousseau@umontreal.ca

Jak je vidět, skupinový screening umožňuje dosáhnout významných úspor, samozřejmě za předpokladu, že senzitivita a specificita testu nebude ovlivněna směsí vzorků, což předpokládáme v této studii a čemuž je tak velice často i v praxi.

Daná laboratoř by také mohla použít stejnou screeningovou strategii pro testování 10 skupin po 10 vzorcích. Pokud by byl nakažen jeden jedinec, bylo by k jeho identifikaci potřeba jen 20 testů. Na druhou stranu k tomu, aby se dospělo k závěru, že infikován nebyl nikdo, by bylo třeba 10 testů.

Která z těchto dvou strategií je lepší? A existují ještě i jiné, které by byly upřednostňované? Odpověď záleží na *prevalenci* nemoci neboli na podílu nakažených jedinců vzhledem k celkové populaci.

Jak v krátkém sdělení publikovaném v roce 1943 v časopise *The Annals of Mathematical Statistics* uvádí Američan Robert Dorfman [1], bylo skupinové testování ve své nejzákladnější formě použito již během druhé světové války, a to k detekci případů syfilidy mezi odvedenci. Tento přístup se prosadil a má dnes mnoho variant, které se používají především v Severní Americe k testování na přítomnost viru HIV, chřipky nebo západonilského viru.

2. Optimalizace algoritmu

Dorfman ukázal, jak určit optimální velikost skupiny na základě prevalence $p \in [0, 1]$ nemoci. Označme velikost skupiny $n \geqq 2$ a předpokládejme, že její členové tvoří náhodný výběr, tedy že jsou vzájemně nezávislí a představují reprezentativní vzorek z populace. Podmínka vzájemné nezávislosti je porušena například v případě, že členové výběru patří do stejné rodiny či školní třídy. Provedeme jeden test na směsi všech vzorků a popřípadě následné testy na jednotlivých vzorcích.

Pokud X označuje neznámý počet infikovaných osob ve skupině, pak má tato náhodná veličina binomické rozdělení[1] s parametry n a p, a tudíž

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

jelikož každý jedinec má pravděpodobnost $1 - p$, že není nakažený. Z toho vyplývá, že

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

Pokud $X = 0$, postačí pouze $N = 1$ test. Pokud je však $X > 0$, bude zapotřebí provést $N = n + 1$ testů. Průměrný počet provedených testů, tj. střední hodnota náhodné veličiny N, se rovná

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

Tato funkce je rostoucí v proměnné p. Pokud $p = 0$, máme $\mathrm{E}(N) = 1$. To je zřejmé i z toho, že nemocný není nikdy nikdo, což se vždy potvrdí jediným testem. Jestliže $p = 1$, máme $\mathrm{E}(N) = n + 1$, protože nakažení jsou všichni a tudíž první test bude nutně pozitivní.

Pro jakoukoli hodnotu $p \in [0, 1]$ je možné určit *relativní náklady* spojené s použitím skupinového screeningu tím, že budeme studovat chování poměru

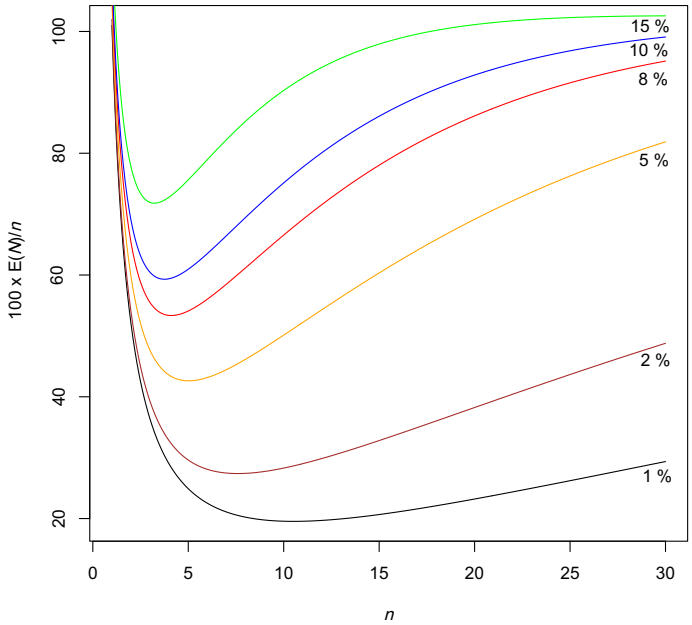
$$\mathrm{E}(N)/n = 1 + 1/n - (1 - p)^n$$

[1] Jedná se o aproximaci, která je opodstatněná, pokud je populace mnohem početnější než jednotlivé skupiny.

jako funkci proměnné n. Čím je $\mathrm{E}(N)/n$ menší, tím víc se vyplatí používat testy skupinové, samozřejmě za předpokladu, že poměr je menší než 1. Když $p = 0$, platí $\mathrm{E}(N)/n = 1/n$, a tudíž se vyplatí volit n co největší. Když $p = 1$, máme vždy

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

neboť skupinový test je vždy pozitivní, čímž ztrácí význam.



Obr. 1. Křivka $100 \times \mathrm{E}(N)/n$ jako funkce proměnné n pro různé hodnoty p

Při pevné hodnotě p představuje funkce $100 \times \mathrm{E}(N)/n$ průměrné procento testů provedených na skupině o velikosti n. Obrázek 1 ukazuje graf této funkce pro různé hodnoty p, které odpovídají prevalenci 1 %, 2 %, 5 %, 8 %, 10 % a 15 %. Jak je vidět, optimální velikost směsí, která odpovídá minimu křivky, se mění podle podílu p infikovaných jedinců v populaci. Tabulka 1, kterou uvádí Dorfman, udává optimální velikost n pro některé hodnoty p.

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

Tab. 1. Optimální volba n a relativní náklady pro některé hodnoty procenta infikovaných p

3. Zobecnění

Výše popsaný testovací protokol je příkladem adaptivního dvoukrokového algoritmu. Nazýváme jej adaptivní, protože výběr (a tedy počet) testů, které mají být provedeny ve druhém kole testování, záleží na výsledku testu provedeného v prvním kole. Výkon algoritmu tohoto typu lze zlepšit několika způsoby, například tím, že zvýšíme počet fází.

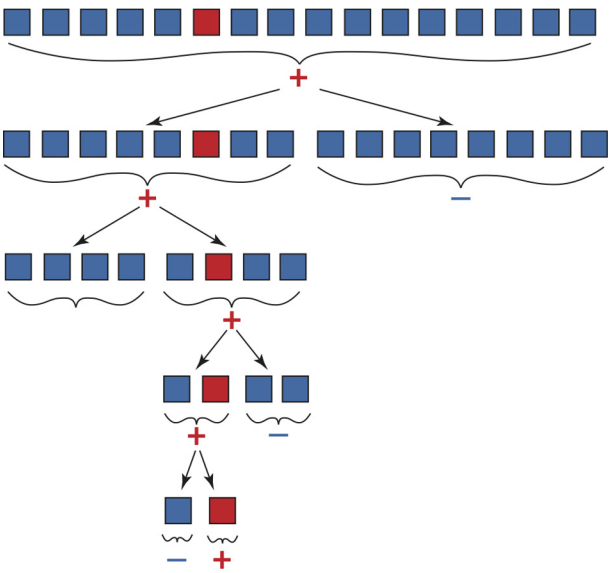
Zde je klasický algoritmus, který budeme nazývat *algoritmus binárního dělení* a který má jisté optimální vlastnosti (viz obrázek 2).

a) Zvolíme celé číslo n ve tvaru $n = 2^s$ a provedeme k fází testů, kde $k \leqq s + 1$.

b) V prvním kole otestujeme směs vzorků celé skupiny.

c) Jestliže je test pozitivní, rozdělíme skupinu na dvě podskupiny o 2^{s-1} vzorcích a otestujeme směs vzorků obou podskupin.

d) Takto budeme pokračovat až do etapy k, ve které jednotlivě otestujeme všechny členy podskupiny prohlášené za pozitivní v předchozím kole. V konkrétním případě $k = s + 1$ má tato podskupina pouze dva členy.



Obr. 2. Grafické znázornění adaptivního pětikrokového algoritmu binárního dělení, aplikovaného na skupinu o $n = 2^4 = 16$ jedincích, z nichž pouze jeden je nakažený

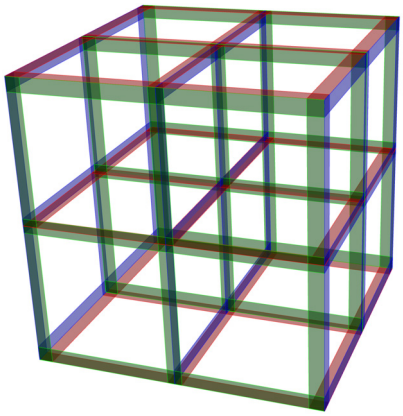
Pokud je ve skupině pouze jeden infikovaný jedinec, najde jej tento algoritmus přesně v $s + 1 = \log_2(n) + 1$ krocích. Obecně platí, že čím vyšší je počet kroků, tím větší jsou úspory vzniklé tímto přístupem. Pokud ovšem vyhodnocení testu trvá 24 až 48 hodin, může být doba získání výsledků tímto způsobem kontraproduktivní. Všimněme si také, že vylepšení výše uvedeným algoritmem i všemi níže uvedenými postupy vyžaduje objemné biologické vzorky; budeme zde předpokládat, že jsou k dispozici.

4. Neadaptivní algoritmus

Za účelem lepší kontroly doby získání výsledků lze také zvážit použití neadaptivních metod při provádění skupinového screeningu. Tyto protokoly mají pouze jednu etapu, což umožňuje provádět všechny testy současně. Jsou také velmi efektivní při detekci případů, pokud máme k dispozici spolehlivý odhad prevalence onemocnění.

Pojďme si vysvětlit tento koncept na následujícím příkladu, který vyvinul rwandský tým vědců jako součást současného boje proti COVID-19 [2]. Nejprve náhodně vybereme $n = 3^m$ jedinců. Potom vytvoříme korespondenci mezi 3^m jednotlivci a body diskrétní hyperkrychle $\{0, 1, 2\}^m$, viz obrázek 3 pro ilustraci v případě $m = 3$.

Navrhovaný postup spočívá v současném provádění $3m$ testů na směsích, z nichž každá obsahuje vzorky 3^{m-1} jedinců. Nicméně směsi jsou vyrobeny podle přesných návodů, a to podle řezů hyperkrychle. Pokud označíme souřadnicové osy hyperkrychle $x_1, \ldots, x_m$, pak každá směs odpovídá 3^{m-1} jedincům umístěným v řezu hypekrychle nadrovinou $x_i = t$, kde $i \in \{1, \ldots, m\}$ a $t \in \{0, 1, 2\}$.



Obr. 3. Diskrétní hyperkrychle $\{0, 1, 2\}^3$. Každý bod mřížky odpovídá jednomu jednotlivci v náhodném výběru $n = 3^3 = 27$ jedinců. Skupiny $3^2 = 9$ jednotlivců, z jejichž vzorků se vytváří jedna z 9 testovacích směsí, určují červené, modré a zelené řezy

Pokud $m = 3$ jako na obrázku 3, pak provedeme $3 \times 3 = 9$ testů na skupinách po $3^2 = 9$ vzorcích. Pokud $m = 4$, což je hodnota zvolená pro využití daného postupu ve Rwandě, vykonáme 12 testů na základě náhodného výběru $n = 81$ jednotlivců. To znamená, že každý vzorek je rozdělen na čtyři stejné části a je využit ve čtyřech různých testech. Každý test je navíc proveden na směsi 27 vzorků.

Tento přístup je založen na technice konstrukce samoopravných kódů, popsané v rámečku. Jednou z jeho velkých výhod je, že pokud je v náhodném výběru pouze jeden infikovaný jedinec, bude s jistotou identifikován. Pokud je však nakažena více než jedna osoba, je třeba provést druhé kolo testování.

Prozkoumejme rwandský příklad v případě náhodného výběru o velikosti $n = 81 = 3^4$. Protože má celkový počet X infikovaných jedinců ve skupině binomické rozdělení s parametry $n = 81$ a p, platí

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

Vytváření neadaptivního algoritmu

Pro testování skupiny lidí chceme vytvořit neadaptivní algoritmus. Tento algoritmus znázorníme tabulkou o T řádcích a n sloupcích, nebo ekvivalentně maticí o rozměrech $T \times n$, jejíž všechny prvky jsou buď 0, nebo 1. V této matici představuje i-tý řádek i-tý test. Prvky rovné 1 v j-tém sloupci představují testy, kterých se účastní j-tý jedinec. To znamená, že $m_{ij} = 1$, pokud se j-tý jedinec podílí na i-tém testu, a $m_{ij} = 0$ v opačném případě.

Vytvořme vektor X o n složkách představující skupinu, kterou budeme testovat: jeho i-tá složka x_i se rovná 1, jestliže je i-tý jednotlivec infikovaný, a 0 v opačném případě. Považujme X za slovo délky n a jeho složky rovné 1 za *chyby* v počátečním slově X_0, jehož všechny složky jsou nulové. Algoritmy samoopravných kódů umožňují opravit chyby, ke kterým mohlo dojít při přenosu X_0, což odpovídá identifikaci složek x_i vektoru X, které se rovnají 1, a to je přesně naším cílem. Obecně platí, že algoritmus samoopravného kódu může opravit nanejvýš k chyb, kde počet k je předem zvolený při konstrukci kódu.

Matice M je *maticí kódu*. Postačující vlastnost, aby kód opravil k chyb, je ta, že matice je k-disjunktní. Definujme tento pojem. Následující situaci se chceme vyhnout: jestliže jednotlivci $j_1, \ldots, j_k$ (ne nutně odlišní) jsou infikovaní, pak nakažený jedinec j_{k+1} nebude odhalen. Vybrat sloupec j (tj. jednotlivce j) je totéž jako vybrat z $\{1, \ldots, T\}$ podmnožinu A_j odpovídající řádkům, které v j-tém sloupci obsahují prvky 1 (tj. indexy všech testů, na kterých se tento jedinec podílí). Pokud jsou jednotlivci $j_1, \ldots, j_k$ infikovaní, potom budou pozitivní všechny testy odpovídající sjednocení $A_{j_1} \cup \ldots \cup A_{j_k}$. Aby infikovaná osoba j_{k+1} nezůstala bez povšimnutí, $A_{j_{k+1}}$ nesmí být podmnožinou $A_{j_1} \cup \ldots \cup A_{j_k}$. Matice M je *k-disjunktní*, jestliže toto platí pro všechna $j_1, \ldots, j_k$ a všechna j_{k+1} odlišná od $j_1, \ldots, j_k$. Matice 12×81 odpovídající výše uvedenému příkladu použitému ve Rwandě by byla maticí 1-disjunktní.

Existují dva hlavní typy metod pro konstrukci k-disjunktních matic. První je pravěpodobnostní: náhodně generujeme matice, jejichž prvky jsou 1 s pravděpodobností q a 0 s pravděpodobností $1 - q$. Pokud jsou n, T a k vhodně zvolené, pravděpodobnost, že matice je k-disjunktní, není nulová; metodou pokus–omyl proto nakonec vygenerujeme k-disjunktní matici.

Druhá, algebraická metoda je převzata z teorie Reedových–Solomonových samoopravných kódů. Tato teorie umožňuje vytvářet matice M, ve kterých všechny řádky mají stejný počet m nenulových prvků a všechny sloupce mají stejný počet c nenulových prvků. Všechny testy se tedy provádějí na podskupinách o velikosti m a každý jednotlivý vzorek je rozdělen na c částí, které jsou zahrnuty do c různých testů.

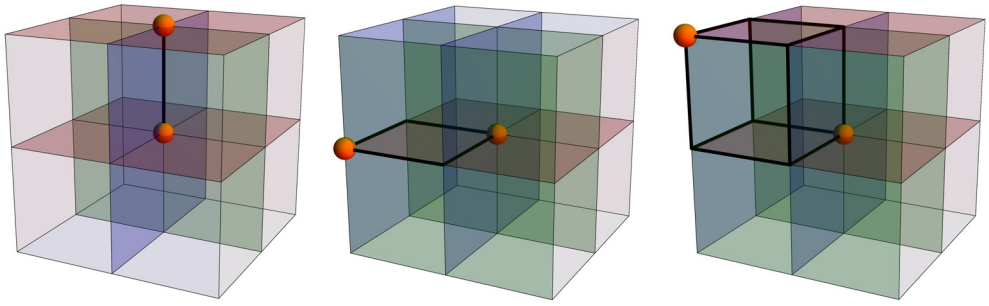
Tato strategie je zajímavá, pokud je velká pravděpodobnost, že $X \leqq 1$. Abychom měli například $P(X \leqq 1) \geqq 0{,}95$, musí ovšem platit, že $p \leqq 0{,}44$ %; jinak řečeno, prevalence nemoci musí být nízká. Již při prevalenci 1 % máme $P(X \leqq 1) = 0{,}806$ a téměř ve 20 % případů bude nutné využít druhé kolo. Nicméně pro prevalenci 1 % platí

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

a tedy $P(X \leqq 2) = 0{,}952$.

Náklady lze snadno udržet pod kontrolou, pokud chytře zvolíme způsob provedení druhého kola v případě, že je infikovaný více než jeden jedinec. Například všechny situace, ve kterých $X = 2$, splňují následující podmínku (viz obrázek 4, kde jsou v případě $m = 3$ znázorněny řezy vedoucí k pozitivním směsím):

> (P) Pro každou hodnotu $i \in \{1, \ldots, m\}$ existují maximálně dva řezy typu $x_i = s$ a $x_i = t$ vedoucí k pozitivnímu testu. Zároveň existuje alespoň jedna hodnota i, pro kterou máme přesně dva řezy tohoto typu vedoucí k pozitivnímu testu.



Obr. 4. Diskrétní hyperkrychle $\{0, 1, 2\}^3$ a tři možnosti, kdy jsou ve výběru přesně dva infikovaní jedinci. Ti se nachází buď na jedné přímce rovnoběžné s jednou ze souřadnicových os (vlevo), nebo v protilehlých vrcholech čtverce v rovině (uprostřed), nebo v protilehlých vrcholech krychle (vpravo). V každém z těchto případů znázorňují zvýrazněné řezy směsi vzorků, které mají pozitivní test. Při vyznačení řezů jsou použity stejné barvevné konvence jako na obrázku 3

Pokud k je počet hodnot $i \in \{1, \ldots, m\}$, pro které máme dva pozitivní testy typu upřesněného v (P), udává počet dalších testů potřebných k identifikaci všech infikovaných jedinců tabulka 2.

k	Počet dodatečných testů
1	0
2	4
3	8
4	16

Tab. 2. Počet dodatečně provedených testů potřebných k identifikaci všech infikovaných jedinců v závislosti na celkovém počtu k hodnot $i \in \{1, \ldots, m\}$, pro které existují dva pozitivní testy mající formu upřesněnou v podmínce (P)

Pokud podmínka (P) neplatí, musí být otestováni všichni jedinci nacházející se v řezech, které vedly ke směsím s pozitivním testem. To znamená, že bude zapotřebí maximálně 81 dalších testů. Výčtem všech možných případů můžeme spočítat, že střední hodnota relativních nákladů se rovná $100 \times 17/81 \approx 21$ %.

5. Perspektivy

Hledání algoritmů uzpůsobených k boji proti koronaviru je v plném proudu. Jeden izraelský výzkumný tým například vyvinul a v laboratoři implementoval následující algoritmus, který spočívá v provedení 48 testů na náhodném výběru o velikosti $8 \times 48 = 384$. Každý jednotlivý vzorek je rozdělen na šest stejných částí. Každý z testů je proveden na směsi 48 těchto částí, z nichž každá patří jiné osobě. Každý vzorek každého jedince je proto přítomen v šesti různých testovacích směsích. Na přípravu 48 směsí byl v laboratoři naprogramován robot. Tento algoritmus dokáže jednoznačně identifikovat až 4 nakažené osoby. Potřebujeme tedy osmkrát méně testů než jednotlivců. Zopakujme tedy ještě jednou, že čím menší je procento infikovaných jedinců, tím lepší je výkonnost algoritmu.

Při vývoji algoritmu pro testování bereme samozřejmě v potaz i další skutečnosti. Pro zjednodušení jsme zde implicitně předpokládali, že použitý test je neomylný. V praxi mohou i ty nejlepší metody vést k falešně pozitivním nebo falešně negativním výsledkům. Senzitivita a specificita testů jsou důležitými prvky, které je třeba vzít v úvahu při doporučení jejich implementace; důležitou roli hraje rovněž jejich proveditelnost z hlediska času, nákladů a složitosti manipulace.

Literatura

[1] Dorfman, R.: *The detection of defective members of large populations.* Ann. Math. Statist. 14 (1943), 436–440.

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