Smart Pooling: AI-powered COVID-19 testing

María Escobar[1], Guillaume Jeanneret[1], Laura Bravo-Sánchez[1], Angela Castillo[1], Catalina Gómez[1], Diego Valderrama[1], Maria F. Roa[1], Julián Martínez[1], Jorge Madrid-Wolff[2], Martha Cepeda[3], Marcela Guevara-Suarez[4], Olga L. Sarmiento[5], Andrés L. Medaglia[6], Manu Forero-Shelton[7], Mauricio Velasco[8], Juan Manuel Pedraza-Leal[7], Silvia Restrepo[4], Pablo Arbelaez[1*]

1 Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Bogotá, 111711, Colombia
2 Laboratory of Applied Photonics Devices, École Polytechnique Fédérale de Lausanne, Lausanne, CH-1015, Switzerland
3 Faculty of Sciences, Universidad de los Andes, Bogotá, 111711, Colombia
4 Department of Chemical Engineering, Universidad de los Andes, Bogotá, 111711, Colombia
5 Faculty of Medicine, Universidad de los Andes, Bogotá, 111711, Colombia
6 Department of Industrial Engineering, Universidad de los Andes, Bogotá, 111711, Colombia
7 Department of Physics, Universidad de los Andes, Bogotá, 111711, Colombia
8 Department of Mathematics, Universidad de los Andes, Bogotá, 111711, Colombia

* mc.escobar11@uniandes.edu.co

Abstract

Massive molecular testing for COVID-19 has been pointed as fundamental to moderate the spread of the disease. Pooling methods can enhance testing efficiency, but they are viable only at very low prevalences of the disease. We propose Smart Pooling, a machine learning method that uses sociodemographic data from patients to increase the efficiency of pooled molecular testing for COVID-19 by arranging samples into all-negative pools. We show efficiency gains of 42% with respect to individual testing at disease prevalence of up to 25%, a regime in which two-step pooling offers marginal efficiency gains. Additionally, we calculate the possible efficiency gains of one- and two-dimensional two-step pooling strategies and present the optimal strategies for disease prevalences up to 25%. We discuss practical limitations to conduct pooling in the laboratory.

1 Introduction

COVID-19 is an acute respiratory illness caused by the novel coronavirus SARS-CoV-2 [1, 2]. The disease has rapidly spread to most countries and territories and has caused 7.5 million confirmed infections and 425,000 deaths as of June 13th 2020 [3]. Countries are rushing to implement massive testing to identify and isolate carriers to tackle the spread of the disease. Nevertheless, massive testing is costly and has increased the demand for scarce reagents. However, there is still a global need to make testing more accessible to larger populations. Thus, strategies to enhance the efficiency of testing, that is, the number of people that can be tested with the same amount of test kits, are urgent.

Pooling methods, which were initially proposed by Dorfman to diagnose syphilis among the US military during World War II [4], allow to test several patients with fewer reagents by combining their samples in a single test tube. Different pooling strategies, such as two-step pooling [4] and matrix pooling [5], offer more significant efficiency gains for different combinations of sensitivity of the test and prevalence of the disease. Pooling has not only been used for diagnosis but also in genetic sequencing [6].

Empirical tests have confirmed the ability of pooled analysis to reliably detect SARS-CoV-2 in a pool comprising one positive specimen and up to 31 negative specimens [7, 8]. Additionally, preliminary trials show that pools of 10 specimens can increase efficiency without strongly compromising sensitivity [9, 10]. Given sufficient sensitivity for the test, pooling methods are effective when the prevalence of the disease is low so that the probability of all samples in a pool being negative is high, but fails as the prevalence increases, as shown in Figure 1a.

Here we present Smart Pooling, a machine learning method that enhances the efficiency of pooling testing strategies. Smart Pooling exploits demographic information of samples to estimate their probability of testing positive for COVID-19. As Figure 1b shows, our method uses these probabilities to arrange all samples into pools that maximize the probability of yielding a negative result within the most number of pools. That is, we group samples such that positive samples are excluded from the pooling process and are evaluated in single tests, thus reducing the number of tests used for the same amount of samples.

Machine learning and regression models have been used to classify the novel pathogen from its genetic sequencing [11], to support diagnosis from CT scans [12], to assist clinical prognosis of patients [13], and to forecast the evolution of the pandemic [14]. Regression models trained with patient-reported symptoms and laboratory test results have been used to predict infection from symptomatology [15]. These strategies risk compromising sensitivity and confidence. In contrast, Smart Pooling does not seek to replace current molecular testing but assist it by improving its efficiency.

Complementary information from samples, such as microscopic inspections, has been effectively used to guide pooling methods in detecting malaria [16]. In the

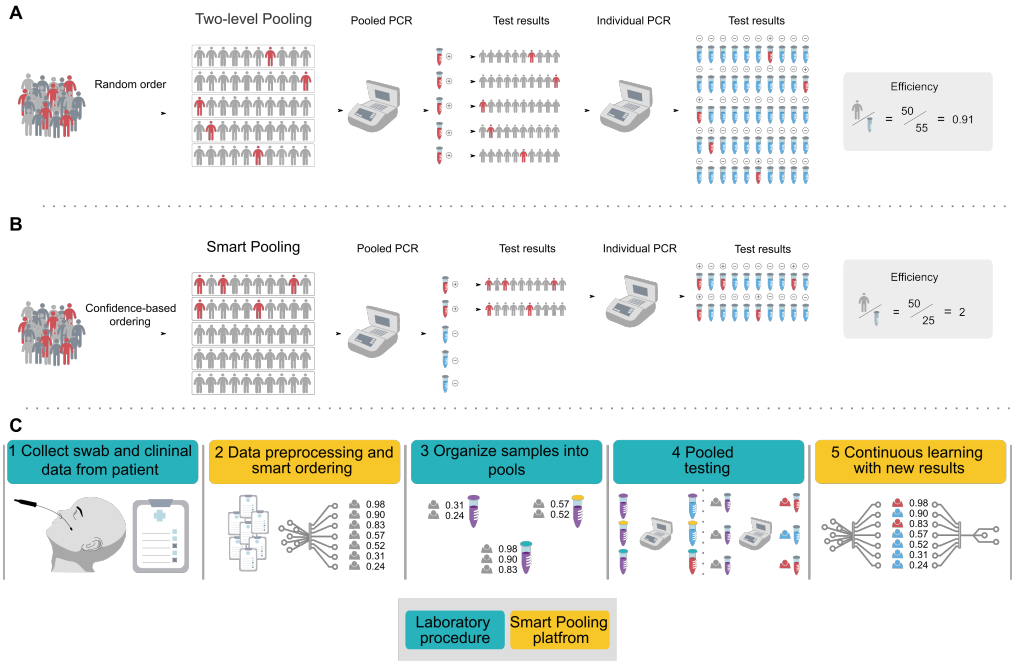

Fig 1. Smart Pooling makes pooling methods efficient even at high prevalence of the disease. a. In standard two-step pooling methods, samples are pooled randomly. When the outcome of a pooled test is negative, all the samples in it are labeled as negative. When the outcome is positive, all samples are tested individually. As prevalence increases, the efficiency of pooling without a priori information drops rapidly and makes the strategy unviable, mainly because the probability of having at least one positive sample in a pool increases. **b.** Smart Pooling tackles this problem by arranging samples into pools that maximize the probability of being all-negative. **c.** *The Smart Pooling pipeline.* Samples and data are collected from patients. The Smart Pooling model processes these data and returns an arrangement with the probability that each sample tests positive. In the lab, samples are pooled based on this arrangement. Subsequently, samples from positive pools are tested individually. Finally, the diagnostic outcome of each sample is fed to the Smart Pooling platform. This process enlarges the dataset and allows for continuous learning.

epidemiological context of the COVID-19 pandemic, testing laboratories may have clinical and demographic data for each individual. These data could be exploited to estimate the probability of a patient yielding a positive result [17]. Thus, an informed guess could be made to exclude a sample from a pool.

Smart Pooling is easily adaptable to any pooling procedure. To use Smart Pooling as a tool for the desired pooling strategy, we propose a five-step pipeline between the laboratory procedures and the Smart Pooling analysis. Figure 1c illustrates this process. First, the laboratory acquires samples and demographic metadata from patients. Secondly, the Smart Pooling platform processes metadata to provide an ordering of the patients into pools. Thirdly, samples are pooled according to this ordering in the laboratory. Then, molecular tests are run on the ordered pools until there is a diagnosis for each sample. Lastly, the laboratory feeds the results of the tested samples into the Smart Pooling platform to continuously improve the model.

2 Results and Discussion

2.1 Smart Pooling increases testing efficiency by excluding positive samples from pools

We identify that using complementary information to arrange pools can improve the efficiency of testing. We do this by training a machine learning algorithm to predict the probability that a sample will test positive for COVID-19 based on sociodemographic data. Testing efficiency increases by using individual testing on high probability samples and arranging the remaining samples into pools simulating a two-step pooling protocol. Figure 2 shows that for disease prevalences ranging from 0% to 25%, Smart Pooling efficiency outperforms the simulated efficiency obtained with Dorfsman's two-step pooling and individual testing.

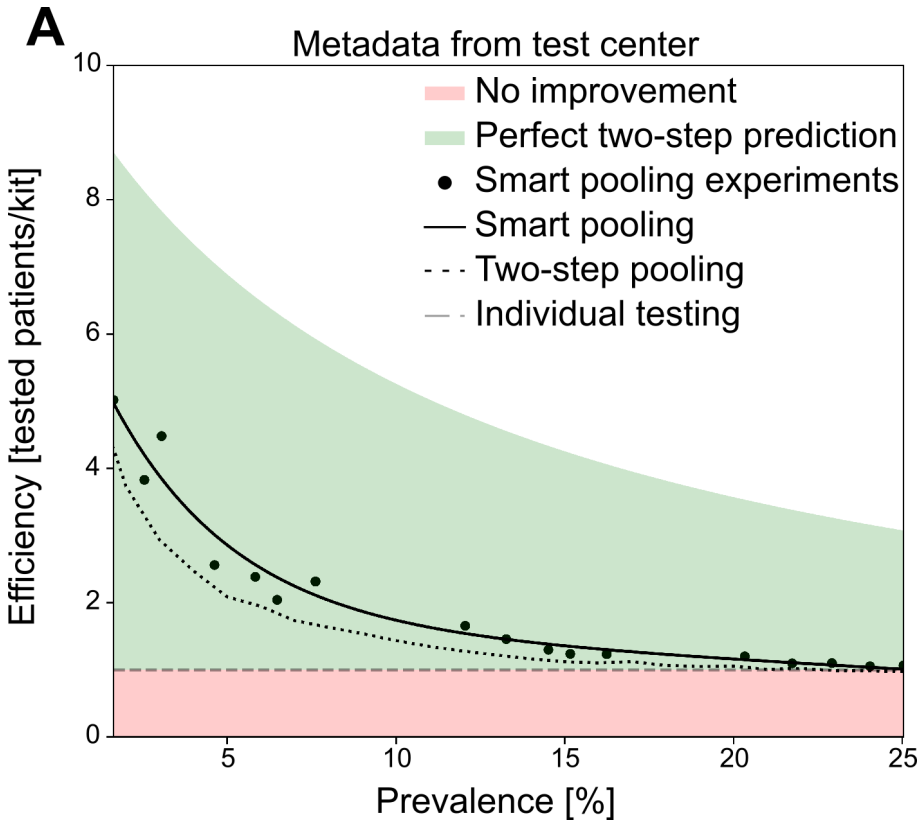

Fig 2. Efficiency of Smart Pooling trained on different data. Smart pooling achieves higher efficiencies than two-step pooling for prevalence of the disease of up to 25%.

For our experiments, we construct a dataset that comprises the results of qPCR tests of samples from the city of Bogotá, Colombia. We gather data at different granularity levels reflecting real data availability and organize it into the *Test Center Dataset*. This data is restricted to information of the test center and the date the samples were taken. We conduct qPCR on DNA extracted from lower or upper respiratory tract samples,

following standard diagnostic protocols for COVID-19 [18, 19]. We perform DNA extraction and amplification individually for each sample in the dataset.

Figure 2 shows the efficiency gains of Smart Pooling. Our computational experiments show that efficiency gains are obtained for all simulated prevalences up to 25% when performing Smart Pooling. For instance, with Smart Pooling at a prevalence of 10% and an efficiency around 2, the estimated number of patients that could be tested with the same number of test kits is doubled compared to individual testing. These results on the *Test Center Dataset* show that Smart Pooling does not depend on the availability of rich complementary data such as patient-level data.

Figure 3 depicts the visualized predictions from our machine learning models trained with test center metadata. We rank samples according to the confidence of the model for predicting a sample as positive. Compared to the random ordered samples, Smart Pooling's predictions of the test center data set most of the positive samples at the top of the arrangement. This figure illustrates the working principle of Smart Pooling: it enhances efficiency by artificially reducing the incidence in the samples sent to pools, by sending the samples most likely to test positive to individual testing.

The machine learning model is inferring the probability that a sample results positive from the metadata. For the *Test Center Dataset*, the machine learning model could be exploiting underlying correlations in the samples. Samples in the dataset were acquired during strict measures limiting mobility in the city of Bogota. It is likely that people tested at the same center share epidemiological factors, such as visiting the same markets, sharing public transport, or being in the same hospital. The model could also be learning the different probability distributions of samples being positive in different parts of the city.

Smart Pooling can be used with multiple pooling strategies and improves efficiency regardless of the strategy. Figure S1 shows the effect of using Smart Pooling with a fixed pool size of 10 and an adaptive pooling strategy based on the optimal strategies previously stated. Both of these alternatives are more efficient than two-step pooling and individual testing when coupled with Smart Pooling. Our simulations show that using adaptive pooling strategies offers efficiency gains than fixed-size pooling for higher prevalences ($p > 15\%$). Ultimately, the pooling strategy used in practice should be determined by the resources available at the testing laboratory.

3 Methods

3.1 Pooling strategies

3.1.1 *Two-step pooling protocols*

By a two-step pooling protocol, we mean a procedure that, given a set of samples, is combined into pools using specific criteria. Once the pools are defined, we carry out two steps. In the first step, we test each pool for disease using qPCR. There are two possible outcomes: (i) The pooled sample is *negative*, which implies that *all* individual

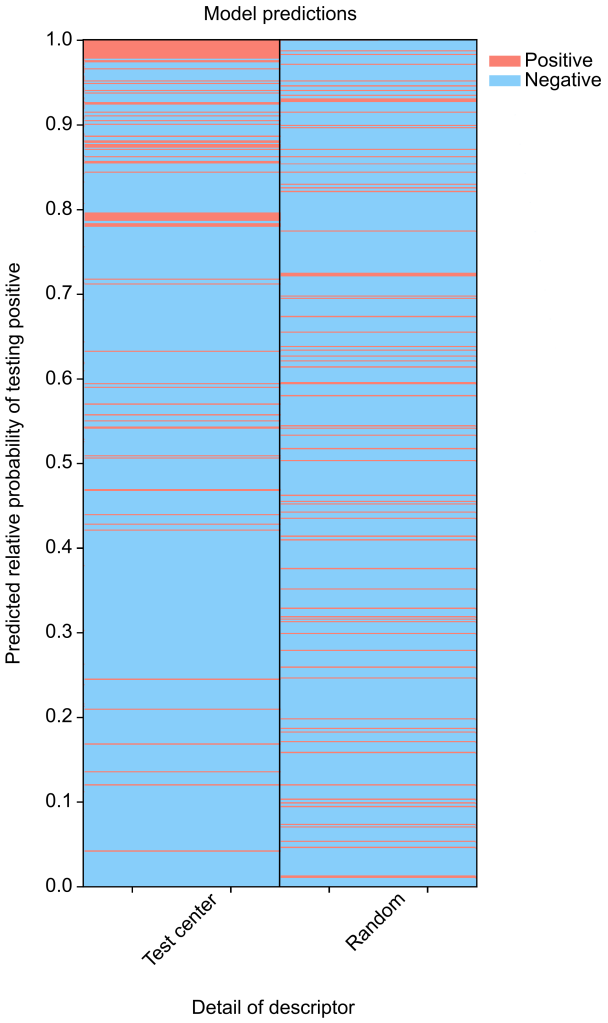

Fig 3. Visualization of Smart Pooling model predictions trained with different descriptors. Samples are arranged by probability of testing positive based on the predictions made by each model compared to a random arrangement.

samples within the pool are negative; or (ii) The pooled sample is *positive*, which implies that *at least one* individual sample within the pool is positive. If the test result of the first step is positive, then it is necessary to proceed to the second step. In the second step, each sample is individually tested, thereby finding the infected samples. More importantly, note that if the result of the first step is negative, by performing a single pooled test, we save all the PCR kits required to test the samples individually (except the one used for the pooled test). Crucially, as we will show later, this gain is obtained without any loss in either test sensitivity or specificity.

We explored the following two kinds of pooling protocols:

1. Dorfman's pooling protocols [4]: Given m samples, we make a single pool with all of them. We denote this protocol by S_m

2. Matrix pooling protocols [5]: Given a collection of $m \times n$ samples, we place them

into a rectangular $m \times n$ array. We create pools by combining samples along the rows and along the columns of this array (for a total of $m + n$ pools). In the second phase, we test each sample at the intersection of the positive columns and rows individually. We denote this protocol by $S_{m\times n}$.

3.1.2 Quantifying pooling efficiency

Pooling efficiency will be our main tool to quantitatively compare different pooling protocols, as it is defined as the number of patients tested per detection kit consumed. Efficiency depends on the prevalence p of the sample (defined as the probability that an individual in the population is ill). Assuming independence among patients, pooling efficiency can be easily computed analytically [4, 5]:

1. For Dorfman's pooling S_m the efficiency is given by

$$E(p) = \frac{1}{\frac{1}{m} + 1 - (1-p)^2}$$

2. For the matrix pooling protocol $S_{m\times n}$ the efficiency is given by

$$E(p) = \frac{1}{\frac{1}{m} + \frac{1}{n} + p + (1-p)(1-(1-p)^{m-1})(1-(1-p)^{n-1})}$$

The efficiency functions above show the key property behind pooling in general, and Smart Pooling in particular. At low prevalences, the efficiency can be considerably greater than one. As the prevalence increases, $E(p)$ decreases, and the efficiency gains become negligible after prevalences around 30%. Figure 4 shows graphically these efficiency gains (and how they vanish) for two-step pooling.

3.1.3 Optimal pooling strategies

What are the best pooling protocols that maximize efficiency for a given prevalence p, assuming a fixed bound c on pool size? The experimental setup unavoidably constrains the maximum pool size. In the context of COVID-19 we will focus on the cases $c = 5$ and $c = 10$ since these are the most useful in practice (see Section S1.1 for details). Finding the optimal pooling strategies for a given prevalence equals finding the pooling protocol of maximum efficiency by comparing the values of $E(p)$ for the different protocols. Figure 5 shows the efficiency curves of the best pooling protocols of the form S_j and $S_{m\times n}$ with maximum pool size $c \leq 10$ and $c \leq 5$. The optimal two-stage protocols exist (i.e., their efficiencies are the highest among all two-stage pooling protocols, not only among those of the form S_m and $S_{m\times n}$). More precisely, table 1 shows the optimal protocols and their respective intervals of optimality when $c = 10$ and $c = 5$ respectively.

Even while using optimal protocols, it is clear that the efficiency gains quickly disappear as the prevalence increases (effectively vanishing when $p \geq 30.66\%$).

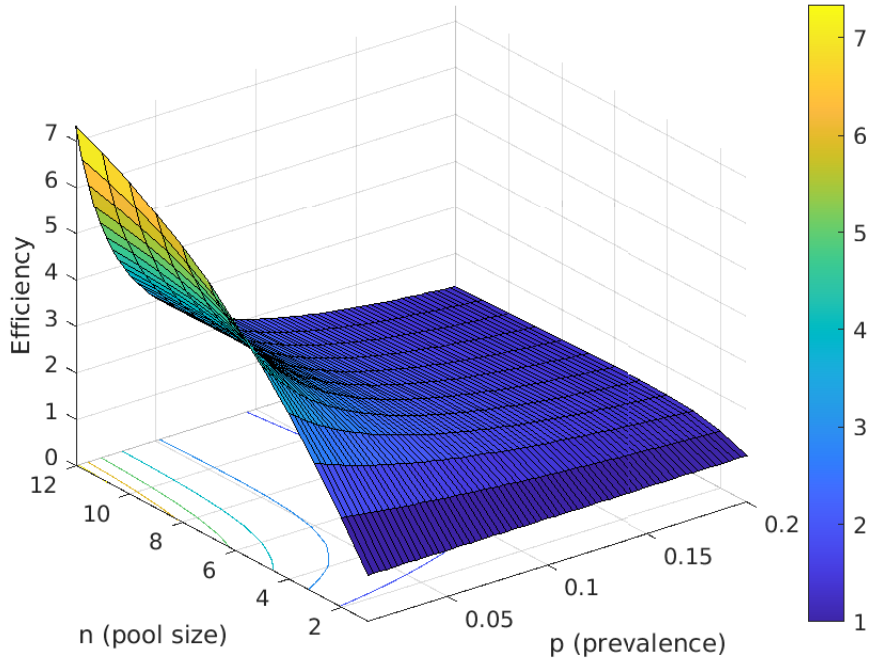

Fig 4. Efficiency of two-step Dorfman's pooling as a function of prevalence p and pool size n

3.2 Test Center Dataset

The data corresponds to the molecular tests conducted by Universidad de los Andes for the Health Authority of the city of Bogotá, Colombia. We organized the data according to the test center that collected the sample. Additionally, metadata from the test centers, such as their location and name, were available. Samples were collected from April 6th to May 25th 2020, with a total of 7162 samples from 101 test centers. To construct the dataset, we tested samples individually following the Berlin Protocol [18] (for dates before April 18th) and the protocol for the U-TOP^{TM} COVID-19 Detection kit from Seasun Biomaterials [19].

3.3 Training of the predictors

3.3.1 *Dataset division*

To experimentally validate the performance of our proposed models and descriptors, we divided the data into a training split containing all the samples until May 7th, and a held-out test split for evaluating the performance of the model with the remaining samples.

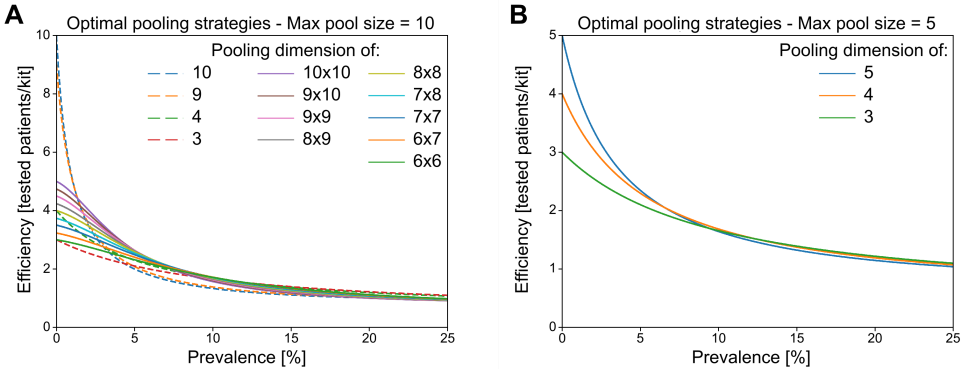

Fig 5. Optimal pooling strategies given a maximum pool size of **a.** 10 and **b.** 5.

Table 1. Optimal pooling strategies restricted by a maximum pool size c.

$c = 10$		
Pooling protocol	Prevalence Interval (%) Lower bound	Upper Bound
S_{10}	0.00	1.25
S_9	1.25	1.375
$S_{10\times10}$	1.375	4.875
$S_{9\times10}$	4.875	5.25
$S_{9\times9}$	5.25	5.875
$S_{8\times9}$	5.875	6.375
$S_{8\times8}$	6.375	7.375
$S_{7\times8}$	7.375	8.125
$S_{7\times7}$	8.125	9.625
$S_{6\times7}$	9.625	10.875
$S_{6\times6}$	10.875	12.125
S_4	12.125	12.375
S_3	12.375	25.00

$c = 5$		
Pooling protocol	Prevalence Interval (%) Lower bound	Upper Bound
S_5	0.00	6.625
S_4	6.625	12.375
S_3	12.375	25.00

Prevalence intervals and their respective optimal pooling strategy. S_m are single pooling protocols and $S_{m\times n}$ are matrix pooling protocols.

3.3.2 *Evaluation*

Quantitative evaluation. We used the standard metrics in machine learning for detection problems: Precision-Recall curves, the maximum F-measure, and Average Precision. Additionally, we measured the efficiency of our models in silico in terms of the number of tests used when employing the model's output as the criterion for Smart Pooling. We calculated the efficiency as the fraction between the initial number of samples and the total number of tests used, following a two-step pooling strategy. First, we sorted the individual tests based on the predicted prevalence by decreasing order. Then, we grouped the samples following the original order and the pool size. Finally, the total number of tests corresponds to the number of pools in the first step added to the individual samples from the second step, that is, samples from the pools that were positive in the first step.

Prevalence and Efficiency graphs For our best model at test time, we simulated different prevalences from the data up to 25%. We calculated the model's efficiency for each prevalence and thus obtained a point in the prevalence efficiency graph. Originally, the prevalence was 22.6% for the *Test Center dataset*. We obtained higher and lower prevalence values by removing positive and negative samples, respectively. Due to the nature of the data, we could not remove individual samples; thus, we eliminated all the samples of a test center on a given date.

3.3.3 *Model training*

We trained a machine learning method for the Test Center dataset. We performed the training of our models in three phases. First, we further split the training data into two disjoint subsets: training and validation. Secondly, we used the AutoML from H2O library [20] to explore multiple machine learning models. We used the performance of the validation subset as the criterion to select the best model. Lastly, we obtained our final model by retraining the best model using all the available training data. During the first training phase, we selected the validation subset such that we predicted as many dates as those in the test set. The maximum number of dates to predict in May was five. We removed the test centers that did not have sufficient dates for the validation split but included them back for the second training phase. Below we explain the details for training.

The level of granularity of this dataset allowed us to have information on the number of positive and total tests per test center on a given date. However, we do not have daily reports from each test center. Thus, we explicitly modeled the data as a sparse time series. We trained the machine learning methods to predict the fraction of positive tests for a center until the current date. Afterward, we assigned the incidence of each sample from a test center on a given date. Said incidence corresponds to the probability of the sample from the test center testing positive on that date. We sorted the tests by decreasing incidence, simulated a two-step pooling protocol, computed the number of used tests, and calculated the efficiency of Smart Pooling. The best model for this task was a Gradient Boosting Machine (GBM) [21] with 50 trees, a constant depth of 5, a minimum of 12 leaves, and a maximum of 29 leaves.

To predict the incidence for a date in the validation or test set, we defined a descriptor calculated from the available training data for each testing center. We included in the dimensions of the features, the cumulative tests of each institution up to every date within the time series, and the total number of tests from all the institutions at the corresponding date. To include the temporal information, we defined as features the date in YY-MM-DD format and the relative date in the number of days since the first date on the time series. Additionally, we created a variable that we named gap, which encodes the distance between two consecutive entries from the same test center. In particular, during test, the gap encodes the distance between the last training date and the date to predict. The gap provided the model with an estimate of the analyzed time window.

As such, we compute the descriptor's features by analyzing the relative differences between variables on the last known date and those from the training days in the current gap. These gap features comprise the cumulative number of tests, total tests, number of positive samples for each institution at the last available date, and the corresponding incidences. We complemented the descriptor with features that indicate the number of days, relative to the current prediction date, from the first n number of positive tests in each test center. For our experiments we defined $n = 1, 5, 10, 100, 500, 1000$. The rationale behind this feature was to provide the model with a temporal encoding of the incidence's evolution. These features were adapted from a public kernel from Kaggle's COVID-19 Global Forecasting Challenge [1].

The key idea of Smart Pooling is maintaining p artificially low, even under scenarios with a high overall prevalence of COVID-19, by reordering the samples according to a priori estimates of prevalence, before the pooling takes place. We demonstrate efficiency gains at simulated prevalences of up to 25%, considerably higher than those applicable for even optimal standard pooling methods.

4 Conclusions

At the prevalence levels for COVID-19 that many diagnostic laboratories are currently managing, two-step pooling efficiency is significantly reduced. Here we show that it is possible to increase pooling efficiency by using machine learning to separate likely positive cases and then pool the rest of the samples using what we show to be an optimal strategy. From our understanding, access to more detailed patient information is very likely to improve predictor performance.

Smart Pooling enhances but does not replace molecular testing. Smart Pooling uses artificial intelligence to enhance the performance of well-established diagnostics. It is an example of how data-driven models can complement, not replace, high-confidence molecular methods. Its robustness to the variability of the available data, prevalence, and model performance and its independence of pooling strategy make it compelling to apply at large scales. Additionally, its continuous learning should make it robust to the pandemic's evolution and our understanding of it.

Smart Pooling could ease access to large scale testing. This pandemic has presented challenges to all nations, regardless of their income. As the number of infected people and the risk of contagion increase, more testing is required. However, the supply of test kits and reagents cannot cope with the demand, with most countries not being able to perform 0.3 new tests per thousand people [3]. Adopting Smart Pooling could translate into more accessible and larger-scale massive testing. In the case of Colombia, this could mean testing 20,000 samples daily, instead of 11,000, in mid-June 2020 [3]. If deployed globally, Smart Pooling truly has the potential to

[1] `https://www.kaggle.com/rohanrao/covid-19-w3-lgb-mad/notebook`

empower humanity to respond to the COVID-19 pandemic. It is an example of how artificial intelligence can be employed to bring social good.

References

1. N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, and W. Tan, "A novel coronavirus from patients with pneumonia in China, 2019," *New England Journal of Medicine*, vol. 382, pp. 727–733, feb 2020.

2. World Health Organization, "Coronavirus disease 2019 (covid-19): situation report, 72," 2020.

3. E. O.-O. Max Roser, Hannah Ritchie and J. Hasell, "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

4. R. Dorfman, "The Detection of Defective Members of Large Populations," *The Annals of Mathematical Statistics*, vol. 14, no. 4, pp. 436–440, 1943.

5. R. M. Phatarfod and A. Sudbury, "The use of a square array scheme in blood testing," *Statistics in Medicine*, vol. 13, no. 22, pp. 2337–2343, 1994.

6. C. Kendziorski, R. A. Irizarry, K.-S. Chen, J. D. Haag, and M. N. Gould, "On the utility of pooling biological samples in microarray experiments," *Proceedings of the National Academy of Sciences*, vol. 102, no. 12, pp. 4252–4257, 2005.

7. B. Abdalhamid, C. R. Bilder, E. L. McCutchen, S. H. Hinrichs, S. A. Koepsell, and P. C. Iwen, "Assessment of Specimen Pooling to Conserve SARS CoV-2 Testing Resources," *American Journal of Clinical Pathology*, vol. 153, pp. 715–718, 2020.

8. I. Yelin, N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli, N. Gandali, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen, M. Szwarcwort-Cohen, and R. Kishony, "Evaluation of COVID-19 RT-qPCR test in multi-sample pools," *medRxiv*, p. 2020.03.26.20039438, 2020.

9. A. M. Eis-Hübinger, M. Hönemann, J. J. Wenzel, A. Berger, M. Widera, B. Schmidt, S. Aldabbagh, B. Marx, H. Streeck, S. Ciesek, U. G. Liebert, D. Huzly, H. Hengel, and M. Panning, "Ad hoc laboratory-based surveillance of SARS-CoV-2 by real-time RT-PCR using minipools of RNA prepared from routine respiratory samples," *Journal of Clinical Virology*, vol. 127, jun 2020.

10. C. A. Hogan, M. K. Sahoo, and B. A. Pinsky, "Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2," *JAMA*, vol. 323, no. 19, pp. 1967–1969, 2020.

11. G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study," *PLoS ONE*, vol. 15, no. 4, p. e0232391, 2020.

12. O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *arXiv:2003.05037*, 2020.

13. L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, *et al.*, "An interpretable mortality prediction model for covid-19 patients," *Nature Machine Intelligence*, vol. 2, p. 283–288, 2020.

14. F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," *PLoS ONE*, vol. 15, no. 3, p. e0231236, 2020.

15. C. Menni, A. M. Valdes, M. B. Freidin, C. H. Sudre, L. H. Nguyen, D. A. Drew, S. Ganesh, T. Varsavsky, M. J. Cardoso, J. S. El-Sayed Moustafa, A. Visconti, P. Hysi, R. C. E. Bowyer, M. Mangino, M. Falchi, J. Wolf, S. Ourselin, A. T. Chan, C. J. Steves, and T. D. Spector, "Real-time tracking of self-reported symptoms to predict potential COVID-19.," *Nature Medicine*, may 2020.

16. S. M. Taylor, J. J. Juliano, P. A. Trottman, J. B. Griffin, S. H. Landis, P. Kitsa, A. K. Tshefu, and S. R. Meshnick, "High-throughput pooling and real-time PCR-based strategy for malaria detection," *Journal of Clinical Microbiology*, vol. 48, no. 2, pp. 512–519, 2010.

17. C. R. Weinberg, "Editorial: Making the Best Use of Test Kits for COVID-19," *American Journal of Epidemiology*, 2020.

18. V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, D. G. Mulders, B. L. Haagmans, B. Van Der Veer, S. Van Den Brink, L. Wijsman, G. Goderski, J. L. Romette, J. Ellis, M. Zambon, M. Peiris, H. Goossens, C. Reusken, M. P. Koopmans, and C. Drosten, "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR," *Eurosurveillance*, vol. 25, jan 2020.

19. SEASUN Biomaterials, "U-TOP™ COVID-19 Detection Kit," tech. rep., USA Food and Drug Administration, 2020.

20. H20ai, *Python Interface for H2O*, 05 2020. Python module version 3.10.0.8.

21. A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, p. 21, 2013.

22. J. Emmanuel, M. Bassett, H. Smith, and J. Jacobs, "Pooling of sera for human immunodeficiency virus (hiv) testing: an economical method for use in developing countries.," *Journal of Clinical Pathology*, vol. 41, no. 5, pp. 582–585, 1988.

23. T. C. Quinn, R. Brookmeyer, R. Kline, M. Shepherd, R. Paranjape, S. Mehendale, D. A. Gadkari, and R. Bollinger, "Feasibility of pooling sera for hiv-1 viral rna to diagnose acute primary hiv-1 infection and estimate hiv incidence," *Aids*, vol. 14, no. 17, pp. 2751–2757, 2000.

24. CNN, C. Devine, D. Griffin, and R. Kuznia, "Shortage of standard health supplies is 'a huge problem'." Library Catalog: edition.cnn.com.

25. I. Torres, E. Albert, and D. Navarro, "Pooling of nasopharyngeal swab specimens for sars-cov-2 detection by rt-pcr," *Journal of Medical Virology*, 2020.

Acknowledgments

ME, GJ, and PA acknowledge special funding from Facultad de Ingeniería, Universidad de los Andes. The authors thank John Mario González from the Faculty of Medicine at Universidad de los Andes and his team CBMU for devoting the laboratory for COVID testing and their contributions during the laboratory certification. The authors also acknowledge the team members of the GenCore Covid Laboratory at Universidad de los Andes. The authors thank Secretaría de Salud de Bogotá and Alcaldía de Bogotá for their support during the laboratory certification process and access to complementary information for the samples.

Ethics Declaration

Competing interests

The authors declare no competing interests.

Author Contributions

ME, GJ, LBS, AC, CG, DV, MFR, and JM designed and trained the AI models, and curated the dataset. MC, MGS, JMPL, and SR performed the molecular tests. ALM MFS, MV, and JMPL developed mathematical formulations to find the optimal pooling strategies. MFS and JMPL evaluated the applicability of the methods in the laboratory. ALM, MV and PA developed the theoretical interpretation of the results. LBS, GJ, JMW, JM, ALM, and MV prepared the figures. ME, GJ, LBS, AC, CG, DV, MFR, JM, JMW, MFS, and MV wrote the manuscript. All authors revised and accepted the final version of the manuscript. PA supervised and directed the research.

Supporting information

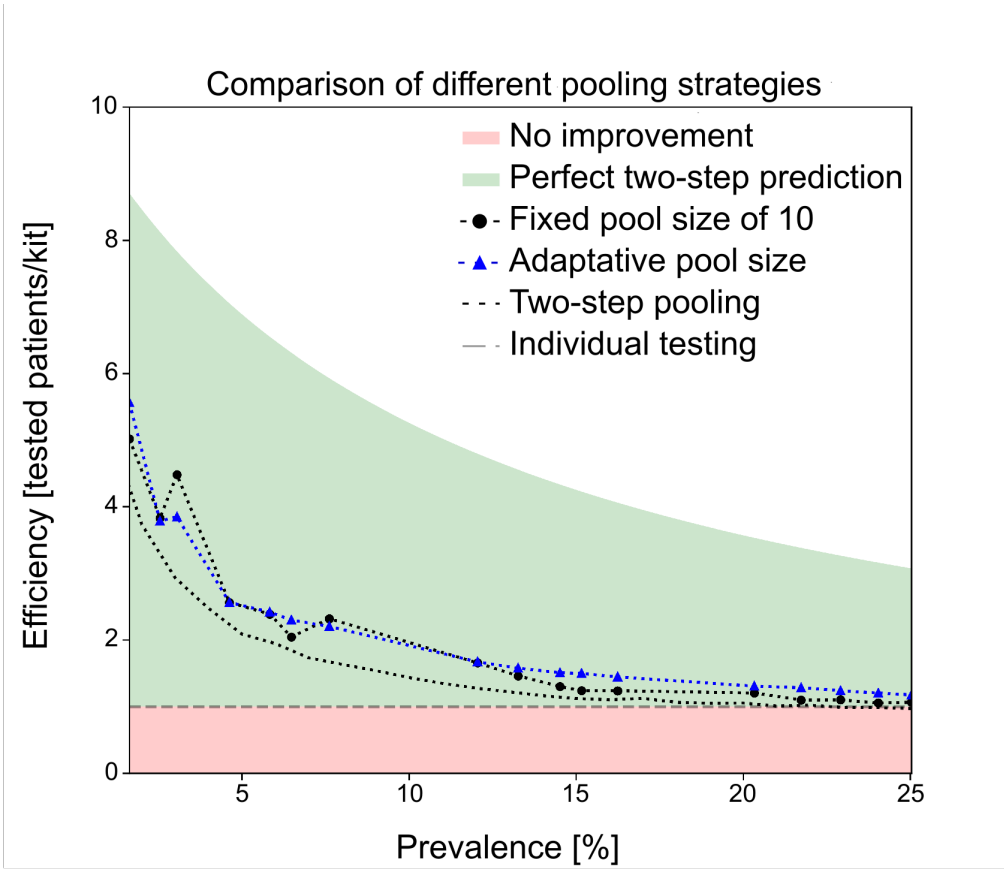

Fig S1. Estimated efficiency of different pooling strategies enhanced by Smart Pooling.

S1.1 Practical considerations for implementation

Historically, pooling was used extensively during the second world war, but since then, it has mostly been implemented in specific niches such as testing blood for diseases and reducing cost in developing countries [16, 22, 23]. In the specific context of SARS-CoV-2, there are several incentives to implement pooling. Most significantly, the current shortage of reagents, especially in developing countries that do not produce these and have limited stocks. Additionally, if implemented correctly, these methods can increase throughput, another motivation for developing countries and locales with a limited number of certifiable qPCR machines. Finally, through the reduced use of reagents and increased throughput, these methods can reduce costs and motivate health care providers.

Here we considered Dorfman and array testing algorithms because they are easily compatible with a manual implementation of pooling. Although it is possible to find the optimal algorithm for pooling at a particular prevalence, it may be cumbersome to implement all the protocols. Fortunately, most of the pooling algorithms are relatively

robust over a broader range than that in which they are optimal, so implementing some should increase efficiency without added complexity. Different algorithms will also affect tip use as well as plating time. Filter tips have been in short supply since the beginning of the pandemic [24], and some of the pooling protocols increase efficiency at the expense of increased tip use. Plating time will depend on both the pooling algorithm and the experience of the personnel doing the pooling. Each lab has to adapt the specifics of the protocol for their needs.

Successful implementation of PCR-based pooling requires understanding the limitations of the method and performing viability tests. Although it is possible to pool reactions including primers that detect the virus only when it is present, it is not possible, to our knowledge, to pool the RNAse P positive control. This phenomenon occurs because the samples are positive unless a problem has occurred with the sample or the extraction of RNA; it involves finding a negative among positives where samples have variable RNA content and an exponential amplification step. In our implementation, the two reactions are separate. The RNAse P control is performed on a different plate, using a faster PCR protocol (1.5 hours instead of 2.5 hours). In kits for which the specific primers and the control are multiplexed, for example, it is not possible to ensure the presence of RNA when pooling.

Sensitivity must also be taken into account when implementing pooling. Each two-fold dilution results in the increase of the Ct value by 1 unit ($1\Delta Ct$), on average. Some have detected dilutions up to 32-fold [8], but only for samples with average Ct, not for samples close to the detection limit. It is necessary to calibrate the dilution process in each lab since it depends on the kit, machine, and operators. Another limit for the size of the pools is the total reagent volume in each pool. Our plates can handle $10\mu l$ of a sample while other plate-kit combinations may handle just $5\mu l$. Depending on the operators, pipetting volumes under $1\mu l$ may cause quality control problems, and since we rely on many volunteers for our testing, we only considered pools of up to 10, which also keeps the maximum Ct of the first round below 42, the limit of cycles in our machine. Additionally, we increased the cutoff in Ct for the first round from 38 to 38+ ΔCt at this dilution, from the calibrations. This modification may increase the number of false positives in the first round, but since the second round samples are sampled individually, the standard cutoff can be used to eliminate potential false positives from the first round.

Pooling before RNA extraction is an attractive option since it reduces reagent use, but we did not implement it for several reasons. The first is that the expected increase in Ct of 5 [25], was above the practical limits for our machine and kit combination. Secondly, it is not possible to flag poorly taken samples in which there is no RNA, for the same reasons, it is not possible to pool the RNAse P control. Finally, we received heterogeneous samples from both the upper and lower respiratory tracts, which resulted in false negatives when pooled together. It is likely that implementing pooling before RNA extraction is possible with other kits or more homogeneous samples.

The long-term management of COVID-19 will likely require the use of

complementary approaches, including testing for antibodies against the virus. The methods presented in this paper can be applied in this context and could increase the measurement of seropositivity in the population.