Hemogram data as a tool for decision-making in COVID-19 management: Applications to resource scarcity scenarios

Eduardo Avila Corresp., Equal first author, 1, 2, 3 , **Alessandro Kahmann** Equal first author, 3, 4 , **Clarice Alho** 1, 3 , **Marcio Dorn** 3, 5

1 Forensic Genetics Laboratory, School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil

2 Technical Scientific Section, Federal Police Department in Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

3 National Institute of Science and Technology - Forensic Science, Porto Alegre, Rio Grande do Sul, Brazil

4 Institute of Mathematics, Statistics and Physics, Federal University of Rio Grande, Rio Grande, Rio Grande do Sul, Brazil

5 Laboratory of Structural Bioinformatics and Computational Biology, Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

Corresponding Author: Eduardo Avila
Email address: e.avila@edu.pucrs.br

Background: COVID-19 pandemics has challenged emergency response systems worldwide, with widespread reports of essential services breakdown and collapse of health care structure. A critical element involves essential workforce management since current protocols recommend release from duty for symptomatic individuals, including essential personnel. Testing capacity is also problematic in several countries, where diagnosis demand outnumbers available local testing capacity.

Purpose: This work describes a machine learning model derived from hemogram exam data performed in symptomatic patients and how they can be used to predict qRT-PCR test results.

Methods: Hemogram exams data from 510 symptomatic patients (73 positives and 437 negatives) were used to model and predict qRT-PCR results through Naïve-Bayes algorithms. Different scarcity scenarios were simulated, including symptomatic essential workforce management and absence of diagnostic tests. Adjusts in assumed *prior* probabilities allow fine-tuning of the model, according to actual prediction context.

Results: Proposed models can predict COVID-19 qRT-PCR results in symptomatic individuals with high accuracy, sensitivity and specificity, yielding a 100% sensitivity and 22.6% specificity with a *prior* of 0.9999; 76.7% for both sensitivity and specificity with a *prior* of 0.2933; and 0% sensitivity and 100% specificity with a *prior* of 0.001. Regarding background scarcity context, resources allocation can be significantly improved when model-based patient selection is observed, compared to random choice.

Conclusions: Machine learning models can be derived from widely available, quick, and inexpensive exam data in order to predict qRT-PCR results used in COVID-19 diagnosis. These models can be used to assist strategic decision-making in resource scarcity scenarios, including personnel shortage, lack of medical resources, and testing insufficiency.

Hemogram Data as a Tool for Decision-making in COVID-19 Management: Applications to Resource Scarcity Scenarios

Eduardo Avila[1,2,3,*]**, Alessandro Kahmann**[3,4,*]**, Clarice Sampaio Alho**[1,3]**, and Márcio Dorn**[3,5]

[1]**Forensic Genetics Laboratory, School of Health and Life Sciences, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre, RS, Brazil**
[2]**Technical Scientific Section, Federal Police Department in Rio Grande do Sul, Porto Alegre, RS, Brazil**
[3]**National Institute of Science and Technology - Forensic Science, Porto Alegre, RS, Brazil**
[4]**Institute of Mathematics, Statistics and Physics, Federal University of Rio Grande, Rio Grande, RS, Brazil**
[5]**Laboratory of Structural Bioinformatics and Computational Biology, Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil**
[*]***These authors contributed equally to this work.***

Corresponding author:
Eduardo Avila.

Email address: e.avila@edu.pucrs.br; mdorn@inf.ufrgs.br; csalho@pucrs.br; alessandrokahmann@furg.br

ABSTRACT

Background: COVID-19 pandemics has challenged emergency response systems worldwide, with widespread reports of essential services breakdown and collapse of health care structure. A critical element involves essential workforce management since current protocols recommend release from duty for symptomatic individuals, including essential personnel. Epidemiological screening is also problematic in several countries, where diagnosis demand outnumbers available local testing capacity.

Purpose: This work describes a machine learning model derived from hemogram exam data performed in symptomatic patients and how they can be used to predict qRT-PCR test results.

Methods: Hemogram exams data from 510 symptomatic patients (73 positives and 437 negatives) were used to model and predict qRT-PCR results through *Naïve-Bayes* algorithms. Different scarcity scenarios were simulated, including symptomatic essential workforce management and absence of diagnostic tests. Adjusts in assumed prior probabilities allow fine-tuning of the model, according to actual prediction context.

Results: Proposed models can predict COVID-19 qRT-PCR results in symptomatic individuals with high accuracy, sensitivity and specificity.yielding a 100% sensitivity and 22.6% specificity with a *prior* of 0.9999; 76.7% for both sensitivity and specificity with a *prior* of 0.2933; and 0% sensitivity and 100% specificity with a *prior* of 0.001. Regarding background scarcity context, resources allocation can be significantly improved when model-based patient selection is observed, compared to random choice.

Conclusions: Machine learning models can be derived from widely available, quick, and inexpensive exam data in order to predict qRT-PCR results used in COVID-19 diagnosis. These models can be used to assist strategic decision-making in resource scarcity scenarios, including personnel shortage, lack of medical resources, and testing insufficiency.

INTRODUCTION

Since its first detection and description (Huang et al., 2020), COVID-19 expansion has brought worldwide concerns to governmental agents, public and private institutions, and health care specialists. Declared as a pandemic, this disease has deeply impacted many aspects of life in affected communities. Relative lack of knowledge about the disease particularities has led to significant efforts devoted to alleviating its effects (Lipsitch et al., 2020).

Alternatives to mitigate the disease spread include social distancing (Anderson et al., 2020). Such a course of action has shown some success in limiting contagion rates (Tu et al., 2020). However, isolation policies manifest drawbacks as economic impact, with significant effects on macroeconomic indicators and unemployment rates (Nicola et al., 2020). To address this, governments worldwide have proposed guidelines to manage the essential workforce, considered pivotal for maintaining strategic services and provide an appropriate response to the pandemics expansion (Black et al., 2020).

Widespread reports of threats to critical national infrastructure have been presented, with significant impact associated with medical attention (Kandel et al., 2020). Significant pressure is being faced by emergency response workers, with some countries on the brink of collapse of their national health systems (Tanne et al., 2020). The main concern associated with COVID-19 is the lack of extensive testing capacity. Shortage of diagnostic material and other medical supplies pose as a major restraining factor in pandemics control (Ranney et al., 2020).

The most common COVID-19 symptoms are similar to other viral infectious diseases, making the prompt clinical diagnostic impractical (Adhikari et al., 2020). Official guidelines emphasize the use of quantitative real-time PCR (qRT-PCR) assays for detection of viral RNA in diagnosis as the primary reference standard (Tahamtan and Ardebili, 2020). In many countries, test results are hardly available within at least a week, forcing physicians and health care providers to take strategic decisions regarding patient care without quality information.

Previous reports have described alterations in laboratory findings in COVID-19 patients. Hematological effects include leukopenia, lymphocytopenia and thrombocytopenia, while biochemical results show variation on alanine and aspartate aminotransferases, creatine kinase and D-dimer levels, among other parameters (Guan et al., 2020; Huang et al., 2020). Some efforts have been applied to evaluate clinical and epidemiological aspects of this disease using computational methods, such as diagnosis, prognosis, symptoms severity, mortality, and response to different treatments. A useful review of some of these methods is presented by Wynants and collaborators (Wynants et al., 2020).

The main objective of this article is to provide insights to healthcare decision-makers facing scarcity situations, as a shortage of test capacity or limitations in the essential workforce. A useful method of doing so is using hemogram test results. This clinical exam is widely available, inexpensive, and fast, applying automation to maximize throughput. To do so, we have analyzed hemogram data from Brazilian symptomatic patients with available test results for COVID-19. We propose a framework using a *Naïve-Bayes* model for machine learning, where test conditions can be adjusted to respond to actual lack of resources problems. Finally, four distinct scarcity scenarios examples are presented, including handling of the essential workforce and shortage of testing and treatment resources.

MATERIAL AND METHODS

Data Collection

5644 patients admitted to the emergency department of *Hospital Israelita Albert Einstein* (HIAE - São Paulo, Brazil) presenting COVID-19-like symptoms were tested via qRT-PCR. A total number of 599 patients (10.61%) presented positive results for COVID-19. The full dataset

contains patients anonymized ID, age, qRT-PCR results, data on clinical evolution, and a total of 105 clinical tests. Not all data was available for all patients, therefore the amount of missing information is significant, with most available parameters informed for a small fraction of subjects. All variables were normalized (i.e. mean = 0 and variance = 1) to maintain anonymity and remove scale effects. No missing data imputation was performed during model generation to avoid *bias*. Considering the significant ammount of missing data, only 510 patients presented values for all 15 parameters evaluated in hemogram results (comprising the following cell counts or hematological measures: hematocrit, hemoglobin, platelets, mean platelet volume, red blood cells, lymphocytes, leukocytes, basophils, eosinophils, monocytes, neutrophils, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), and red blood cell distribution width (RDW). Data for the above parameters were used in model construction, along with qRT-PCR COVID-19 test results. No baseline characteristics detailing can be presented for the evaluated cohort, once additional patients description is not accessible. Even the limited provided individual information (as subjects ages, for instance) was normalized within the dataset, and therefore cannot be inferred from available data. The full dataset is available in `https://www.kaggle.com/einsteindata4u/covid19` and can also be found in Supplementary Materials.

Machine Learning Analysis - *Naïve Bayes* Classifier

Machine learning (ML) is a field of study in computer science and statistics dedicated to the execution of computational tasks through algorithms that do not require explicit instructions but instead rely on learning patterns from data samples to automate inferences (Mitchell, 1997). These algorithms can infer input-output relationships without explicitly assuming a pre-determined model (Géron, 2017; Hastie et al., 2009). There are two learning paradigms: supervised and unsupervised. Supervised learning is a process in which the predictive models are constructed through a set of observations, each of those associated with a known outcome (*label*). In opposition, in unsupervised learning, one does not have access to the labels, it can be viewed as the task of "spontaneously" finding patterns and structures in the input data.

Our objective with this study is to predict in advance the results of the qRT-PCR test with a supervisedmachine learning model using data from hemogram tests performed on symptomatic patients. The main process can be divided into four steps: (1) *pre-processing of the data* (2) *selection of an appropriate classification algorithm*, (3) *model development and validation*, i.e., the process of using the selected characteristics to separate the two groups of subjects (positive for COVID-19 vs. negative for COVID-19 in qRT-PCR test), and (4) test generated model with additional data. Steps are detailed as follows:

Data Pre-processing: Samples presenting a missing value in any of the 15 evaluated features were removed, in order to avoid bias introduction in model . A total of 510 patients (73 positives for COVID-19 and 437 negatives) presented complete data and were considered for the model construction.

Classification Algorithm: In this work, we use a Gaussian *Naïve Bayes* (NB) classifier, which is a probabilistic machine learning model used for classification tasks. The main reasons for choosing this classifier are due to its low computational cost, its ability to handle missing data and because it presented better classification performance when compared to other evaluated ML techniques, for this particular dataset (data not shown). In medicine, the first computer-learn attempts in decision support were based mainly on the Bayes theorem, in order to aggregate data information to physicians' previous knowledge (Martin et al., 1960). The *Naïve Bayes* (NB) method combines the previous probability of an event (also called *prior probability*, or simply *prior*) with additional evidence (as, for example, a set of clinical data from a patient) to calculate a combined, conditional probability that includes the *prior* probability given the extra

information. The result is the *posterior probability* of an outcome, or simply *posterior*. This classifier is called "naïve" because it considers that each exam result (variables) is independent of each other. Since this situation is not realistic in medicine, the model should not be interpreted (Schurink et al., 2005). Besides this drawback, it can outperform more robust alternatives in classification tasks, and once it reflects the uncertainty involved in the diagnosis, Bayesian approaches are more suitable than deterministic techniques (Gorry and Barnett, 1968; Hastie et al., 2009).

Model Development and Validation: A classifier is an estimator with a predict method that takes an input array (test) and makes predictions for each sample in it. In supervised learning estimators (our case), this method returns the predicted labels or values computed from the estimated model (for this work, positive or negative for COVID-19). *Cross-validation* is a model evaluation method that allows one to evaluate an estimator on a given dataset reliably. It consists of iteratively fitting the estimator on a fraction of the data, called training set, and testing it on the left-out unseen data, called test set. Several strategies exist to partition the data. In this work, we used the *Leave-one-out* (LOO) cross-validation model, as in Chang et al. (Chang et al., 2003), since this method is appropriate to handle small sample size datasets. The number of data points was split N times (number samples). The method was trained on all the data except for one point, and a prediction was made for that point. The proposed approach was implemented in *Python v.3* (https://www.python.org) code using *Scikit-Learn v. 0.22.2* (Pedregosa et al., 2011) as a backend.

Missing Data Tolerance: In order to evaluate the adequacy and generalization power of the proposed model, as well as its tolerance to handle samples containing missing data (i.e. at least one variable with no informed values), an additional set of 92 samples (10 positives for COVID-19 and 82 negatives) was obtained from the patient database. Those samples were not initially employed in model delineation, considering they present a single missing value among all 15 employed hemogram parameters. This "incomplete dataset", comprising 92 samples with a single missing value information per sample, was then submitted to the previously generated model, in order to evaluate classification performance and missing values handling ability for the model.

RESULTS

Descriptive Analysis

For data description, probability density function (PDF) of all 15 hemogram parameters were estimated through the original sample by kernel density estimator. Some hemogram parameters present notable differences between the distributions of positive and negative results, mainly regarding its modal value (distribution peak value) and variance (distribution width). Differences are summarized in Table 1. Regarding basophiles, eosinophils, leukocytes and platelets counts, qRT-PCR positive group distribution shows lower modal value and lower variance. On the other hand, monocyte count displays opposite behavior, once lower modal value and variance are observed for the qRT-PCR positive group. The remaining nine hemogram parameters did not show a notable difference between negative and positive groups. All variables contribute to the classification model, and despite the fact classification can be perfomed without the complete set of parameters (i.e. including missing data), a most sucessfull prediciton is achieved when complete hemogram information is used as input. PDF analysis results are presented in Figure 1.

Naïve Bayes Model Results

A NB classifier based on training set hemogram data was developed. Under the model, the complete range of *prior* probabilities (from 0.0001 to 0.9999 by 0.0001 increments) was scrutinized, and *posterior* probability of each class was computed for different *prior* conditions. A
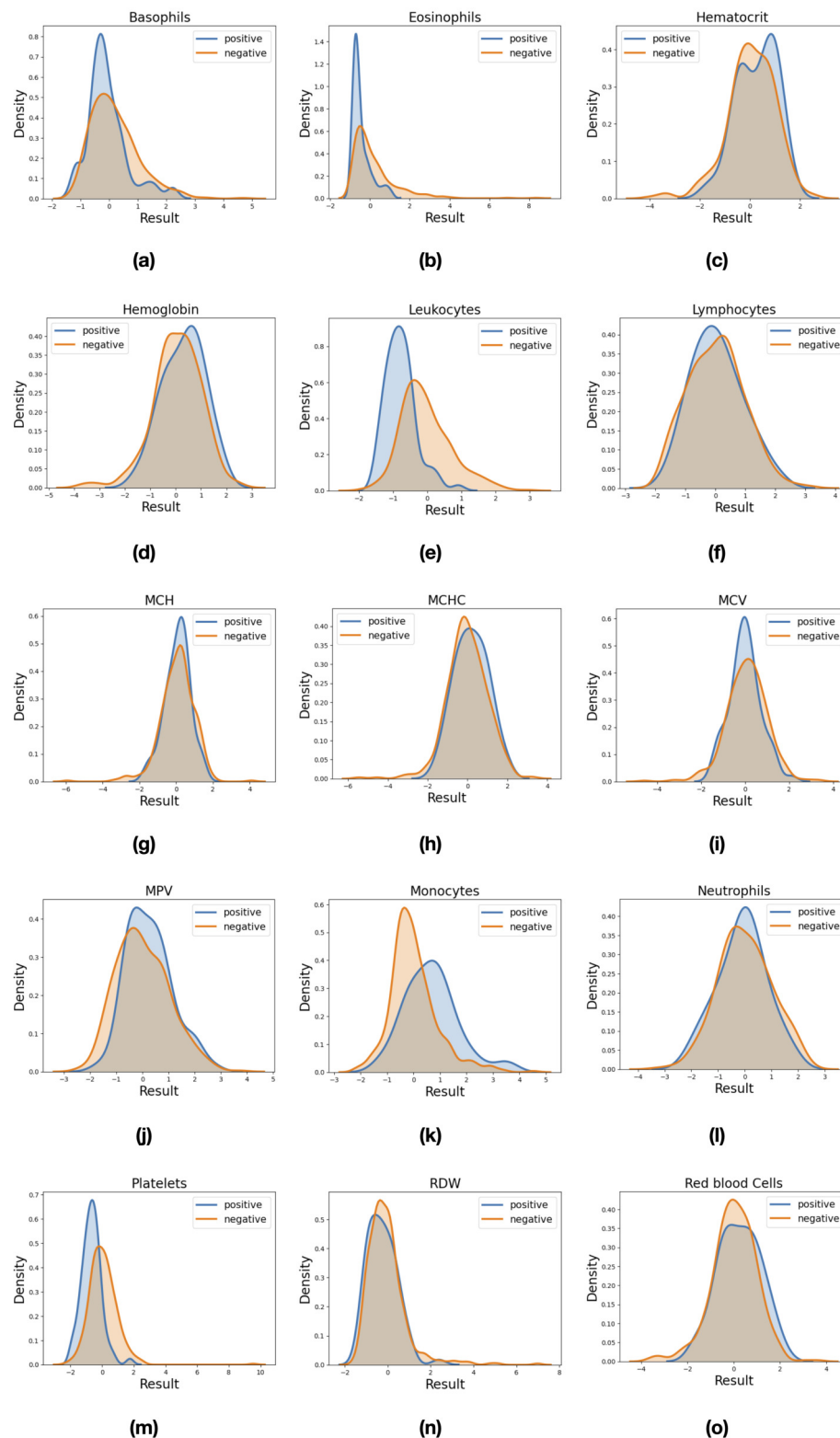
Figure 1. Probability density function (PDF) of all 15 hemogram parameters.
(a) Basophils; **(b)** Eosinophils; **(c)** Hematocrit; **(d)** Hemoglobin; **(e)** Leukocytes; **(f)** Lymphocytes; **(g)** MCH; **(h)** MCHC; **(i)** MCV; **(j)** MPV; **(k)** Monocytes; **(l)** Neutrophils; **(m)** Platelets; **(n)** RDW; **(o)** Red Blood Cells

Table 1. Descriptive analysis of hemogram parameters used in present study.

Parameter	Modal value	Variance
Basophiles	Reduced in positive cases	Reduced in positive cases
Eosinophiles	Reduced in positive cases	Reduced in positive cases
Leukocytes	Reduced in positive cases	Reduced in positive cases
Monocytes	Augmented in positive cases	Augmented in positive cases
Platelets	Reduced in positive cases	Reduced in positive cases
Parameters not shown displayed no difference between groups		

188 *posterior* probability value of 0.5 was defined as the classification threshold in one of the positive
189 or negative predicted groups. Resulting model showed a good predictive power of the qRT-PCR
190 test result based on hemogram data. Figure 2 shows the accuracy, sensitivity, and specificity
191 curves derived from the model for different *prior* probabilities of each class (positive or negative
192 for COVID-19). Reported *prior* probabilities refer to positive COVID-19 condition.
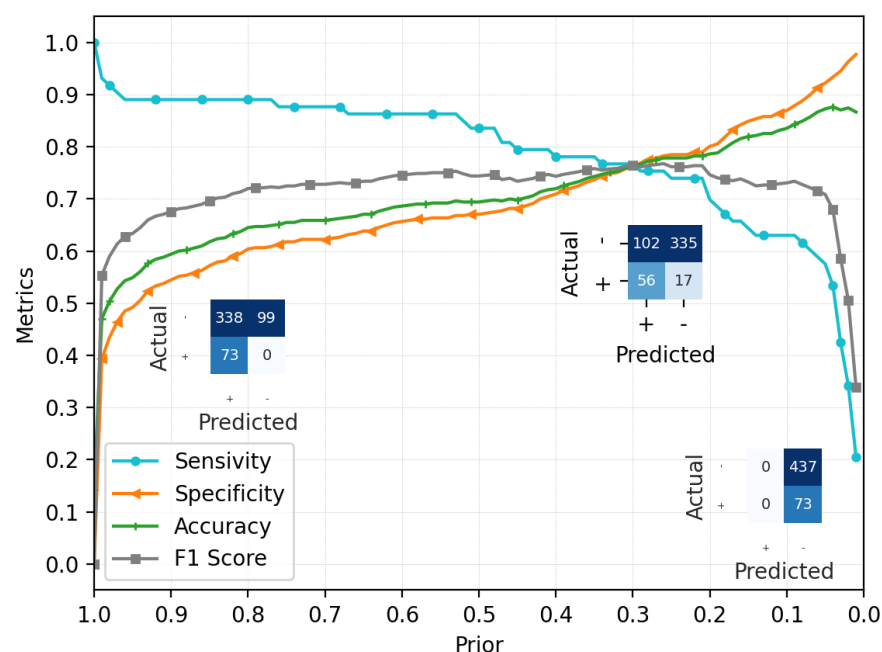
Figure 2. Performance metrics of proposed Naive-Bayes model. *prior* probabilities are
presented in reference to positive qRT-PCR prediction. Confusion matrices (left to right) are
presented for 0.9999, 0.2933 and 0.0001 *Prior* probabilities, respectively. Sensitivity=True
Positive Ratio; Specificity=True Negative Ratio

193 When setting the *prior* probability to the maximum defined value (0.9999), the NB classifier
194 correctly diagnosed all PCR positive cases. On the other hand, such configuration improperly
195 predicted 77.3% of negative PCR results as positive. Regarding the lower possible *prior*
196 probability setting, it does not classify a single observation as positive. This result can be
197 explained by the unbalanced number of observations for each class, tending to over classify
198 samples as the class with more observation, i.e. negative results. Such characteristics can also be

noticed in the general accuracy, since smaller values for the *prior* used in the classifier tend to diagnose all observations as belonging to the dominant class (negative) and consequently raising the total of correctly classified samples. The break-even point is met when *prior* probability is set to 0.2933. Under this condition, all metrics are approximately 76.6%.

Regarding the model sensitivity, the rate of positive samples correctly classified is over 85% within 0.999 to 0.5276 range, with small decrease of it when the *prior* probability of positive result is diminished within this range. When *prior* is set to values under 0.1, the number of positive predicted samples decreases rapidly, yielding lower sensitivity. As for specificity, it presents linear growth as tested *priors* decrease. Ultimately, the accuracy results profile are similar to specificity, due to the negative patients dominance. Figure 3 shows additional results for the classification performance presented by the proposed method. Figure 3a displays the ROC (receiver operating characteristics) curve for the methods. Obtained measure of the area under the curve (AUROC) is equivalent to 0.84, suggesting excellent prediction performance for the model. Figures 3b to 3d presents prediction results for the baseline model, using a dummy classifier for the most frequent class (negative qRT-PCR, Figure 3b), and for stratified (Figure 3c) and uniform (Figure 3d) evaluations.

As mentioned above, *prior* probability choice has a critical relevance in proposed model use. It is clear that, when extreme values of positive probability are applied (close to 0 or 1), specific classes (positive or negative qRT-PCR test results predictions) are favored, increasing its ability of correct detection. As an example, when a value of 0.9999 is set for *prior* probability of positive result is set, an increase in misclassification in negative class results is observed. At the same time, it is possible to properly identify samples where hemogram evidence strongly indicates a negative result, according to the model. This is based on the fact that evidence used in the model construction (in present case, hemogram data) must strongly support the reduction of *posterior* probability of disease to values under 0.5, therefore leading to a negative result. This logic can be applied to fine tune the *prior* probability used in the model, in order to improve correct classification of positive or negative groups prediction. Examples of how to use this feature is provided in the "Discussion" section. Samples including a single missing value (n=92, including 10 qRT-PCR positives) were used to test the missing data tolerance presented by the proposed model. Figure 4 presents results obtained from the model application to this incomplete (lacking information for one variable per sample) dataset.

DISCUSSION

Laboratory findings can provide vital information for pandemics surveillance and management (Lippi and Plebani, 2020). Hemogram data have been previously proposed as useful parameters in diagnosis and management of viral pandemics (Shimoni et al., 2013). In the present work, an analysis concerning hemogram data from symptomatic patients suspected of COVID-19 infection was executed. A machine learning model based on *Naïve Bayes* method is proposed in order to predict actual qRT-PCR from such patients. The presented model can be applied to different situations, aiming to assist medical practitioners and management staff in key decisions regarding this pandemic, especially in conditions of limited access to medical resources (Brown et al., 2020). Figure 5 summarizes model construction and application. Predictions are not intended to be used as a diagnostic method since this technique was designed to anticipate qRT-PCR results only. As such, it is highly dependant on factors affecting qRT-PCR efficiency, and its prediction capability is dependent on the sensitivity, accuracy, and specificity of the original laboratory exam (Sethuraman et al., 2020).

Descriptive analysis of hemogram clinical findings shows differences in blood cell counts and other hematological parameters among COVID-19 positive and negative patient results.
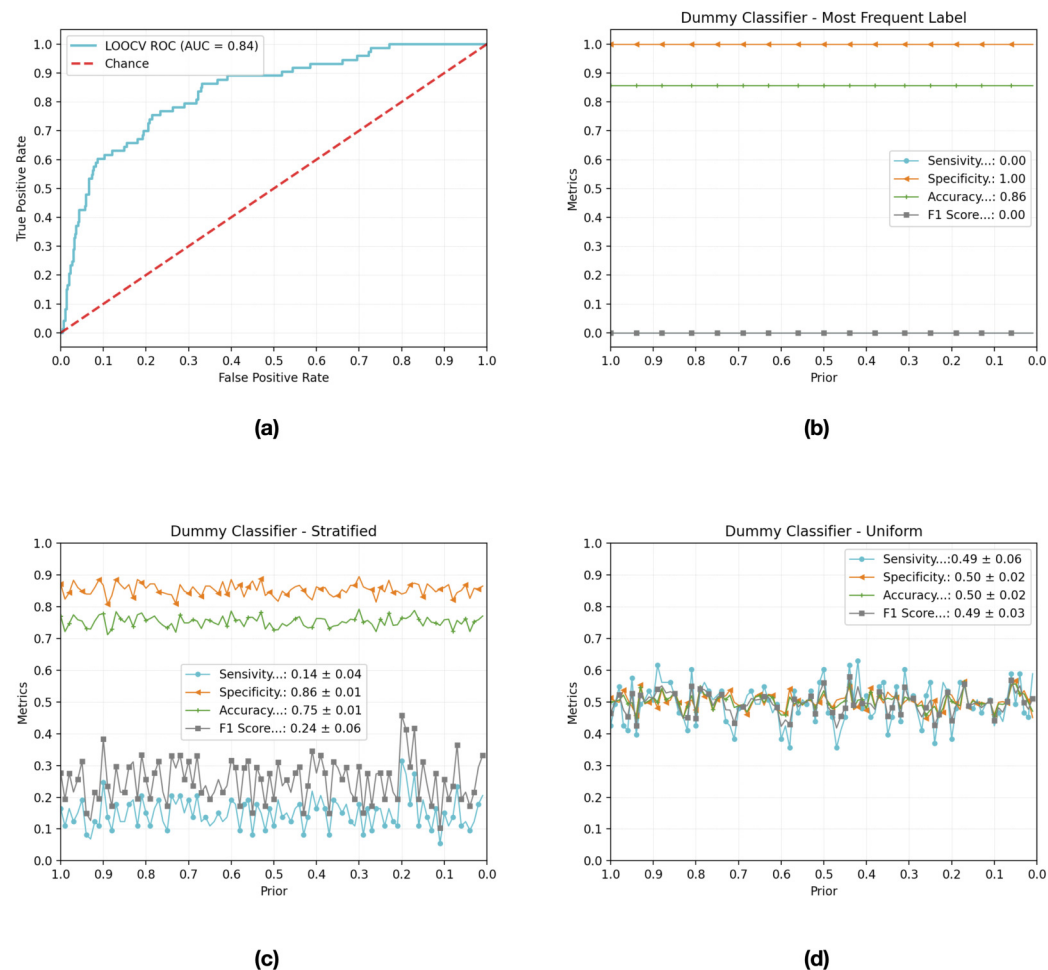
Figure 3. Classification performance evaluation. Classification perfomance was evaluated thorugh ROC curve **(a)** or against the baseline model using a dummy classifier for **(b)** the most frequent class (negative); **(c)** stratified data; **(d)** uniform data. AUC=Area Under the Curve; Sensitivity=True Positive Ratio; Specificity=True Negative Ratio

246 Differences are conspicuous among three measures (leukocytes, monocytes and platelets) and
247 more discrete to additional two (basophiles and eosinophiles). It is possible that differences
248 are also present across the complete data spectrum, even though they are not clearly visualized
249 with PDF data. These results are in accordance to previous reports of changes in laboratory
250 findings in COVID-19 infected patients, where conditions as leukopenia, lymphocytopenia and
251 thrombopenia were reported (Fan et al., 2020; Ding et al., 2020). It is important to highlight
252 that data analysis is not sufficient to characterize clinical hematological alterations in evaluated
253 patients (when compared to demographic hematologic parameters data), once data was normal-
254 ized for the evaluated sample set only. However, even within this particular quota of population
255 (individuals presenting COVID-19-like symptoms), differences were found between individuals
256 presenting negative or positive qRT-PCR COVID test results. The proposed NB-ML model can
257 be helpful in accessing different levels of information from hemogram results, through inferring
258 non-evident patterns and parameter relationships from this data. Also, our simulations suggest
259 that the NB model has at least some degree of tolerance to missing data values, which can be
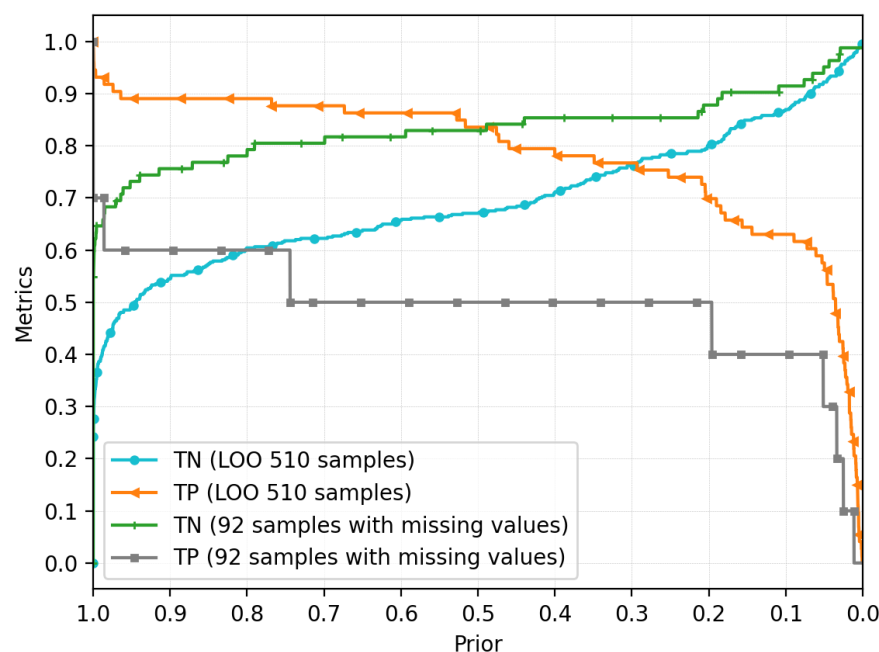
Figure 4. Classification performance for training (LOO) and missing value-containing datasets. Results presented for the complete *prior* probability range. Results are presented as the percentage of correctly predicted qRT-PCR exams. Informed *prior* probability refers to positive outcome. TN: true negative; TP: true positive.
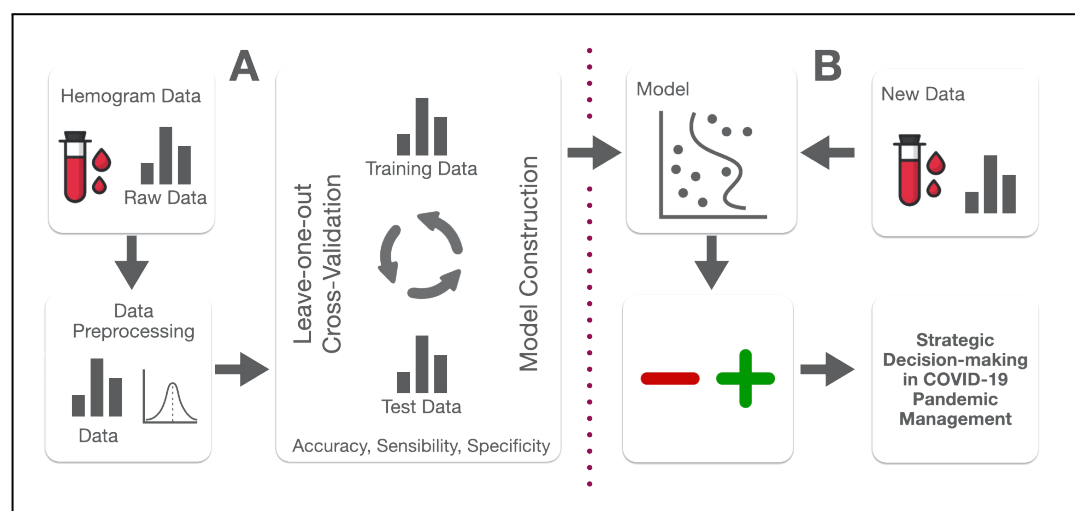
Figure 5. NB Model construction and application diagram.

260 advantgeous when compared to other ML techniques.

261   Bayesian techniques are based on the choice of a *prior* probability of an event (in present

262 case, positive result for qRT-PCR test). The method considers actual evidence (hemogram data)

263 to result in a *posterior* probability of the outcome (prediction of a positive result). By changing

264 the selected *prior* probability, we can derive an uncertainty analysis of the model to understand

265 its distribution. Uncertainty can be then applied to adequately adapt the classifier to a particular
266 ongoing context. This option allows the evaluation of different decision-making scenarios
267 concerning diverse aspects of pandemics management. During a crisis situation, measures should
268 be taken seeking to maximize benefits and achieve a fair resource allocation (Emanuel et al.,
269 2020). To illustrate the model flexibility and how it can be used to help on this matter, a general
270 framework of application is proposed, followed by a simulation of four scenarios where resource
271 scarcity is assumed.

Application Framework

273 The proposed NB model can be applied in two distinct situations. When clinical data is available
274 for a particular patient, it is highly recommended that medical staff determine the *prior* probability
275 on a case-by-case basis. When no clinical or medical data is available, or when decisions
276 regarding resource management involving multiple symptomatic patients are necessary, the
277 model can be used in multiple individuals simultaneously, aiming to identify those with higher
278 probabilities of presenting positive qRT-PCR results.

Individual Assessment

280 Individual risk management and personal evaluation is essential for COVID-19 response (Gasmi
281 et al., 2020). Individuals presenting COVID-19 symptoms are medically evaluated where no
282 COVID-19 test is available for appropriate diagnosis confirmation. Medical practitioners can
283 determine a probability of disease based on anamnesis, symptoms, clinical exams, laboratory
284 findings and other available data. This probability of infection, as determined by the physician
285 or medical team, can be considered as the *prior* probability. Using hemogram data as input,
286 and informing the *prior* probability of COVID-19 based on medical findings, the model will
287 consider hemogram data to inform a *posterior* probability, which can be higher or lower than the
288 original, and based on the hemogram alterations caused by the virus infection. It is important that
289 hemogram data would not be included in original medical assessment and *prior* determination,
290 in order to avoid bias and reduce model overfit.

Multiple Patients Evaluation

292 It can be used in situations where decisions are necessary for resource management including
293 multiple individuals. Choice of a target group (positive or negative qRT-PCR result prediction)
294 should be defined. The model can be applied to multiple individuals simultaneously, with
295 the choice of *prior* probability carefully adjusted to result in a specific number of predicted
296 individuals from the target group, according to the desired outcome. This method increases
297 the correct selection of candidates belonging to the target group, when compared to random
298 selection. Individual results can be ranked based on *posterior* probability of a positive or negative
299 result, and results stratified according to convenience, as a way to evaluate a particular scenario
300 of interest. When additional clinical data is available, or become available later, patients selected
301 during bulk evaluation should be reassessed individually as proposed in the general framework,
302 in order to reduce misclassifications.

Applications to Scarcity Scenarios:

304 Examples of proposed model use are presented for some specific scarcity scenarios in Table 2.
305 As can be seen, the model sensitivity can be adjusted by selecting *prior* probability employed,
306 according to desired outcome or interest group. *Prior* selection should be carefully decided,
307 based on current context or situation proposed, and must consider the classification group where
308 higher accuracy is intended.
309 High accuracy in qRT-PCR result prediction is achieved based on hemogram information
310 only. Further analysis performed on the original data (not shown) suggest that additional clinical
311 results can improve prediction efficiency. This conclusion is in accordance with previous findings

Table 2. Strategies for NB-ML model applications and symptomatic patient selection in scarcity conditions. Hemogram test results are available for all symptomatic patients. Scenarios proposed for situations where test results are not available (no testing or waiting qRT-PCR test results). Prediction results were appraised in a binary form, with positive or negative classification based on *posterior* probability threshold of 0.5. Results are presented in reference to random patient selection.

Condition	Context example	Objective	Strategy	Action	Starting / fixed *prior*	Results in training set (positive misclassified among cleared)	Results in missing-data set (positive misclassified among cleared)
Testing shortage	Testing capacity is limited to a third of candidates only	Maximize number of infected patients tested	Prioritize TP identification	Fine-tune *prior* until positive reach testing capacity	0.5	130% increase in actual infected patients tested (*prior*=0.3482)	100% increase in actual infected patients tested (*prior*=0.9706)
Lack of essential workforce	Professionals with high risk of nocosomial or work-related transmission	Keep symptomatic, non-infected essential workers in duty	Search for evident non-infected workers (TN identification)	All workers are considered as infected, unless model says otherwise	0.9999	19.4% of total workforce cleared (0%)	52% of total workforce cleared (6.25%)
Lack of essential workforce	Professionals with medium to low risk of transmission	Keep symptomatic, non-infected essential workers in duty	Find ideal balance to simultaneously, maximize both TN and TP	Use intersection of sensitivity and specificity curves from training set	0.2933	69.0% of total workforce cleared (5%)	81.5% of total workforce cleared (6.6%)
Limited medical access	Medical assistance limited to 20% of symptomatic individuals only	Avoid contagion exposure of non-infected patients in ER during medical assistance	Eliminate non-infected from candidates for medical assistance (TN identification)	Fine-tune *prior* to select most likely negative results. Select remaining set for medical assistance	0.5	35.6% decrease in non-infected patients exposure (*prior*=0.0954)	18.8% decrease in non-infected patients exposure (*prior*=0.4888)

TP: True Positive; TN: True Negative

312 suggesting biochemical and immunological abnormalities, in addition to hematologic alterations,
313 can be caused by COVID-19 disease (Henry et al., 2020). In this context, the relevance of data
314 employed to generate ML models is emphasized. The use of large and comprehensive datasets,
315 containing as much information as possible regarding clinical and laboratory findings, symptoms,
316 disease evolution, and other relevant aspects, is crucial in devising useful and adequate models.
317 The development of nationwide or regional databases based on local data is essential, in order to
318 capture epidemiological idiosyncrasies associated with such populations (Terpos et al., 2020).
319 Also, natural differences in hemogram results from distinct demographic groups (as seen in
320 reference values according to age, sex, or other physiological factors) can aggregate noise to
321 the model, which can be reduced when large database are employed in model construction, and
322 results can be devised for each ethnographic strata.

323 Despite having high overall accuracy, performance metrics obtained with proposed model
324 show unequal ability to predict positive or negative results. This situation is caused by a
325 significant imbalance in number of samples belonging to each of this qRT-PCR result groups
326 in original data. The use of balanced data in machine learning model design is important to
327 assure high prediction quality (Krawczyk, 2016). The option of maintaining original data in
328 model construction was adopted, since it better represents actual COVID-19 prevalence among
329 symptomatic patients, and therefore seems to represent a more realistic situation. Additional
330 simulations applying a balanced model (data not shown) using positive group oversampling
331 (to compensate its insufficiency in original data) have devised alternative models with superior
332 predictive power. Alternative balanced model results are presented in Supplementary Material
333 Figure S1.Therefore, additional positive samples will be added to the data and used in future
334 model versions.

335 As a perspective, collection of hemogram results from asymptomatic patients (in addition to
336 symptomatic individuals) can be used to evaluate the utility of this approach on the detection
337 of asymptomatic infections, in order to provide alternatives in diagnostics, especially in a
338 context of testing deficiency. A web-based application was developed by the authors, in which
339 hemogram data can be introduced for a single individual, along with prior probability of infection,
340 based on data used to generate the present model. The online tool is available at `http:`
341 `//sbcb.inf.ufrgs.br/covid`. Future implementation will allow the upload of multiple
342 patients simultaneously, and construction or testing of user data-derived models. This service
343 will allow easy access and practical application of the proposed model.

CONCLUSION

345 Machine learning models based on hemogram data can be employed in COVID-19 pandemics
346 management, in order to assist strategical medical decisions in different scarcity scenarios. The
347 proposed Naïve-Bayes model has the flexibility to be applied in a large variety of possible
348 critical conditions, and can be adjusted to improve classification accurracy for a particular target
349 group. Even though the method proposed in this work is not suitable to be used as a diagnostic
350 technique, it can be employed to provide aditional, useful information regarding data-driven
351 resource allocation, in shortage conditions.

REFERENCES

353 Adhikari, S. P., Meng, S., Wu, Y.-J., Mao, Y.-P., Ye, R.-X., Wang, Q.-Z., Sun, C., Sylvia, S.,
354 Rozelle, S., Raat, H., and Zhou, H. (2020). Epidemiology, causes, clinical manifestation and
355 diagnosis, prevention and control of coronavirus disease (covid-19) during the early outbreak
356 period: a scoping review. *Infectious Diseases of Poverty*, 9(1-12):29.

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the covid-19 epidemic? *The Lancet*, 395(10228):931–934.

Black, J. R. M., Bailey, C., Przewrocka, J., Dijkstra, K. K., and Swanton, C. (2020). Covid-19: the case for health-care worker screening to prevent hospital transmission. *The Lancet*, n/a(n/a):1–2. ahead of print.

Brown, C. S., Ravallion, M., and van de Walle, D. (2020). Can the world's poor protect themselves from the new coronavirus? Working Paper 27200, National Bureau of Economic Research.

Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R., Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C., and O'Connell, P. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet*, 362(9381):362–369.

Ding, X., Yu, Y., Lu, B., Huo, J., Chen, M., Kang, Y., Lou, J., and Liu, Z. (2020). Dynamic profile and clinical implications of hematological parameters in hospitalized patients with coronavirus disease 2019. *Clinical Chemistry and Laboratory Medicine*, n/a(n/a):20200411.

Emanuel, E. J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C., Boyle, C., Smith, M., and Phillips, J. P. (2020). Fair allocation of scarce medical resources in the time of covid-19. *New England Journal of Medicine*, n/a(n/a):1–7. Sounding Board.

Fan, B. E., Ong, K. H., Chan, S. S. W., Young, B. E., Chong, V. C. L., Chen, S. P. C., Lim, S. P., Lim, G. P., and Kuperan, P. (2020). Blood and blood product use during covid-19 infection. *American Journal of Hematology*, n/a(n/a):1–3.

Gasmi, A., Noor, S., Tippairote, T., Dadar, M., Menzel, A., and Bjorklund, G. (2020). Individual risk management strategy and potential therapeutic options for the covid-19 pandemic. *Clinical Immunology*, n/a(n/a):1–37. ahead of print.

Gorry, G. and Barnett, G. (1968). Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, 1(5):490 – 507.

Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D. S., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y., Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P., Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y., and Zhong, N.-s. (2020). Clinical characteristics of coronavirus disease 2019 in china. *New England Journal of Medicine*, n/a(n/a):1–13. ahead of print.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*. O'Reilly, 1 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition.

Henry, B. M., Benoit, S. W., [de Oliveira], M. H. S., Hsieh, W. C., Benoit, J., Ballout, R. A., Plebani, M., and Lippi, G. (2020). Laboratory abnormalities in children with mild and severe coronavirus disease 2019 (covid-19): A pooled analysis and review. *Clinical Biochemistry*.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, 395(10223):497–506.

Kandel, N., Chungong, S., Omaar, A., and Xing, J. (2020). Health security capacities in the context of covid-19 outbreak: an analysis of international health regulations annual report data from 182 countries. *The Lancet*, 395(10229):1047–1053.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions.

Progress in Artificial Intelligence, 5(4):221–232.

Lippi, G. and Plebani, M. (2020). The critical role of laboratory medicine during coronavirus disease 2019 (covid-19) and other viral outbreaks. *Clinical Chemistry and Laboratory Medicine*, n/a(n/a):1–7. ahead of print.

Lipsitch, M., Swerdlow, D. L., and Finelli, L. (2020). Defining the epidemiology of covid-19 - studies needed. *New England Journal of Medicine*, 382(13):1194–1196.

Martin, W., Apostolakos, P., and Roazen, H. (1960). Clinical versus acturial prediction in the differential diagnosis of jaundice. a stdy of the relative accuracy of predictions made by physicians and by a statistically derived formula in differentiating parenchymal and obstructive jaundice. *The American Journal of the Medical Sciences*, 240:571–578.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York, 1 edition.

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., and Agha, R. (2020). The socio-economic implications of the coronavirus and covid-19 pandemic: A review. *International Journal of Surgery*, n/a(n/a):1–31. ahead of print.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Ranney, M. L., Griffeth, V., and Jha, A. K. (2020). Critical supply shortages - the need for ventilators and personal protective equipment during the covid-19 pandemic. *New England Journal of Medicine*, Perspective(n/a):1–3.

Schurink, C., Lucas, P., Hoepelman, I., and Bonten, M. (2005). Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. *The Lancet Infectious Diseases*, 5(5):305–312.

Sethuraman, N., Jeremiah, S. S., and Ryo, A. (2020). Interpreting Diagnostic Tests for SARS-CoV-2. *JAMA*, n/a(n/a):200101. ahead of print.

Shimoni, Z., Glick, J., and Froom, P. (2013). Clinical utility fo the full blood count in identifying patients with pandemic influenza a (h1n1). *Journal of Infection*, 66(6):545–547.

Tahamtan, A. and Ardebili, A. (2020). Real-time rt-pcr in covid-19 detection: issues affecting the results. *Expert Review of Molecular Diagnostics*, n/a(n/a):1–2.

Tanne, J. H., Hayasaki, E., Zastrow, M., Pulla, P., Smith, P., and Rada, A. G. (2020). Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. *BMJ*, 368:1–5.

Terpos, E., Ntanasis-Stathopoulos, I., Elalamy, I., Kastritis, E., Sergentanis, T. N., Politou, M., Psaltopoulou, T., Gerotziafas, G., and Dimopoulos, M. A. (2020). Hematological findings and complications of covid-19. *American Journal of Hematology*, n/a(n/a):1–32. ahead of print.

Tu, H., Tu, S., Gao, S., Shao, A., and Sheng, J. (2020). The epidemiological and clinical features of covid-19 and lessons from this global infectious public health event. *The Journal of Infection*, n/a(n/a):1–20. ahead of print.

Wynants, L., Van Calster, B., Bonten, M. M. J., Collins, G. S., Debray, T. P. A., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G. M., Riley, R. D., Schuit, E., Smits, L. J. M., Snell, K. I. E., Steyerberg, E. W., Wallisch, C., and van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ*, 369:1–11.