

Hemogram data as a tool for decision-making in COVID-19 management: Applications to resource scarcity scenarios (#48815)

1

First submission

Guidance from your Editor

Please submit by **30 May 2020** for the benefit of the authors (and your \$200 publishing discount) .



Structure and Criteria

Please read the 'Structure and Criteria' page for general guidance.



Author notes

Have you read the author notes on the [guidance page](#)?



Raw data check

Review the raw data.



Image check

Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the [materials page](#).




Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. BASIC REPORTING
2. EXPERIMENTAL DESIGN
3. VALIDITY OF THE FINDINGS
4. General comments
5. Confidential notes to the editor






 You can also annotate this PDF and upload it as part of your review

When ready [submit online](#).





Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your [guidance page](#).





BASIC REPORTING

-  Clear, unambiguous, professional English language used throughout.
-  Intro & background to show context. Literature well referenced & relevant.
-  Structure conforms to [Peerj standards](#), discipline norm, or improved for clarity.
-  Figures are relevant, high quality, well labelled & described.
-  Raw data supplied (see [Peerj policy](#)).

EXPERIMENTAL DESIGN

-  Original primary research within [Scope of the journal](#).
-  Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
-  Rigorous investigation performed to a high technical & ethical standard.
-  Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

-  Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
-  All underlying data have been provided; they are robust, statistically sound, & controlled.
-  Speculation is welcome, but should be identified as such.
-  Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

3



The best reviewers use these techniques

Tip

Support criticisms with evidence from the text or from other sources

Example

Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.

Give specific suggestions on how to improve the manuscript

Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).

Comment on language and grammar issues

The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.

Organize by importance of the issues, and number your points

1. Your most important issue
2. The next most important item
3. ...
4. The least important points

Please provide constructive criticism, and avoid personal opinions

I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC

Comment on strengths (as well as weaknesses) of the manuscript

I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.

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

Eduardo Avila^{Corresp., 1, 2, 3}, Marcio Dorn^{2, 4}, Clarice Sampaio Alho^{1, 2}, Alessandro Kahmann^{2, 5}

¹ School of Health and Life Sciences, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Brazil

² National Institute of Science and Technology - Forensic Sciences, Porto Alegre, Rio Grande do Sul, Brazil

³ Technical Scientific Section, Federal Police, Porto Alegre, Rio Grande do Sul, Brazil

⁴ Institute of Informatics, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

⁵ Institute of Mathematics, Statistics and Physics, Universidade Federal do Rio Grande, Rio Grande, 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: A Naïve-Bayes model for machine learning is proposed for handling different scarcity scenarios, including managing symptomatic essential workforce and absence of diagnostic tests. Hemogram result data was used to predict qRT-PCR results in situations where the latter was not performed, or results are not yet available. 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. Data assessment can be performed in an individual or simultaneous basis, according to desired outcome. Based on hemogram data and background scarcity context, resource distribution is significantly optimized when model-based patient selection is observed, compared to random choice. The model can help manage testing deficiency and other critical circumstances.

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,*}, Márcio Dorn^{3,4,*}, Clarice Sampaio Alho^{1,3}, and Alessandro Kahmann^{3,5,*}

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

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

³National Institute of Science and Technology - Forensic Science, Porto Alegre, RS, Brazil

⁴Laboratory of Structural Bioinformatics and Computational Biology, Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

⁵Institute of Mathematics, Statistics and Physics, Federal University of Rio Grande, Rio Grande, 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. 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: A *Naïve-Bayes* model for machine learning is proposed for handling different scarcity scenarios, including managing symptomatic essential workforce and absence of diagnostic tests. Hemogram result data was used to predict qRT-PCR results in situations where the latter was not performed, or results are not yet available. 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. Data assessment can be performed in an individual or simultaneous basis, according to desired outcome. Based on hemogram data and background scarcity context, resource distribution is significantly optimized when model-based patient selection is observed, compared to random choice. The model can help manage testing deficiency and other critical circumstances.

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 Brazil, 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 **number** of missing information is significant. All **variables were normalized** to maintain anonymity and remove scale effects. No missing data imputation was performed during model generation to avoid *bias*. Considering the significant amount 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. The full dataset is available in <https://www.kaggle.com/einsteindata4u/covid19>.

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 machine learning models 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**. 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 the *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 their low computational cost and **clear interpretation**. 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. **Once** 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

141 takes an input array (test) and makes predictions for each sample in it. In supervised learning
 142 estimators (our case), this method returns the predicted labels or values computed from the
 143 estimated model (positive or negative for COVID-19). *Cross-validation* is a model evaluation
 144 method that allows one to evaluate an estimator on a given dataset reliably. It consists of
 145 iteratively fitting the estimator on a fraction of the data, called training set, and testing it on the
 146 left-out unseen data, called test set. Several strategies exist to partition the data. In this work, we
 147 used the *Leave-one-out* (LOO) cross-validation model, as in Chang et al. (Chang et al., 2003).
 148 The number of data points was split N times (number samples). The method was trained on all
 149 the data except for one point, and a prediction was made for that point. The proposed approach
 150 was implemented in *Python v.3* (<https://www.python.org>) code using *Scikit-Learn v.*
 151 *0.22.2* (Pedregosa et al., 2011) as a backend.
 152 *Model Test:* In order to evaluate the adequacy and generalization power of the proposed model, a
 153 set of 92 samples (10 positives for COVID-19 and 82 negatives) was extracted from the patient
 154 database. Those samples were not initially employed in model delineation, considering they
 155 present a single missing value among all 15 employed hemogram parameters. Missing data
 156 for this training set was imputed using the average value of the missing parameter within the
 157 resulting group (positive or negative). The test set was submitted to the previously generated
 158 model in order to evaluate classification performance.

159 RESULTS

160 Descriptive Analysis

161 For data description, probability density function (PDF) of all 15 hemogram parameters were
 162 estimated through the original sample by kernel density estimator. Some hemogram parameters
 163 present notable differences between the distributions of positive and negative results, mainly
 164 regarding its modal value (distribution peak value) and variance (distribution width). Differences
 165 are summarized in Table 1. Regarding basophiles, eosinophils, leukocytes and platelets counts,
 166 qRT-PCR positive group distribution shows lower modal value and lower variance. On the
 167 other hand, monocyte count displays opposite behavior, once lower modal value and variance
 168 are observed for the qRT-PCR positive group. Lower variance may depict a condition pattern,
 169 therefore it is expected that negative cases present higher variance once it may contain a higher
 170 variety of conditions (reasons for symptom presence). The remaining nine hemogram parameters
 171 did not show a notable difference between negative and positive groups. PDF analysis results are
 172 presented in Supplementary Material Figure S1.

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		

173 Naïve Bayes Model Results

174 A NB classifier based on training set hemogram data was developed. Under the model, the
 175 complete range of *prior* probabilities (from 0.0001 to 0.9999 by 0.0001 increments) was scruti-
 176 nized, and *posterior* probability of each class was computed for different *prior* conditions. A

177 *posterior* probability value of 0.5 was defined as the classification threshold in one of the positive
 178 or negative predicted groups. Resulting model showed a good predictive power of the qRT-PCR
 179 test result based on hemogram data. Figure 1 shows the accuracy, sensitivity, and specificity
 180 curves derived from the model for different *prior* probabilities of each class (positive or negative
 181 for COVID-19). Reported *prior* probabilities refer to positive COVID-19 condition.

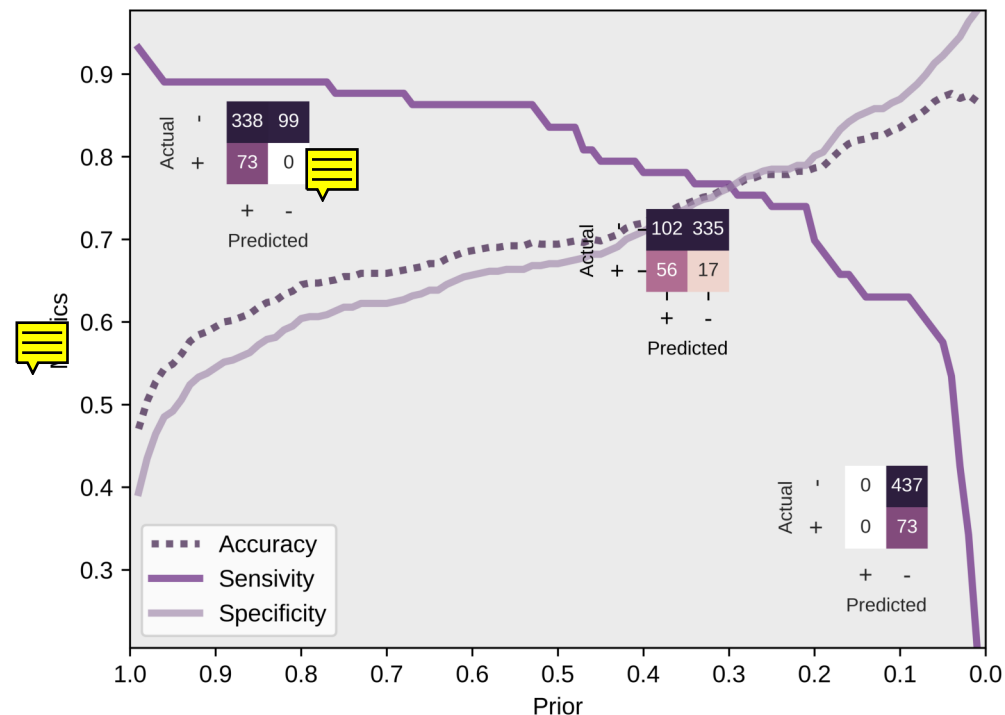


Figure 1. Performance metrics of proposed Naive-Bayes model. *prior* probabilities are presented in reference to positive qRT-PCR prediction. Confusion matrices (left to right) are presented for 0.9999, 0.2933 and 0.0001 *prior* probabilities, respectively. Sensitivity=True Positive Ratio; Specificity=True Negative Ratio

182 When setting the *prior* probability to the maximum defined value (0.9999), the NB classifier
 183 correctly diagnosed all PCR positive cases. On the other hand, such configuration improperly
 184 predicted 77.3% of negative PCR results as positive. Regarding the lower possible *prior*
 185 probability setting, it does not classify a single observation as positive. This result can be
 186 explained by the unbalanced number of observations for each class, tending to over classify
 187 samples as the class with more observation, i.e. negative results. Such characteristics can also
 188 be noticed in the general accuracy, since the decrease in the *prior* ponce the classifier tends to
 189 diagnose all observations as belonging to the dominant class (negative) and consequently raising
 190 the total of correctly classified samples. The break-even point is met when *prior* probability is
 191 set to 0.2933. Under this condition, all metrics are approximately 76.6%.

192 Regarding the model sensitivity, the rate of positive samples correctly classified is over 85%
 193 within 0.999 to 0.5276 range, with small decrease of it when the *prior* probability of positive
 194 result is diminished within this range. When *prior* is set to under 0.0606, the number of positive
 195 predicted samples decrease rapidly, yielding lower sensitivity. As for specificity, it presents
 196 linear growth as tested *priors* decrease. Ultimately, the accuracy results profile are similar to

specificity, due to the negative patients dominance.

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 favoured, 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. Test samples (n=92, including 10 qRT-PCR positives) were used to test the proposed model. Figure 2 presents results obtained from the model application to test dataset.

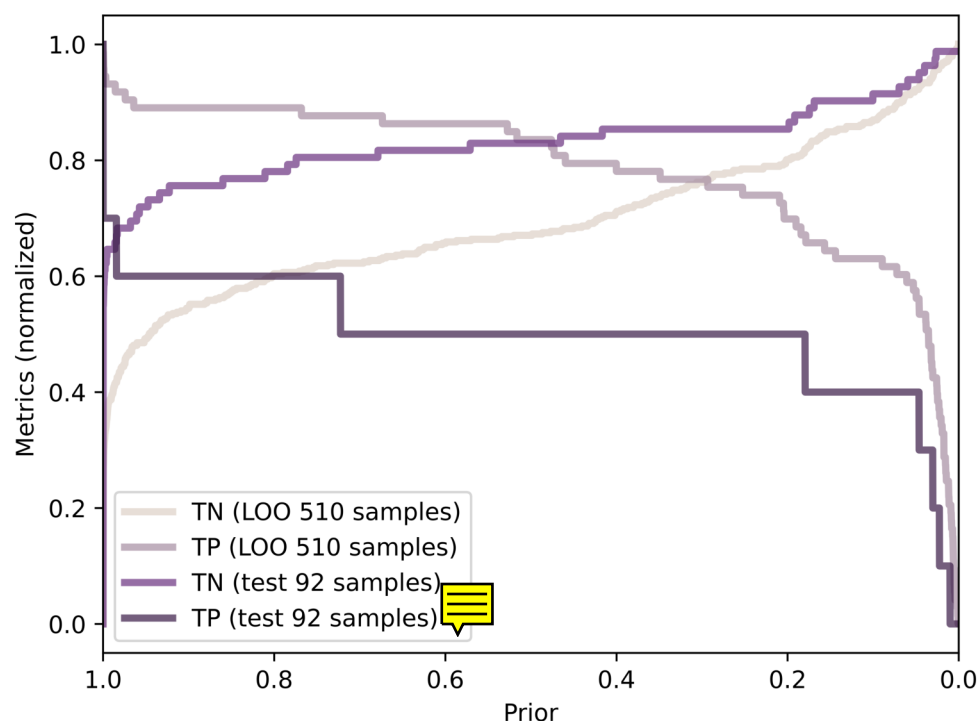


Figure 2. Classification performance for training (LOO) and test datasets. Results presented for the complete *prior* probability range. Results are presented as the percentage of correctly predicted qRT-PCR exams. Informed *prior* probability refers to positive outcome. TN: true negative; TP: true positive.

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. Figure 3 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 dependable on factors affecting qRT-PCR efficiency, and its prediction capacity is dependent on the sensitivity, accuracy, and specificity of the original laboratory exam (Sethuraman et al., 2020).

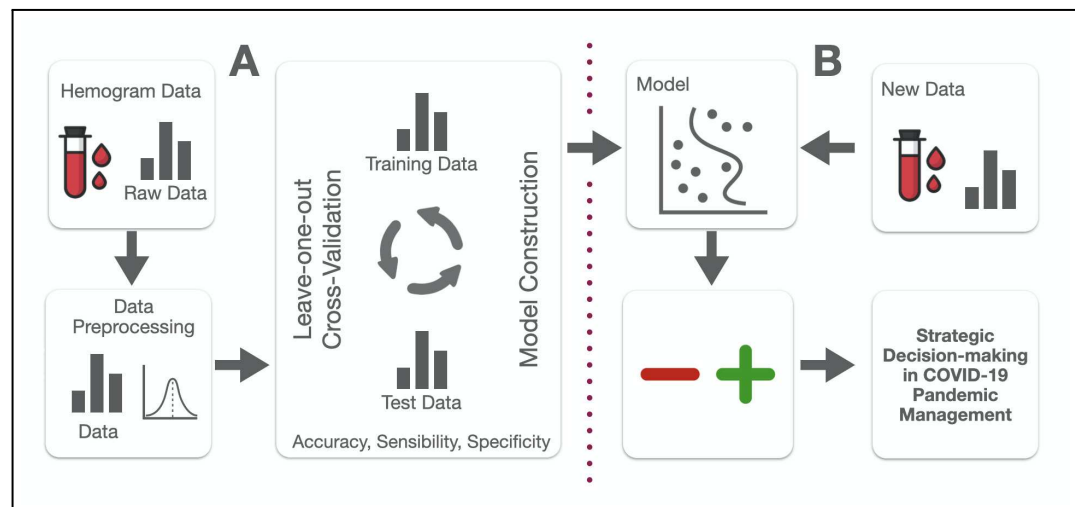


Figure 3. NB Model construction and application diagram.

Descriptive analysis of hemogram clinical findings shows differences in blood cell counts and other hematological parameters among COVID-19 positive and negative patient results. Differences are conspicuous among three measures (leukocytes, monocytes and platelets) and more discrete to additional two (basophiles and eosinophiles). It is possible that differences are also present across the complete data spectrum, even though they are not clearly visualized with PDF data. These results are in accordance to previous reports of changes in laboratory findings in COVID-19 infected patients, where conditions as leukopenia, lymphocytopenia and thrombopenia were reported (Fan et al., 2020). It is important to highlight that data analysis is not sufficient to characterize clinical hematological alterations in evaluated patients (when compared to demographic hematologic parameters data), once data was normalized for the evaluated sample set only. However, even within this particular quota of population (individuals presenting COVID-19-like symptoms), differences were found between individuals presenting negative or positive qRT-PCR COVID test results. The proposed NB-ML model can be helpful in accessing different levels of information from hemogram results, through inferring non-evident patterns and parameter relationships from this data.

Bayesian techniques are based on the choice of a *prior* probability of an event (in present case, positive result for qRT-PCR test). The method considers actual evidence (hemogram data) to result in a *posterior* probability of the outcome (prediction of a positive result). By changing the selected *prior* probability, we can derive an uncertainty analysis of the model to understand its distribution. Uncertainty can be then applied to adequately adapt the classifier to a particular ongoing context. This option allows the evaluation of different decision-making scenarios concerning diverse aspects of pandemics management. During a crisis situation, measures should

247 be taken seeking to maximize benefits and achieve a fair resource allocation (Emanuel et al.,
248 2020). To illustrate the model flexibility and how it can be used to help on this matter, a general
249 framework of application is proposed, followed by a simulation of four scenarios where resource
250 scarcity is assumed.

251 **Application Framework**

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

258 **Individual Assessment**

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

270 **Multiple Patients Evaluation**

271 It can be used in situations where decisions are necessary for resource management including
272 multiple individuals. Choice of a target group (positive or negative qRT-PCR result prediction)
273 should be defined. The model can be applied to multiple individuals simultaneously, with
274 the choice of *prior* probability carefully adjusted to result in a specific number of predicted
275 individuals from the target group, according to the desired outcome. This method increases
276 the correct selection of candidates belonging to the target group, when compared to random
277 selection. When additional clinical data is available, or become available later, patients selected
278 during bulk evaluation should be reassessed individually as proposed in the general framework,
279 in order to reduce misclassifications.

280 **Applications to Scarcity Scenarios:**

281 Examples of proposed model use are presented for some specific scarcity scenarios in Table 2.
282 As can be seen, the model sensitivity can be adjusted by selecting *prior* probability employed,
283 according to desired outcome or interest group. *prior* selection should be carefully decided,
284 based on current context or situation proposed, and must consider the classification group where
285 higher accuracy is intended.

286 High accuracy in qRT-PCR result prediction is achieved based on hemogram information
287 only. Further analysis performed on the original data (not shown) suggest that additional clinical
288 results can improve prediction efficiency. This conclusion is in accordance with previous findings
289 suggesting biochemical and immunological abnormalities, in addition to hematologic alterations,
290 can be caused by COVID-19 disease (Henry et al., 2013). In this context, the relevance of data
291 employed to generate ML models is emphasized. The use of large and comprehensive datasets,
292 containing as much information as possible regarding clinical and laboratory findings, symptoms,
293 disease evolution, and other relevant aspects, is crucial in devising useful and adequate models.

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 test 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 <i>prior</i> until positive reach testing capacity	0.5	130% increase in actual infected patients tested (<i>prior</i> =0.3482)	100% increase in actual infected patients tested (<i>prior</i> =0.9607)
Lack of essential workforce	Professionals with high risk of nosocomial 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%)	50% of total workforce cleared (6.5%)
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 <i>prior</i> to select most likely negative results. Select remaining set for medical assistance	0.5	35.6% decrease in non-infected patients exposure (<i>prior</i> =0.0954)	18.8% decrease in non-infected patients exposure (<i>prior</i> =0.4652)

TP: True Positive; TN: True Negative

294 The development of nationwide or regional databases based on local data is essential, in order to
295 capture epidemiological idiosyncrasies associated with such populations (Terpos et al., 2020).
296 Also, natural differences in hemogram results from distinct demographic groups (as seen in
297 reference values according to age, sex, or other physiological factors) can aggregate noise to
298 the model, which can be reduced when large database are employed in model construction, and
299 results can be devised for each ethnographic strata.

300 Despite having high overall accuracy, performance metrics obtained with proposed model
301 show unequal ability to predict positive or negative results. This situation is caused by a
302 significant imbalance in number of samples belonging to each of this qRT-PCR result groups
303 in original data. The use of balanced data in machine learning model design is important to
304 assure high prediction quality (Krawczyk, 2016). The option of maintaining original data in
305 model construction was adopted, since it better represents actual COVID-19 prevalence among
306 symptomatic patients, and therefore seems to represent a more realistic situation. Additional
307 simulations applying a balanced model (data not shown) using positive group oversampling
308 (to compensate its insufficiency in original data) have devised alternative models with superior
309 prediction power. Alternative balanced model results are presented in Supplementary Material
310 Figure S2. Therefore, additional positive samples will be added to the data and used in future
311 model versions.

312 As a perspective, collection of hemogram results from asymptomatic patients (in addition to
313 symptomatic individuals) can be used to evaluate the utility of this approach on the detection
314 of asymptomatic infections, in order to provide alternatives in diagnostics, especially in a
315 context of testing deficiency. A web-based application was developed by the authors, in which
316 hemogram data can be introduced for a single individual, along with prior probability of infection,
317 based on data used to generate the present model. The online tool is available at <http://sbcbr.inf.ufrgs.br/covid>. Future implementation will allow the upload of multiple
318 patients simultaneously, and construction or testing of user data-derived models. This service
319 will allow easy access and practical application of the proposed model.
320

321 REFERENCES

- 322 Adhikari, S. P., Meng, S., Wu, Y.-J., Mao, Y.-P., Ye, R.-X., Wang, Q.-Z., Sun, C., Sylvia, S.,
323 Rozelle, S., Raat, H., and Zhou, H. (2020). Epidemiology, causes, clinical manifestation and
324 diagnosis, prevention and control of coronavirus disease (covid-19) during the early outbreak
325 period: a scoping review. *Infectious Diseases of Poverty*, 9(1-12):29.
326 Anderson, R. M., Heesterbeek, H., Klinkenberg, D., and Hollingsworth, T. D. (2020). How
327 will country-based mitigation measures influence the course of the covid-19 epidemic? *The*
328 *Lancet*, 395(10228):931–934.
329 Black, J. R. M., Bailey, C., Przewrocka, J., Dijkstra, K. K., and Swanton, C. (2020). Covid-
330 19: the case for health-care worker screening to prevent hospital transmission. *The Lancet*,
331 n/a(n/a):1–2. ahead of print.
332 Chang, J. C., Wooten, E. C., Tsimelzon, A., Hilsenbeck, S. G., Gutierrez, M. C., Elledge, R.,
333 Mohsin, S., Osborne, C. K., Chamness, G. C., Allred, D. C., and O’Connell, P. (2003). Gene
334 expression profiling for the prediction of therapeutic response to docetaxel in patients with
335 breast cancer. *The Lancet*, 362(9381):362–369.
336 Emanuel, E. J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., Zhang, C., Boyle,
337 C., Smith, M., and Phillips, J. P. (2020). Fair allocation of scarce medical resources in the
338 time of covid-19. *New England Journal of Medicine*, n/a(n/a):1–7. Sounding Board.
339 Fan, B. E., Ong, K. H., Chan, S. S. W., Young, B. E., Chong, V. C. L., Chen, S. P. C., Lim, S. P.,

- 340 Lim, G. P., and Kuperan, P. (2020). Blood and blood product use during covid-19 infection.
341 *American Journal of Hematology*, n/a(n/a):1–3. corresponde.
- 342 Gasmi, A., Noor, S., Tippairrote, T., Dadar, M., Menzel, A., and Bjorklund, G. (2020). Individual
343 risk management strategy and potential therapeutic options for the covid-19 pandemic. *Clinical*
344 *Immunology*, n/a(n/a):1–37. ahead of print.
- 345 Gorry, G. and Barnett, G. (1968). Experience with a model of sequential diagnosis. *Computers*
346 *and Biomedical Research*, 1(5):490 – 507.
- 347 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,
348 D. S., Du, B., Li, L.-j., Zeng, G., Yuen, K.-Y., Chen, R.-c., Tang, C.-l., Wang, T., Chen, P.-y.,
349 Xiang, J., Li, S.-y., Wang, J.-l., Liang, Z.-j., Peng, Y.-x., Wei, L., Liu, Y., Hu, Y.-h., Peng, P.,
350 Wang, J.-m., Liu, J.-y., Chen, Z., Li, G., Zheng, Z.-j., Qiu, S.-q., Luo, J., Ye, C.-j., Zhu, S.-y.,
351 and Zhong, N.-s. (2020). Clinical characteristics of coronavirus disease 2019 in china. *New*
352 *England Journal of Medicine*, n/a(n/a):1–13. ahead of print.
- 353 Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts,*
354 *tools, and techniques to build intelligent systems*. O'Reilly, 1 edition.
- 355 Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data*
356 *mining, inference and prediction*. 2 edition.
- 357 Henry, B. M., de Oliveira, M. H. S., Benoit, S., Plebani, M., and Lippi, G. (2013). Hematologic,
358 biochemical and immune biomarker abnormalities associated with severe illness and mortality
359 in coronavirus disease 2019 (covid-19): a meta-analysis. *Clinical Chemistry and Laboratory*
360 *Medicine (CCLM)*, n/a(n/a):20200369. ahead of print.
- 361 Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X.,
362 Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao,
363 H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., and Cao, B. (2020).
364 Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*,
365 395(10223):497–506.
- 366 Kandel, N., Chungong, S., Omaar, A., and Xing, J. (2020). Health security capacities in the
367 context of covid-19 outbreak: an analysis of international health regulations annual report
368 data from 182 countries. *The Lancet*, 395(10229):1047–1053.
- 369 Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions.
370 *Progress in Artificial Intelligence*, 5(4):221–232.
- 371 Lippi, G. and Plebani, M. (2020). The critical role of laboratory medicine during coronavirus
372 disease 2019 (covid-19) and other viral outbreaks. *Clinical Chemistry and Laboratory*
373 *Medicine*, n/a(n/a):1–7. ahead of print.
- 374 Lipsitch, M., Swerdlow, D. L., and Finelli, L. (2020). Defining the epidemiology of covid-19 -
375 studies needed. *New England Journal of Medicine*, 382(13):1194–1196.
- 376 Martin, W., Apostolakos, P., and Roazen, H. (1960). Clinical versus actuarial prediction in
377 the differential diagnosis of jaundice. a study of the relative accuracy of predictions made by
378 physicians and by a statistically derived formula in differentiating parenchymal and obstructive
379 jaundice. *The American Journal of the Medical Sciences*, 240:571–578.
- 380 Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York, 1 edition.
- 381 Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M., and Agha,
382 R. (2020). The socio-economic implications of the coronavirus and covid-19 pandemic: A
383 review. *International Journal of Surgery*, n/a(n/a):1–31. ahead of print.
- 384 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
385 Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher,
386 M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal*
387 *of Machine Learning Research*, 12:2825–2830.
- 388 Ranney, M. L., Griffeth, V., and Jha, A. K. (2020). Critical supply shortages - the need for

- 389 ventilators and personal protective equipment during the covid-19 pandemic. *New England*
390 *Journal of Medicine*, Perspective(n/a):1–3.
- 391 Schurink, C., Lucas, P., Hoepelman, I., and Bonten, M. (2005). Computer-assisted decision
392 support for the diagnosis and treatment of infectious diseases in intensive care units. *The*
393 *Lancet Infectious Diseases*, 5(5):305–312.
- 394 Sethuraman, N., Jeremiah, S. S., and Ryo, A. (2020). Interpreting Diagnostic Tests for SARS-
395 CoV-2. *JAMA*, n/a(n/a):200101. ahead of print.
- 396 Shimoni, Z., Glick, J., and Froom, P. (2013). Clinical utility fo the full blood count in identifying
397 patients with pandemic influenza a (h1n1). *Journal of Infection*, 66(6):545–547.
- 398 Tahamtan, A. and Ardebili, A. (2020). Real-time rt-pcr in covid-19 detection: issues affecting
399 the results. *Expert Review of Molecular Diagnostics*, Editorial(n/a):1–2.
- 400 Tanne, J. H., Hayasaki, E., Zastrow, M., Pulla, P., Smith, P., and Rada, A. G. (2020). Covid-19:
401 how doctors and healthcare systems are tackling coronavirus worldwide. *BMJ*, 368:1–5.
- 402 Terpos, E., Ntanasis-Stathopoulos, I., Elalamy, I., Kastritis, E., Sergentanis, T. N., Politou, M.,
403 Psaltopoulou, T., Gerotziafas, G., and Dimopoulos, M. A. (2020). Hematological findings and
404 complications of covid-19. *American Journal of Hematology*, n/a(n/a):1–32. ahead of print.
- 405 Tu, H., Tu, S., Gao, S., Shao, A., and Sheng, J. (2020). The epidemiological and clinical
406 features of covid-19 and lessons from this global infectious public health event. *The Journal*
407 *of Infection*, n/a(n/a):1–20. ahead of print.
- 408 Wynants, L., Van Calster, B., Bonten, M. M. J., Collins, G. S., Debray, T. P. A., De Vos, M.,
409 Haller, M. C., Heinze, G., Moons, K. G. M., Riley, R. D., Schuit, E., Smits, L. J. M., Snell,
410 K. I. E., Steyerberg, E. W., Wallisch, C., and van Smeden, M. (2020). Prediction models for
411 diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ*,
412 369:1–11.

413

Supplementary Material

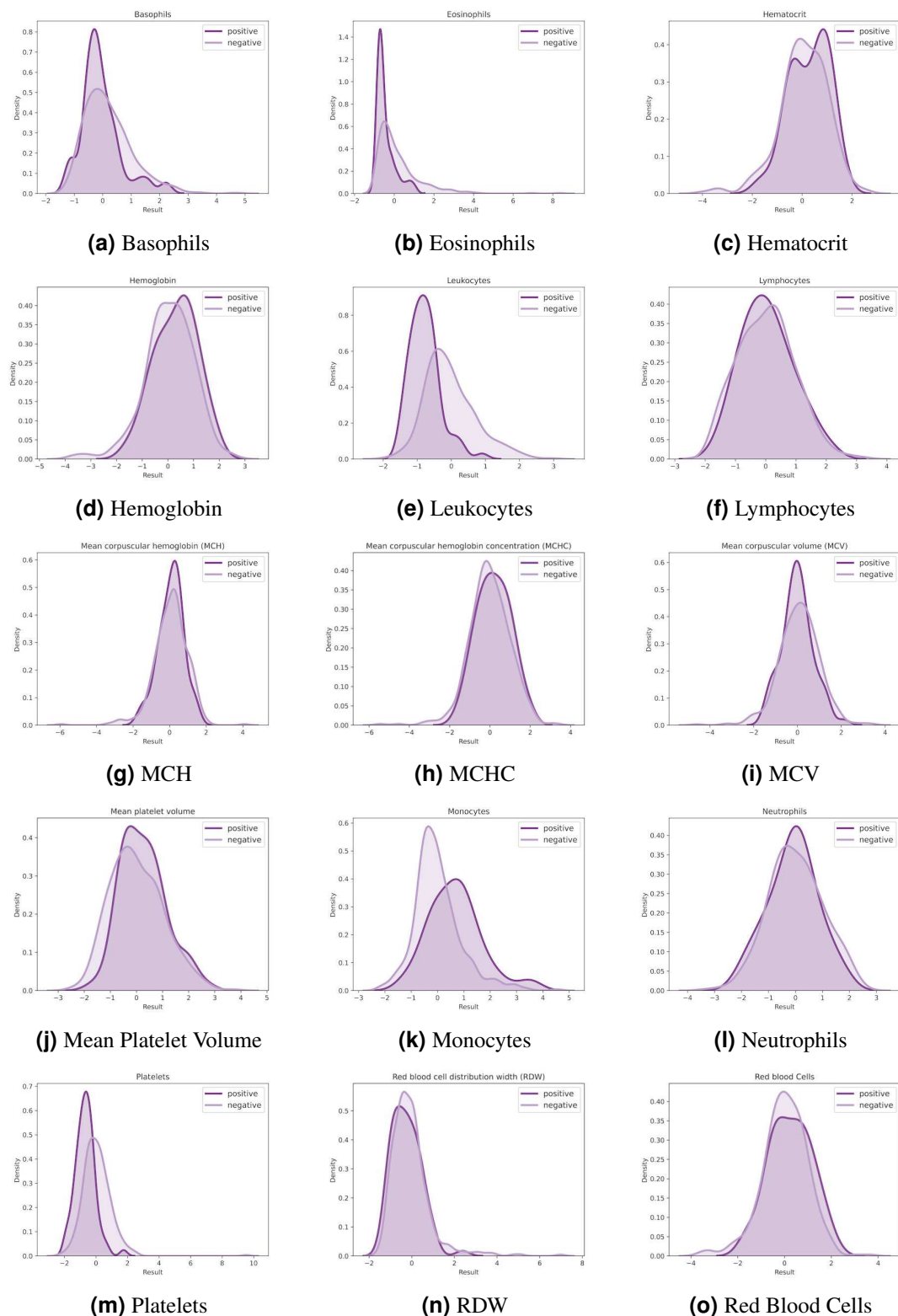


Figure S1. Probability density function (PDF) of all 15 hemogram parameters.

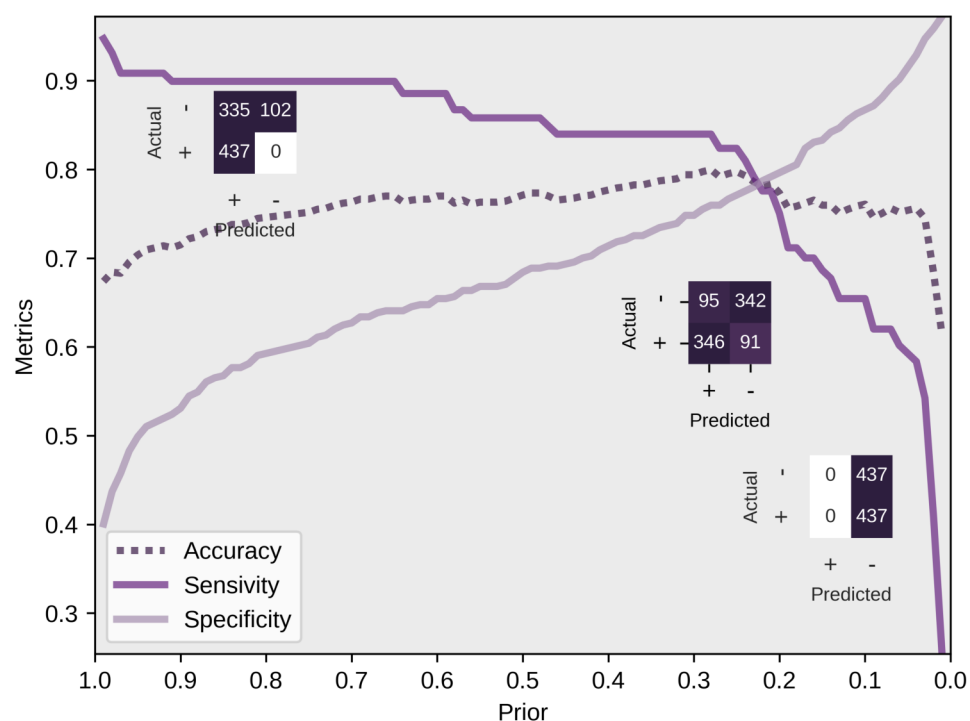


Figure S2. Performance metrics of alternative balanced Naive-Bayes model. In this case, random oversampling of positive results was employed, until sample number in each class is identical. *Prior* probabilities are presented in reference to positive qRT-PCR prediction. Confusion matrices (left to right) are presented for 0.9999, 0.2237 and 0.0001 *prior* probabilities, respectively. Sensitivity=True Positive Ratio; Sensitivity=True Negative Ratio. Random seed was set to 0 for replication purposes.