medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Title: **Nonspecific blood tests as proxies for COVID-19 hospitalization: are there plausible associations after excluding noisy predictors?**

Authors: G. Ishikawa[1], G. Argenti[2], C. B. Fadel[3]

Affiliations:

1. Professor and researcher at Universidade Tecnologica Federal do Parana (UTFPR) in Ponta Grossa, Brazil
2. Researcher enrolled in the postgraduate programme in health sciences at Universidade Estadual de Ponta Grossa (UEPG) in Ponta Grossa, Brazil
3. Professor and researcher at Universidade Estadual de Ponta Grossa (UEPG) in Ponta Grossa, Brazil

Corresponding author: G. Ishikawa

E-mail address: gersonishikawa@utfpr.edu.br

Word count (summary): 195

Word count (text): 3933

No. of references: 40

No. of tables: Four (4)

No. of figures: Six (6)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

SUMMARY

This study applied causal criteria in directed acyclic graphs for handling covariates in associations for prognosis of severe COVID-19 (Corona virus disease 19) cases. To identify nonspecific blood tests and risk factors as predictors of hospitalization due to COVID-19, one has to exclude noisy predictors by comparing the concordance statistics (AUC) for positive and negative cases of SARS-CoV-2 (acute respiratory syndrome coronavirus 2). Predictors with significant AUC at negative stratum should be either controlled for their confounders or eliminated (when confounders are unavailable). Models were classified according to the difference of AUC between strata. The framework was applied to an open database with 5644 patients from Hospital Israelita Albert Einstein in Brazil with SARS-CoV-2 RT-PCR (Reverse Transcription – Polymerase Chain Reaction) exam. C-reactive Protein (CRP) was a noisy predictor: hospitalization could have happen due to causes other than COVID-19 even when SARS-CoV-2 RT-PCR is positive and CRP is reactive, as most cases are asymptomatic to mild. Candidates of characteristic response from moderate to severe inflammation of COVID-19 were: combinations of eosinophils, monocytes and neutrophils, with age as risk factor; and creatinine, as risk factor, sharpens the odds ratio of the model with monocytes, neutrophils, and age.

Keywords: COVID-19; Prediction; Hospitalization; Laboratory tests; Creatinine; Eosinophils; Monocytes; Neutrophils; C-protein reactive

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

INTRODUCTION

COVID-19 (Corona virus disease 19) caused by SARS-CoV-2 (acute respiratory syndrome coronavirus 2) stands out for its high rate of hospitalization and long hospital stay and in intensive care units (ICU). COVID-19 disease severity can be mild, moderate, severe, and critical [1]. While 81% of those infected with COVID-19 have mild or moderate symptoms, World Health Organization estimates that 14% of those infected with COVID-19 are severe and require hospitalization and oxygen support, and 5% are critical and admitted to intensive care units [1]. Reported median hospital length of stay (LoS) was from 4 to 21 days (outside China) and ICU LoS was from 4 to 19 days [2].

The severity of COVID-19 states is associated with many risk factors. Early reports suggest advanced age, morbidities, multi-morbidities, and immunosuppression [3,4]. The enlarging list includes cardiac, chronic lung, cerebrovascular, chronic kidney, and liver diseases, cancer, diabetes, obesity, hypertension, dyspnea, fatigue, and anorexia [1,5,6].

Early identification of severe cases allows for optimizing emergency care support [1] and improving patient outcomes [7]. However, patients who do not yet meet supportive care criteria may fail to receive the necessary care, when there is rapid deterioration or inability to promptly go to a hospital. In the transition from moderate to severe cases there can be avoidable delays in life support interventions with non-optimized treatments.

Together with high hospitalization rates [1] and lengthy stay [2], the superposition of COVID-19 waves and sustained transmission [8] are causing prolonged depletions of health care resources in many countries. Prognosis tools may play a role in planning and in improving the access to supportive treatments by allowing timely allocation of scarce resources to better cope with COVID-19. Indeed, there is widespread interest in predictive models of COVID-19 outcomes [7,9], but a review of 50 prognostic models concluded that they are at high risk of bias [9]. As they focus on statistical findings, our concern is with lack of minimum causal criteria to identify associations that are effectively related to COVID-19.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

In this context, a path to optimized supportive treatments is more reliable assessments of the transition from moderate to severe cases of COVID-19 inflammation. We choose nonspecific blood tests as they are widely available, and hospitalization decision as a proxy to characterize the transition from moderate to severe cases (when not constrained by inpatients availability). After formalizing an analytical framework with causal reasoning, the goal is to identify candidate sets of blood tests associated with hospitalization (with risk factors), excluding noisy predictors that are not related to COVID-19 inflammation.

METHODS

Whereas causal effects are clearly predictive, prediction studies usually refer to noncausal analysis that uses observational data to make predictions beyond the observed ones and confounding bias is generally considered a nonissue [10]. But when one needs more reliable predictions, confounding bias and causality should be accounted for in associations. This study applies analytical tools from the causal effect estimation of directed acyclic graph theory [11] to investigate associations considering covariates.

The strength of the association depends on the specificity and sensitivity of the inflammation pattern, as a kind of distinctive signature of the disease. A low association can also occur and means that the pattern with that set of variables allows weak inferences. If a substantial association due to COVID-19 is identified and it is also stable and representative of the target population, then these blood tests may be useful as proxies in surveillance protocols and screening interventions.

Theoretical framework

The theory of directed acyclic graph (DAG) provides graphical notation and a non-parametric probabilistic terminology to describe and evaluate causal relationships [11]. The use of DAGs in epidemiology is emergent [12] and it is especially helpful with multiple potential confounders [12,13] that may introduce systematic bias [10,14]. In DAGs, confounding associations between two variables may come from unblocked backdoor paths [13] that can be graphically identified because they share parent nodes. With a formal definition of backdoor path, for instance, DAG provides a

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

general explanation of the Simpson's paradox [15], where a phenomenon appears to reverse the sign of the estimated association in disaggregated subsets in comparison to the whole population. As a framework, DAG supplies analytical tools to evaluate which adjustment is mandatory (to predict a noncausal sign reverse) and which covariate should be omitted (to estimate the causal effect), thereby enforcing the elicitation of qualitative causal assumptions [11,12,14].

A hypothetical DAG model with latent variable was conceived to evaluate the influence of various types of covariates on the focal association. Initially we drew the main causal path from exposure to outcome. The DAG in Figure 1 starts from the infection by SARS-CoV-2 (exposure E) that, in some cases, leads to "Moderate to Severe Inflammation due to COVID-19" (MSIC, hypothetical latent variable [E→MSIC]), and that inflammation causes two outcomes (mutual dependent relationship [H←MSIC→B]): (H) hospitalization decision; and ($B=\{B_1,\ldots,B_k\}$) blood tests measured at hospital admission. The blood tests are selected according to their strength with hospitalization. The focal outcomes under investigation are hospitalization (H) and blood tests (B).

Considering the initial DAG plausible, we hypothesized candidate covariates that are parents of the variables and may open back-door paths, Figure 1 shows one risk factor (RF3) and one confounder (BOC1). Figure 2 is an enhancement of the initial DAG with potential risk factors, confounders of the focal association, and other covariates. Risk factors contribute directly to the development of COVID-19 inflammation ($RF=\{RF_1,\ldots,RF_L\}$, mutual causation relationships [RF_i→MSIC←RF_j]) and they can also affect other variables. Figure 2 also distinguishes the covariates in terms of their confounding potential on the association between H and B. Covariates that affect both focal outcomes are identified as Both-Outcomes-Confounders ($BOC=\{BOC_1,\ldots,BOC_m\}$), as they are correlated to the focal outcomes but not to COVID-19, and when affect one outcome as Single-Outcome-Covariate ($SOC=\{SOC_1,\ldots,SOC_n\}$). These covariates are not exhaustive but to generate causal graph criteria for handling confounding factors.

Causal relationships in DAGs are defined with the *do*(.) operator that performs a theoretical intervention by holding constant the value of a chosen variable [11,16]. The association caused by

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

COVID-19 inflammation can be understood as a comparison of the conditional probabilities of hospitalization (H) given a set of blood tests (B) under intervention to SARS-CoV-2 infection (*do*(SARS-CoV-2)=1) and intervention without infection (*do*(SARS-CoV-2)=0):

$$P[H|B=b,do(\text{SARS-CoV-2}=1)] \quad (1)$$

$$P[H|B=b',do(\text{SARS-CoV-2}=0)] \quad (2)$$

Where P[H|B=*b*,*do*(SARS-CoV-2=1)] represents the population distribution of H (hospitalization) given a set of blood tests equal to *b*, if everyone in the population had been infected with SARS-CoV-2. And P[H|B=*b′*,*do*(SARS-CoV-2=0)] if everyone in the population had not been infected. Of interest is the comparison of these distributional probabilities for each intervention.

The interventions with *do*(.) generate two modified DAGs:

- The *do*(SARS-CoV-2=0) eliminates all arrows directed towards SARS-CoV-2 and to MSIC (Figure 3). Ignoring the floating covariates, there are single arrow covariates pointing to hospitalization (RF3, RF4A, SOC1, SOC3) and to blood tests (RF4B, SOC2, SOC4) and fork covariates pointing to both outcomes (BOC1, BOC2, RF5).
- Similarly, the modified graph of *do*(SARS-CoV-2=1) is equal to the former by adding single arrows from RF1 and RF2 to MSIC; and converting RF3, RF4A, RF4B, and RF5 to fork types with arrows directed to MSIC.

As most covariates are either unmeasured or unknown, the effect of their absence can be evaluated following the d-separation concept [11]. This concept attempts to separate (make independent) two focal sets of variables by blocking the causal ancestors (or back-door paths) and by avoiding statistical control for mutual causal descendants [11]. Differently, to preserve the association between descendants of MSIC (Figure 2), the focal outcomes (H and B) must remain d-connected (dependent on each other only through MSIC) and their relations with other covariates (that may introduce systematic bias) should be d-separated (conditionally independent). Figure 3, at the negative stratum, shows the confounders that may introduce systematic bias into both outcomes: BOC1, BOC2, RF5. The influence of these confounders on the focal association can be estimated with

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

the modified model at the negative strata. A strong association of the outcomes without infection can be due to these confounders and suggest efforts to measure and control for them (as they have to be d-separated). Another pragmatic possibility is to exclude the noisy exams affected by these confounders. The other covariates are single arrows or they affect only one outcome (H or B) – their absence should not be critical because they are likely to be discarded due to poor discriminative performance.

Model assessment with naïve estimation

A naïve estimation of equations (1) and (2) is to assume that they are equal to their conditional probabilities available in a given dataset at each stratum. The cost of this simplification is that the analysis is no longer causal (in a counterfactual sense, because we are not contrasting the whole population infected and the whole population not infected [10,11,16]) and the estimation becomes an association between two disjoint sets that each represents separate parts of the target population.

$$P[H|B=b,do(\text{SARS-CoV-2}=1)]=P[H|B=b,\text{SARS-CoV-2}=1] \quad (3)$$

$$P[H|B=b',do(\text{SARS-CoV-2}=0)]=P[H|B=b',\text{SARS-CoV-2}=0] \quad (4)$$

As Hospitalization is a dichotomous variable, this conditional probability, $P[H|B=b,\text{SARS-CoV-2}=1]$, can be computed through a logistic regression of Hospitalization (dependent variable) given a set of blood tests at SARS-CoV-2=1. From the modified graph with intervention, $P[H|B=b',\text{SARS-CoV-2}=0]$ is calculated with the same model parameters but applied to cases at the negative stratum. It is implicit that there is the conditioning by a proper set of covariates for each model.

The concordance statistic of a logistic regression model is a measure of its predictive accuracy and is calculated as the Area Under of the receiver operating characteristic Curve (AUC) [10,17]. A way to compare the discriminative ability of (3) and (4) is to subtract the AUC values at each stratum. A difference of 0.0 means no specific association with COVID-19 (i.e., equivalent responses for both strata) and 0,5 means perfect focal association of the outcomes and perfect differentiation among strata (i.e., perfect response at the positive stratum and random response at the negative).

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

$$\Delta_{\text{Discriminative Ability Naïve}}=\text{AUC}(P[H|B=b,\text{SARS-CoV-2}=1])-\text{AUC}(P[H|B=b',\text{SARS-CoV-2}=0]) \quad (5)$$

The comparison of the models with AUC values at the negative stratum of SARS-CoV-2 is a necessary improvement in the assessment of prognostic models. This is similar to the null values concept in measures of associations of two groups with two outcomes [10], but generalized for continuous multivariable prognostic models.

Model selection criteria

The above framework guided our approach to identify sets of blood tests associated with the hospitalization due to COVID-19 together with:

- Acceptable overall statistical properties of each model at the positive stratum of SARS-CoV-2, without and with bootstrap procedure;
- Consistency of the blood test coefficients across models with one variable and with multiple variables: considering causal effects, coefficients should not change signal when properly conditioned across models [15]; and
- Elimination of models with high AUC at the negative stratum of SARS-CoV-2 and classification of the sets of blood tests by the difference of AUC between strata.

Source dataset

We identified one public observational database in which, at least partially, we could apply the framework and generate candidate prognostic models. Hospital Israelita Albert Einstein (HIAE), Sao Paulo/Brazil, made public a database (HIAE_dataset)[18] in the *kaggle* platform of 5644 patients screened with SARS-CoV-2 RT-PCR (Reverse Transcription–Polymerase Chain Reaction) exam and a few collected additional laboratory tests during a visit to this hospital from February to March, 2020. All blood tests were standardized to have mean of zero and unitary standard deviation. As this research is based on public and anonymized dataset, it was not revised by any institutional board. The logistic regression models were evaluated with IBM SPSS version 22.0 and the causal map with DAGitty.net version 3.0.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

RESULTS

Of the 5644 patients, 558 presented positive results for SARS-CoV-2 RT-PCR. Of the 170 patients hospitalized (in regular ward, semi-intensive unit or intensive care unit), 52 were positive (9,3% rate of hospitalization due to COVID-19). Patient age quantile, from 0 to 19, with sample mean of 9,32, was the only demographic variable available. Age was not conditionally independent with SARS-CoV-2 RT-PCR exam. Only 0,9% were positive in the age quantile 0, 1, and 2 (8 positive cases in 883 exams) while the incidence (not weighted) in the age quantile from 3 to 19 was 11,7% ± 2,6%.

In the first round, 15 blood tests were discarded because of poor performance of the univariate model when SARS-CoV-2=1 (Table 1). The remaining blood tests were creatinine, C-Reactive Protein (CRP), eosinophils, lymphocytes, monocytes, and neutrophils (Table 1). Only creatinine was not related with the immune system directly and was evaluated as a risk factor. Of the 5644 patients, 602 patients presented values of eosinophils, 602 lymphocytes, 601 monocytes, 513 neutrophils, 506 CRP, and 424 creatinine. Regarding missing cases, all observations with the required data were included (available-case analysis).

CRP is a biomarker of various types of inflammation [19,20]. At SARS-CoV-2=1, the model with CRP and age had good discriminative ability with AUC of ,872. But at SARS-CoV-2=0, AUC=,680 was also substantial and the difference of the discriminative ability Δ=,192 was moderate (candidate models should present higher differences); the corresponding ROC curve in Figure 4 show overlapping curves up to sensitivity of 0,5 to 0,6. Models with CRP demonstrated sensitivity to resampling within the dataset [17], the coefficient significance moved from ,005 to ,144. Similar effects were found in models that include CRP with other blood tests and sensitivity to bootstrapping was reduced by dichotomizing CRP (reactive/not-reactive). Models with CRP_reactive, neutrophils, and age generated AUC of ,901 and ,730 in the positive and negative strata (Δ=0,171), and CRP_reactive, monocytes, neutrophils, and age generated AUC of ,921 and ,706, respectively (Δ=,215). CRP is a predictor of hospitalization in general, but high levels of AUC at the negative stratum mean that CRP

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

is a response with significant bias due to other causes than COVID-19. Differently from other prognostic studies [21,22,23,24,25,26], CRP was excluded as candidate.

The Neutrophils to Lymphocytes Ratio (NLR) is considered a possible indicator of severity [21,24,27,28] of COVID-19, but NLR could not be evaluated as all variables were standardized (division by zero). Lymphocytes presented inconsistent behavior across models. Single exam models indicated lymphopenia at SARS-CoV-2=1, as expected [29,30]. But lymphocytes reversed the sign in the model with neutrophils and age (SARS-CoV-2=1), possibly, due to collinearity between them (Pearson correlation of -,925 and -,937 at positive and negative strata, both significant at ,01 (2-tail)). As there are indications of collinearity issues at both strata, lymphocyte and neutrophils should not be in the same model as independent variables, and this is an indication that NLR may be a noisy association with hospitalization. As models with combinations of neutrophils were slightly better than with lymphocyte, lymphocyte was dropped from analysis.

In the second round, combinations of eosinophils, monocytes, and neutrophils with age were tested systematically. Table 2 presents parameters of models combining eosinophils, monocytes, and neutrophils (with age) and the best model with creatinine (as risk factor). Table 3 presents AUCs for each model with the difference of discriminative ability between strata.

Considered individually, eosinophils, monocytes, and neutrophils generated models with good performance to estimate the probability of hospitalization (models 1, 2, 3 with AUC>,810 at positive stratum). The combinations of these blood tests generated models (4, 5, 6, 7) with better discriminative ability (AUC>,856 at SARS-CoV-2=1). The AUC at SARS-CoV-2=0 is a simplified measure of the systematic bias in both outcomes: models 1, 2, and 4 presented low values (with AUC<,564) and the others presented relevant noisy associations (AUC from ,600 up to ,665), but with better difference in discriminative ability Δ>,252 in models with two or more exams.

Two patterns of associations were more salient: (1) age as a risk factor with combinations of eosinophils, monocytes, and neutrophils as predictors; (2) age and creatinine as risk factors with monocytes and neutrophils as predictors. The interpretation of the conditional probabilities will

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

focus on models 7 and 8, but models with at least two blood tests (4 to 8) are potential candidate associations. Considering creatinine as a marker of the renal function, model 8 is the overall best model with significant coefficients at p<,05 and has the highest difference of discriminative ability between strata (Δ=,313). Comparative ROC curves for models 7 and 8 are shown in Figures 5 and 6, where there is a substantial discriminative difference between both strata of SARS-CoV-2; confidence intervals at 95% of AUC values are in Table 3.

When the coefficients of model 7 (Table 2) are converted to conditional probabilities we find that at average age quantile (9,32) and average monocyte and neutrophil levels, there is a hospitalization probability of 51,1% with eosinophils at -1 standard deviation (SD); and 90,2% when age quantile is 15. Model 8 with creatinine has different responses: age quantile coefficient is more pronounced and the odds ratio of creatinine is steep (8,338), so average levels of creatinine result in a probability of hospitalization >50% for age quantile >9 (with monocytes and neutrophils at average). When creatinine is +1 SD at age quantile 9, hospitalization probability is 85,9% (monocytes and neutrophils at average). Only below average levels of creatinine lower hospitalization probabilities. Monocytes and neutrophils are also steeper than model 7. At age quantile 9, +1/2 SD of creatinine, -1/2 SD of monocytes, and +1/2 SD of neutrophils result in a hospitalization probability of 92,5%.

Model biases may be due to missing cases selection. Most likely, missing data are not at random (MNAR). We performed the bootstrapping procedure to identify potential sensitivity to resampling and, indirectly, to selection bias. The selected models maintained the magnitude and statistical significance of the coefficients. Apparently, as no significant deviation was detected, the missing cases bias may not be an issue.

ROC and AUC calculations used the same data for model fitting. Because of limited sample size, it was not suitable to apply the approach of splitting the database for training and then prediction. After dividing the sample in two groups, most coefficients were not significant at p>0,10 (Table 4) at least in one group. Notwithstanding, classification tables were coherent between sub-sets and we found no clear indication of model misspecification.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

DISCUSSION

We focused on models with discriminative ability to identify peculiar responses in the transition from moderate to severe inflammation only due to COVID-19. The AUC evaluation at the negative SARS-CoV-2 stratum to estimate the influence of unwanted confounders into the focal association together with equivalent criteria of severity state at both strata is, to the best of our knowledge, a needed improvement in prognosis studies of COVID-19.

In comparison to other prediction studies, we identified a few focused on the transition from moderate to severe cases of COVID-19 [21,22,23,24,25,26,27,28]. None of them considered data from the negative stratum of SARS-CoV-2, therefore, these models are biased by not excluding noisy predictors.

We eliminated variables with "high" AUC at SARS-CoV-2=0, so that variables with more peculiar responses to COVID-19 were included. Reactive levels of CRP together with SARS-CoV-2 RT-PCR exam may be a predictor of hospitalization, but this can happen due to causes other than COVID-19 (most cases of COVID-19 are asymptomatic to mild). To include it in a model, one should control for all other causes of CRP reactive.

We evaluated age and creatinine as risk factors. Controlling for age improved the AUC of all models at the positive stratum of SARS-CoV-2. The difference between risk factor and outcome among blood tests is subtle. The emergent literature is cautious about whether eosinopenia may be a risk factor [31] and whether creatinine (and other renal markers) may be associated with COVID-19 renal inflammatory response [32]. As an acute inflammatory kidney response to COVID-19, the interpretation changes and further refinement of the framework is necessary. If eosinopenia is a risk factor, the prevalence of this condition should be considered and must be properly diagnosed at admission, and the models should be reviewed with new data.

As we drop noisy predictors, we are effectively dealing with hypothesis about the physiopathology of COVID-19 inflammation. Although not as frequent as the mentions of neutrophils, there are studies on the complex role of eosinophils [31,33] and monocytes [34,35] in COVID-19 inflammation

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

indicating eosinopenia in severe cases and monocytopenia in some phase of the cytokine storm and other COVID-19 pathologies [36].

We selected two patterns of blood tests that are associated with hospitalization due to COVID-19 inflammation: age with combinations of eosinophils, monocytes and neutrophils; and age and creatinine with monocytes and neutrophils. The model findings are aligned with the known physiopathology of COVID-19 but in a more integrative framework of analysis (not as individual predictors, but as a set that is related to risk factors). The selected blood tests are broadly available even in regions with scarce health care resources. It is unlikely that we will have just one or two overall best models; given different sets of risk factors, we should expect a few representative patterns of the COVID-19 inflammation from moderate to severe.

Limitations and future directions

The models are candidates only and the results cannot be representative beyond the patient health profiles of this reference hospital in Sao Paulo/Brazil that attends a high social-economic segment [37]. The sample refers to the initial phase of the pandemics in Brazil and the patterns may change with medicine prescriptions and other adaptations of SARS-CoV-2. The reduced quantity of available cases did not allow the dataset split for training and prediction. Further efforts are needed to increase internal and external validity across populations, as the prognostic ability is also a function of the variability of the development of COVID-19 inflammation.

As there is no unambiguous way to characterize “moderate to severe COVID-19 inflammation”, the inclusion of an unmeasured variable reduces the predicted conditional independences from the DAG. But still this framework can help in the identification and estimation of risk factors. This cross-sectional data (single point time) cannot inform if creatinine (or eosinophil) is risk factor or effect of COVID-19 inflammation. In future data collection efforts, participants should be followed over time, from diagnosis to hospitalization; ideally from exposure throughout the lifecycle and also with the follow-up of negative cases.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Causal studies are intrinsically predictive [10], therefore we need to advance prognosis research within causal frameworks. As most studies will be observational, data collection with ample selection of variables for matching estimators (e.g., stratification) [16] will be required to reduce systematic bias.

All candidate models can be reproduced from the dataset [18]. We believe most hospitals can apply this framework to generate similar models appropriate to the target population in which they are inserted by making efforts to collect blood tests and potential risk factors at admission, and other clinical data. By making these databases public (anonymized and with standardized data), they will allow future external validation in larger target populations.

Finally, in the wider context of COVID-19 epidemiology, the collapse of health systems due to opportunistic pathogens is a symptom of threats that requires system-level measures during and after the pandemics [38]. This research is concerned with hospital care. As a bottleneck, even small gains may have multiplicative effects on health systems. In countries with porous containment efforts, hospital occupancy is a critical metric [39] to alternate between "soft lockdown" and economic activity with "constrained mobility". As some regions with sustained transmission are hesitant and being pushed towards these states, they are poorly capturing the benefits of the switching strategy (Parrondo's paradox applied to epidemics [40]) – because they are struggling in trial and error mode to establish thresholds of when to restrain (and open) and at what pace. Due to the fast saturation of hospital infra-structures with overshooting in these regions, the tendency of excessive losses in each transition is hard to manage. In this context, we believe that the application of prognosis tools can improve the timely access to supportive care in countries with sustained COVID-19 transmission.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Acknowledgements:

We are grateful to Antonio Magno Lima Espeschit and Sonia Mara de Andrade who contributed with suggestions to this research. We are also indebted to Hospital Israelita Albert Einstein for making the dataset available, and the referees for their detailed comments.

Author Contributions:

G. Ishikawa: Conceptualization, methodology, and formal analysis

G. Argenti: Conceptualization, formal analysis, and clinical and epidemiological validation

C. B. Fadel: Clinical and epidemiological validation and critical review

All authors: Writing, editing, visualization, review and final approval of manuscript

Statements:

The authors declare no conflicts of interest.

This paper has not been published previously in whole or part.

The data that support the results of this study are openly available in reference number [18].

Although this research received no specific grant from any funding agency, commercial or not-for-profit sectors, as institutionally required we inform that "this study was financed in part by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001".

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

REFERENCES

1 - **World Health Organization**. Clinical management of COVID-19: Interim guidance. WHO publications [Internet]. 2020 [cited 2020 May 27]; Available from: https://www.who.int/publications/i/item/clinical-management-of-covid-19

2 – **Rees EM, et al**. COVID-19 length of hospital stay: a systematic review and data synthesis. *BMC Medicine*. 18, 270 (2020). doi: https://doi.org/10.1186/s12916-020-01726-3

3 - **Guan W, et al**. Clinical characteristics of coronavirus disease 2019 in China. *The New England Journal of Medicine* [Internet]. 2020. doi: https://doi.org/10.1056/NEJMoa2002032

4 - **Italy: SARS-CoV-2 Surveillance Group**. Characteristics of COVID-19 patients dying in Italy. Epidemiology for public health: Istituto Superiore di Sanità [Internet]. 2020. [cited 2020 Apr 24]; Available from: https://www.epicentro.iss.it/en/coronavirus/sars-cov-2-analysis-of-deaths

5 – **CDC**. Human infection with 2019 novel coronavirus person under investigation (PUI) and case report form. Atlanta, GA: US Department of Health and Human Services, CDC; 2020. https://www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf

6 – **Liu X, et al**. Risk factors associated with disease severity and length of hospital stay in COVID-19 patients. *Journal of Infection*. 2020;81(1):e95-e97. doi: https://doi.org/10.1016/j.jinf.2020.04.008

7 – **Marin BG, et al**. Predictors of COVID-19 severity: A literature review. *Reviews in Medical Virology*. 2020;e2146. doi: https://doi.org/10.1002/rmv.2146

8 – **Lai JW, Cheong KH**. Superposition of COVID-19 waves, anticipating a sustained wave, and lessons for the future. *BioEssays*, 2020, 42, e2000178. doi: https://doi.org/10.1002/bies.202000178

9 – **Wynants l, et al**. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020; 369 :m1328. doi: https://doi.org/10.1136/bmj.m1328

10 – **Westreich D**. *Epidemiology by design*: a causal approach to the health sciences. 1st ed. New York: Oxford University Press, 2020, pp. 41-104.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

11 – **Pearl J**. *Causality:* models, reasoning, and inference. 2nd ed. Cambridge: Cambridge University Press, 2009, pp. 65-106.

12 – **Foraita R, Spallek J, Zeeb H**. Directed Acyclic Graph. In: Ahrens W, Pigeot I. *Handbook of Epidemiology*. 2nd ed. New York: Springer, 2014, pp. 1481-1518.

13 – **Greenland S, Pearl J, Robins JM**. Causal diagrams for epidemiologic research. *Epidemiology*. 1999 Jan; 10(1):37-48. PMID: 9888278. doi: https://doi.org/10.1097/00001648-199901000-00008

14 – **Glymour MM, Greenland S**. Causal Diagrams. In: Rothman KJ, Greenland S, Lash TL (Ed). *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008, pp. 183-212.

15 – **Pearl J**. Comment: Understanding Simpson's Paradox. 2014. *The American Statistician*. 68:1, 8-13, doi: https://doi.org/10.1080/00031305.2014.876829

16 – **Morgan SL, Winship C**. *Counterfactuals and Causal Inference*: Methods and principles for social research. 2nd ed. New York: Cambridge University Press, 2015, pp. 77-139.

17 – **Hosmer DW, Lemeshow S, Sturdivant RX**. *Applied Logistic Regression*. 3rd ed. Hoboken: John Willey & Sons, 2013, pp.173-181.

18 – **Hospital Israelita Albert Einstein**. Diagnosis of COVID-19 and its clinical spectrum: AI and Data Science supporting clinical decisions (from 28th Mar to 3st Apr). Kaggle [Internet]. 2020 [cited 2020 Apr 8]; Available from: https://www.kaggle.com/dataset/e626783d4672f182e7870b1bbe75fae66bdfb232289da0a61f08c2ceb01cab01

19 – **Lelubre C, et al**. Interpretation of C-reactive protein concentrations in critically ill patients, *BioMed Research International*. vol. 2013, Article ID 124021, 11 pages, 2013. doi: https://doi.org/10.1155/2013/124021

20 – **Black S, Kushner I, Samols D**. C-reactive protein. *Journal of Biological Chemistry*. 2004 Nov 19;279(47):48487-90. doi: https://doi.org/10.1074/jbc.R400025200.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

21 – **Cheng B, et al**. Predictors of progression from moderate to severe coronavirus disease 2019: a retrospective cohort. *Clinical Microbiology Infection*. 2020; 26(10):1400-1405. doi: https://doi.org/10.1016/j.cmi.2020.06.033

22 – **Tan L, et al**. Validation of predictors of disease severity and outcomes in COVID-19 patients: A descriptive and retrospective study [published online ahead of print, 2020 May 19]. *Med* (NY). 2020;10.1016/j.medj.2020.05.002. doi: https://doi.org/10.1016/j.medj.2020.05.002

23 – **Zhu Z, et al**. Clinical value of immune-inflammatory parameters to assess the severity of coronavirus disease 2019. *International Journal of Infectious Disease*. 2020; 95:332-339. doi: https://doi.org/10.1016/j.ijid.2020.04.041

24 – **Shang W, et al**. The value of clinical parameters in predicting the severity of COVID-19 [published online ahead of print, 2020 May 21]. *Journal of Medical Virology*. 2020; 10.1002/jmv.26031. doi: https://doi.org/10.1002/jmv.26031

25 – **Zhou C, et al**. Predictive factors of severe coronavirus disease 2019 in previously healthy young adults: a single-center, retrospective study. *Respiratory Research*. 2020; Res 21, 157. https://doi.org/10.1186/s12931-020-01412-1

26 – **Bhargava A, et al**. Predictors for severe COVID-19 infection, *Clinical Infectious Diseases*. 2020, ciaa674. doi: https://doi.org/10.1093/cid/ciaa674

27 – **Yang AP, Liu JP, Tao WQ, Li HM**. The diagnostic and predictive role of NLR, d-NLR and PLR in COVID-19 patients. *International Immunopharmacology*. 2020;84:106504. doi: https://doi.org/10.1016/j.intimp.2020.106504

28 – **Wang C, et al**. Preliminary study to identify severe from moderate cases of COVID-19 using combined hematology parameters. *Annals of Translational Medicine*. 2020; 8(9):593. doi: https://doi.org/10.21037/atm-20-3391

29 – **Zhao Q, et al**. Lymphopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A systemic review and meta-analysis. *International Journal of Infectious Diseases*.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

2020, Volume 96, Pages 131-135, ISSN 1201-9712, doi: https://doi.org/10.1016/j.ijid.2020.04.086

30 – **Huang I, Pranata R**. Lymphopenia in severe coronavirus disease-2019 (COVID-19): systematic review and meta-analysis. *Journal of Intensive Care*. 2020; 8, 36. https://doi.org/10.1186/s40560-020-00453-4

31 – **Lindsley AW, Schwartz JT, Rothenberg ME**. Eosinophil responses during COVID-19 infections and coronavirus vaccination. *The Journal of Allergy and Clinical Immunology*. 2020; 146(1):1-7. doi: https://doi.org/10.1016/j.jaci.2020.04.021

32 – **Qian JY, Wang B, Liu BC**. Acute kidney injury in the 2019 novel coronavirus disease. *Kidney Diseases*. 2020;6:318-323. doi: https://doi.org/10.1159/000509086

33 – **Xie G, et al**. The role of peripheral blood eosinophil counts in COVID-19 patients. *Allergy*. 2020; 00: 1– 12. doi: https://doi.org/10.1111/all.14465

34 – **Alzaid F, et al**. Monocytopenia, monocyte morphological anomalies and hyperinflammation characterise severe COVID-19 in type 2 diabetes [published online ahead of print, 2020 Aug 20]. *EMBO Molecular Medicine*. 2020; e13038. doi: https://doi.org/10.15252/emmm.202013038

35 – **Martinez F, et al**. Monocyte activation in systemic Covid-19 infection: Assay and rationale. *EBioMedicine*. 2020; 59. 102964. doi: https://doi.org/10.1016/j.ebiom.2020.102964

36 – **Pence, B**. Severe COVID-19 and aging: are monocytes the key?. *GeroScience*. 2020. 42. doi: https://doi.org/10.1007/s11357-020-00213-0

37 – **Teich VD, et al**. Epidemiologic and clinical features of patients with COVID-19 in Brazil. *einstein* (Sao Paulo). 2020;18:eAO6022. doi: http://dx.doi.org/10.31744/einstein_journal/2020AO6022

38 – **Cheong KH, Jones MC**. Introducing the 21st Century's New Four Horsemen of the Coronapocalypse. *BioEssays*, 2020, 42, 2000063. doi: https://doi.org/10.1002/bies.202000063

39 – **Martins CM, et al**. Predictive model for COVID-19 incidence in a medium-sized municipality in Brazil (Ponta Grossa, Parana). *Texto & Contexto - Enfermagem*, 2020, 29, e20200154. doi: https://dx.doi.org/10.1590/1980-265x-tce-2020-0154

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

40 – **Cheong KH, Wen T, Lai JW**. Relieving Cost of Epidemic by Parrondo's Paradox: A COVID-19 Case Study. *Advanced Science*, 2020, 2002324. doi: https://doi.org/10.1002/advs.202002324

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 1 – Univariate logistic regression models with blood tests for predicting hospitalization

	At SARS-CoV-2=1						At SARS-CoV-2=0					
	N	B	p	OR	OR 95% C.I.		N	B	p	OR	OR 95% C.I.	
					Lower	Upper					Lower	Upper
zBasophils	83	-,374	,229	,688			519	-,375	,010	,687		
zHematocrit	83	-,123	,658	,884			520	-,976	,000	,377		
zHemoglobin	83	-,073	,785	,930			520	-1,009	,000	,365		
zLeukocytes	83	,617	,167	1,854			519	,658	,000	1,931		
zMCH	83	-,253	,280	,776			519	-,289	,011	,749		
zMCHC	83	,118	,629	1,126			519	-,259	,023	,772		
zMCV	83	-,331	,176	,718			519	-,196	,094	,822		
zMPV	81	-,465	,079	,628			518	-,229	,062	,795		
zPlatelets	83	-,272	,433	,762			519	,101	,363	1,107		
zPotassium	58	-,482	,145	,618			313	,161	,210	1,174		
zRed_blood_cells	83	,087	,707	1,091			519	-,791	,000	,453		
zRDW	83	,140	,560	1,150			519	,648	,000	1,912		
zSerum_glucose	33	-,172	,734	,842			175	,713	,001	2,041		
zSodium	58	-,530	,097	,589			312	-,232	,077	,793		
zUrea	59	,468	,275	1,597			338	,403	,004	1,496		
Age_quantile *	558	0,199	,000	1,220	1,137	1,310	5086	-0,03	,044	0,968	0,938	0,999
zCreatinine **	62	1,002	,019	2,723	1,177	6,301	362	-,116	,367	,891	,693	1,145
zCRP **	70	1,857	,004	6,406	1,805	22,73	436	1,012	,000	2,751	2,015	3,756
zEosinophils **	83	-2,768	,001	,063	,012	,332	519	-,312	,036	,732	,547	,980
zLymphocytes **	83	-,794	,006	,452	,256	,796	519	-,537	,000	,584	,451	,758
zMonocytes **	83	-,629	,006	,533	,339	,838	518	-,321	,021	,726	,552	,953
zNeutrophils **	75	1,412	,000	4,104	1,957	8,605	438	,509	,001	1,663	1,244	2,224

Legend:

SARS-CoV-2 (acute respiratory syndrome coronavirus 2): result of the exam for SARS-CoV-2 RT-PCR (0=negative; 1=positive) (reverse transcription – polymerase chain reaction)

N: Cases included in the analysis; B: coefficient of the univariate logistic regression; p: coefficient significance; OR: odds ratio (exp(B)); CI: confidence interval

MCH: Mean corpuscular hemoglobin; MCHC: Mean corpuscular hemoglobin concentration; MCV: Mean corpuscular volume; MPV: Mean platelet volume; RDW: Red blood cell distribution width

zName: means that the variable was converted and made available in a standardized format (mean=0; standard deviation=1)

* Age was converted in quantiles in the range of 0 to 19, mean value is 9,32.

** Blood tests selected for screening as potential predictors of COVID-19 inflammation

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 2 – Potential candidate logistic regression models for predicting hospitalization with blood tests and age quantile (different models for each stratum)

		At SARS-CoV-2 = 1					At SARS-CoV-2 = 0				
					OR 95% C.I.					OR 95% C.I.	
		B	p	OR	Lower	Upper	B	p	OR	Lower	Upper
Model 1	Age_quantile	,223	,001	1,250	1,091	1,432	,002	,906	1,002	,963	1,043
	zEosinophils	-2,506	,004	,082	,015	,441	-,314	,036	,731	,545	,980
	Constant	-4,233	,000	,015			-1,650	,000	,192		
Model 2	Age_quantile	,249	,000	1,282	1,120	1,468	,000	,995	1,000	,961	1,041
	zMonocytes	-,693	,008	,500	,300	,834	-,321	,021	,726	,552	,954
	Constant	-2,931	,002	,053			-1,668	,000	,189		
Model 3	Age_quantile	,303	,001	1,354	1,137	1,612	,055	,050	1,057	1,000	1,117
	zNeutrophils	1,299	,002	3,665	1,617	8,308	,493	,001	1,637	1,223	2,192
	Constant	-3,940	,002	,019			-2,687	,000	,068		
Model 4	Age_quantile	,240	,001	1,271	1,103	1,466	,003	,885	1,003	,963	1,044
	zEosinophils	-2,109	,012	,121	,023	,630	-,290	,050	,748	,560	1,000
	zMonocytes	-,506	,057	,603	,358	1,015	-,292	,032	,746	,572	,975
	Constant	-4,005	,000	,018			-1,701	,000	,183		
Model 5	Age_quantile	,299	,002	1,349	1,119	1,626	,053	,058	1,055	,998	1,115
	zEosinophils	-2,004	,025	,135	,023	,780	,191	,181	1,211	,915	1,603
	zNeutrophils	1,175	,010	3,240	1,319	7,954	,586	,001	1,797	1,292	2,500
	Constant	-4,927	,001	,007			-2,712	,000	,066		
Model 6	Age_quantile	,362	,001	1,436	1,166	1,770	,056	,050	1,057	1,000	1,118
	zMonocytes	-1,010	,014	,364	,163	,816	-,018	,919	,982	,697	1,384
	zNeutrophils	,968	,033	2,632	1,080	6,413	,487	,002	1,628	1,191	2,224
	Constant	-4,089	,005	,017			-2,687	,000	,068		
Model 7	Age_quantile	,363	,001	1,437	1,149	1,797	,053	,059	1,055	,998	1,115
	zEosinophils	-1,951	,036	,142	,023	,884	,194	,183	1,214	,913	1,615
	zMonocytes	-,925	,023	,397	,178	,882	,018	,920	1,018	,716	1,448
	zNeutrophils	,897	,069	2,453	,933	6,447	,593	,001	1,810	1,264	2,592
	Constant	-5,174	,003	,006			-2,712	,000	,066		
Model 8	Age_quantile	,470	,006	1,600	1,148	2,230	,071	,023	1,074	1,010	1,142
	zCreatinine	2,121	,020	8,338	1,400	49,648	-,267	,166	,766	,525	1,117
	zMonocytes	-1,540	,013	,214	,064	,724	-,076	,690	,927	,639	1,344
	zNeutrophils	1,981	,018	7,251	1,401	37,528	,560	,001	1,751	1,249	2,454
	Constant	-4,542	,031	,011			-2,512	,000	,081		

Legend:

SARS-CoV-2 (acute respiratory syndrome coronavirus 2): result of the exam for SARS-CoV-2 RT-PCR (0=negative; 1=positive) (reverse transcription – polymerase chain reaction);

B: coefficient of the variable;

P: value of the statistical significance of the coefficient;

OR: odds ratio of B (exp(B));

C.I.: confidence interval

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 3 – Discriminative ability of potential candidate models for predicting hospitalization from nonspecific blood tests

			Model							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Variables included in the model:		zEosinophils	●			●	●		●	
		zMonocytes		●		●		●	●	●
		zNeutrophils			●		●	●	●	●
		Age quantile (0 - 19)	●	●	●	●	●	●	●	●
		Creatinine								●
Model for positive stratum (SARS-CoV-2=1)	AUC (Area under ROC curve)		,839	,810	,862	,856	,899	,897	,910	,940
	Standard Error		,046	,049	,044	,043	,036	,036	,034	,029
	Asymptotic Significance		,000	,000	,000	,000	,000	,000	,000	,000
	AUC 95% CI Asymptotic	Lower bound	,748	,715	,775	,772	,828	,826	,844	,883
		Upper bound	,929	,906	,948	,941	,970	,967	,976	,997
	Classification table (cut value = 0,5)	Percentage correct H=0	70,0	70,0	75,0	75,0	72,2	72,2	75,0	81,0
		Percentage correct H=1	79,1	79,1	84,6	83,7	87,2	82,1	89,7	82,9
		Overall percentage	74,7	74,7	80,0	79,5	80,0	77,3	82,7	82,1
	Cases included in the analysis	H=0	40	40	36	40	36	36	36	21
		H=1	43	43	39	43	39	39	39	35
		Total	83	83	75	83	75	75	75	56
Same model of positive stratum applied to SARS-CoV-2=0 cases	AUC		,562	,542	,665	,564	,603	,645	,600	,627
	Standard Error		,036	,043	,045	,037	,044	,047	,044	,049
	Asymptotic Significance		,069	,214	,000	,061	,012	,000	,016	,003
	AUC 95% CI Asymptotic	Lower bound	,492	,459	,576	,492	,518	,553	,513	,532
		Upper bound	,632	,626	,754	,636	,689	,737	,686	,723
	Cases included in the analysis	H=0	433	432	382	432	382	382	382	244
		H=1	86	86	56	86	56	56	56	55
		Total	519	518	438	518	438	438	438	299
Difference of the discriminative ability (naïve)			0,277	0,268	0,197	0,292	0,295	0,252	0,310	0,313
Overall discriminative performance order			5	6	8	4	3	7	2	1

Legend:

AUC: Area under the receiver operating characteristic curve;

CI: confidence interval;

H: Hospitalization (0=false; 1=regular ward, semi-intensive care, or intensive care unit);

SARS-CoV-2 (acute respiratory syndrome coronavirus 2): result of the exam for SARS-CoV-2 RT-PCR (0=negative; 1=positive) (reverse transcription – polymerase chain reaction)

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 4 – Tentative parameters for models 4 to 8 with split dataset at the positive stratum of SARS-CoV-2: sample size is unsuitable for training and then prediction

		Cases 1 to 5030					Cases 5031 to 5644				
					OR 95% CI					OR 95% CI	
		B	p	OR	Lower	Upper	B	p	OR	Lower	Upper
Model 4	Age_quantile	,441	,002	1,555	1,178	2,052	,022	,849	1,022	,817	1,278
	zEosinophils	-1,730	,083	,177	,025	1,250	-6,905	,026	,001	,000	,431
	zMonocytes	-,586	,144	,556	,253	1,221	,175	,717	1,192	,461	3,079
	Constant	-6,673	,002	,001			-4,040	,117	,018		
	Cases (N)			47					36		
	% correct (H=0)			86,2					54,5		
	% correct (H=1)			72,2					92,0		
	Overall correct (%)			80,9					80,6		
Model 5	Age_quantile	,419	,004	1,520	1,140	2,027	,049	,792	1,050	,731	1,509
	zEosinophils	-1,259	,196	,284	,042	1,916	-7,798	,026	,000	,000	,389
	zNeutrophils	,612	,284	1,844	,603	5,640	2,301	,038	9,987	1,131	88,22
	Constant	-6,403	,004	,002			-4,037	,222	,018		
	Cases (N)			45					30		
	% correct (H=0)			77,8					88,9		
	% correct (H=1)			77,8					100,0		
	Overall correct (%)			77,8					96,7		
Model 6	Age_quantile	,467	,005	1,596	1,153	2,209	,228	,211	1,256	,879	1,795
	zMonocytes	-,916	,068	,400	,149	1,071	-,999	,235	,368	,071	1,915
	zNeutrophils	,548	,327	1,729	,579	5,165	1,617	,058	5,036	,949	26,73
	Constant	-5,790	,008	,003			-1,731	,500	,177		
	Cases (N)			45					30		
	% correct (H=0)			88,9					77,8		
	% correct (H=1)			72,2					100,0		
	Overall correct (%)			82,2					93,3		
Model 7	Age_quantile	,504	,005	1,655	1,162	2,359	,060	,774	1,062	,703	1,605
	zEosinophils	-1,475	,201	,229	,024	2,200	-8,005	,042	,000	,000	,747
	zMonocytes	-,898	,069	,407	,154	1,074	-1,024	,402	,359	,033	3,933
	zNeutrophils	,342	,564	1,407	,441	4,492	2,076	,065	7,974	,882	72,13
	Constant	-6,988	,008	,001			-3,731	,264	,024		
	Cases (N)			45					30		
	% correct (H=0)			88,9					88,9		
	% correct (H=1)			72,2					100,0		
	Overall correct (%)			82,2					96,7		
Model 8	Age_quantile	2,321	,155	10,184	,415	249,9	,245	,279	1,278	,819	1,994
	zMonocytes	-4,518	,183	,011	,000	8,467	-1,689	,168	,185	,017	2,039
	zNeutrophils	3,774	,280	43,55	,047	40739	2,758	,035	15,769	1,211	205,3
	zCreatinine	4,304	,221	73,96	,075	72999	2,420	,205	11,240	,266	475,4
	Constant	-25,16	,155	,000			-,919	,772	,399		
	Cases (N)			29					27		
	% correct (H=0)			92,9					85,7		
	% correct (H=1)			93,3					95,0		
	Overall correct (%)			93,1					92,6		

Legend:

SARS-CoV-2: acute respiratory syndrome coronavirus 2; RT-PCR: reverse transcription – polymerase chain reaction; B: coefficient of the variable; p: value of the statistical significance of the coefficient; OR: odds ratio of B (it is equal to exp(B)); C.I.: confidence interval.

Results of classification table cut-off value of 0,5 with percentage of correct non-hospitalization (H=0) and correct hospitalization (H=1).

Note: The cut off at 5030 cases was selected to generate valid parameters with similar quantities of available cases at SARS-CoV-2=1 because lower/higher thresholds generated invalid parameters for model 8 due to perfect discrimination.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

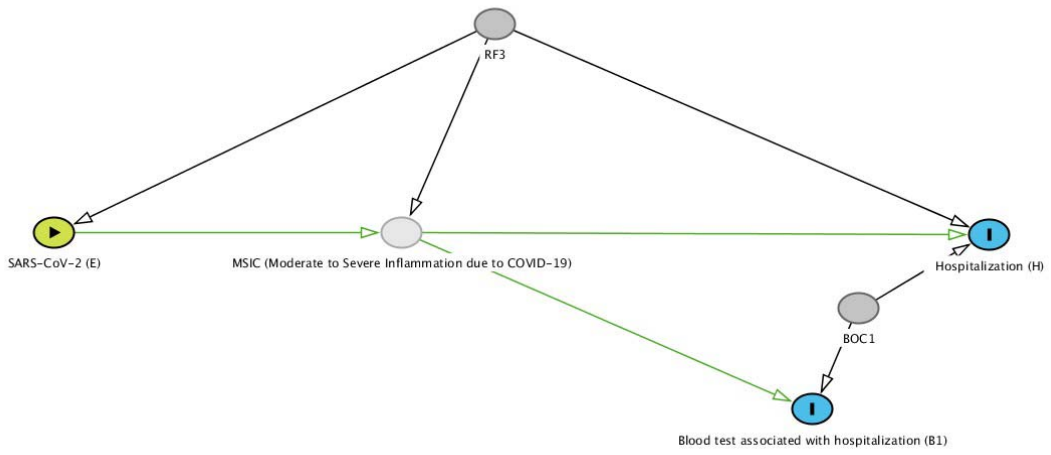

Figure 1 – Initial hypothetical directed acyclic diagram with the main causal path of a moderate to severe COVID-19 inflammation (MSIC), one risk factor (RF3) and one confounder (BOC1) of the focal outcomes (H and B1).

Legend: MSIC is a latent variable (unmeasured); Outcomes are H: hospitalization (H={regular ward, semi-intensive care, intensive care unit}); and B: blood test ($B=\{B_1\}$)

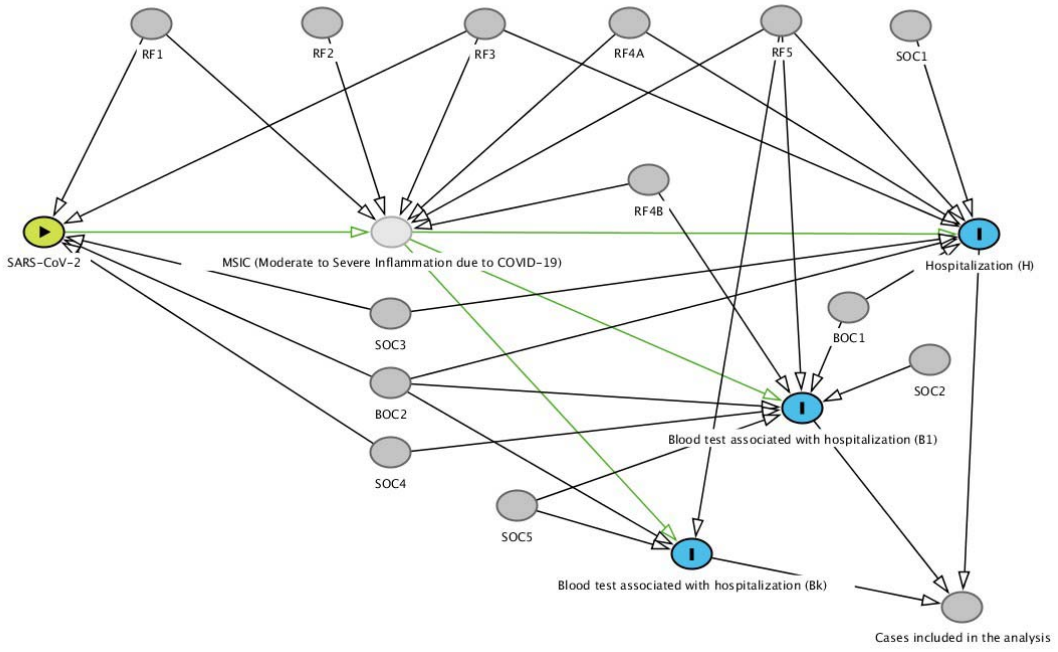

Figure 2 – Hypothetical directed acyclic diagram of a COVID-19 inflammation causal path with risk factors, confounders and other covariates

Legend: Exposure = SARS-CoV-2 (E) (acute respiratory syndrome coronavirus 2); Outcomes are H: hospitalization (H={regular ward, semi-intensive care, intensive care unit}), and B: blood tests ($B=\{B_1,\ldots,B_K\}$); Covariates are RF: risk factor ($RF=\{RF_1,\ldots,RF_{4A}, RF_{4B},RF_5\}$), SOC: single outcome covariate ($SOC=\{SOC_1,\ldots,SOC_5\}$), and BOC: both outcomes confounder ($BOC=\{BOC_1,BOC_2\}$)

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

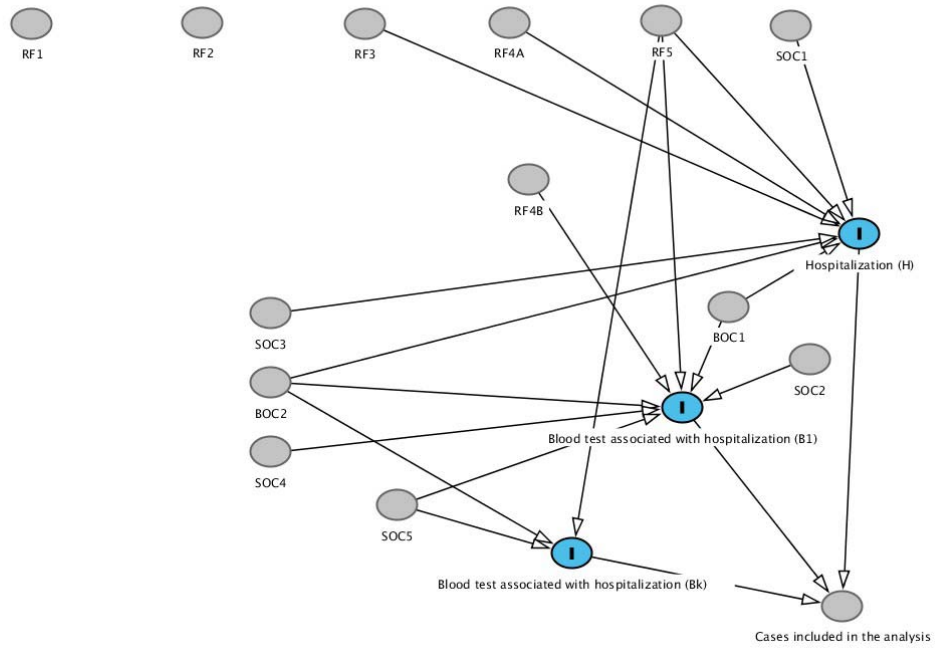

Figure 3 – Modified directed acyclic diagram with intervention at no exposure (*do*(SARS-CoV-2=0)) to evaluate the influence of covariates on the focal outcomes (H and B)

Legend: Exposure = SARS-CoV-2 (E) (acute respiratory syndrome coronavirus 2); Outcomes are H: hospitalization (H={regular ward, semi-intensive care, intensive care unit}), and B: blood tests ($B=\{B_1,\ldots,B_K\}$); Covariates are RF: risk factor ($RF=\{RF_1,\ldots,RF_{4A}, RF_{4B},RF_5\}$), SOC: single outcome covariate ($SOC=\{SOC_1,\ldots,SOC_5\}$), and BOC: both outcomes confounder ($BOC=\{BOC_1,BOC_2\}$)

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

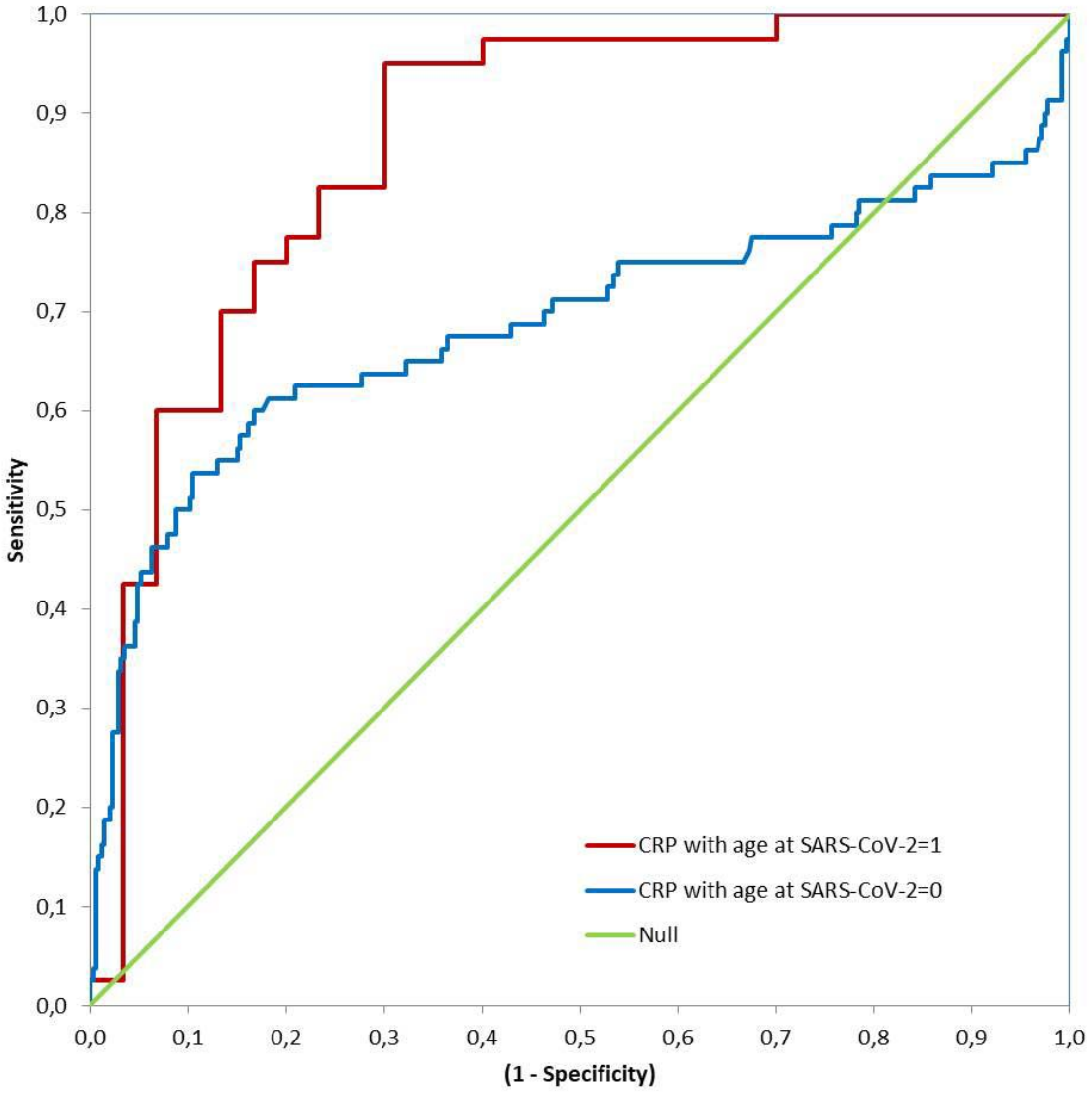

Figure 4 – ROC (Receiver Operating Characteristic) curves of the logistic regression model for hospitalization prediction with C-reactive Protein (CRP) controlled for age quantile at both strata (with and without exposure to SARS-CoV-2)

Legend: Null – Area of the null hypothesis model is 0,5

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

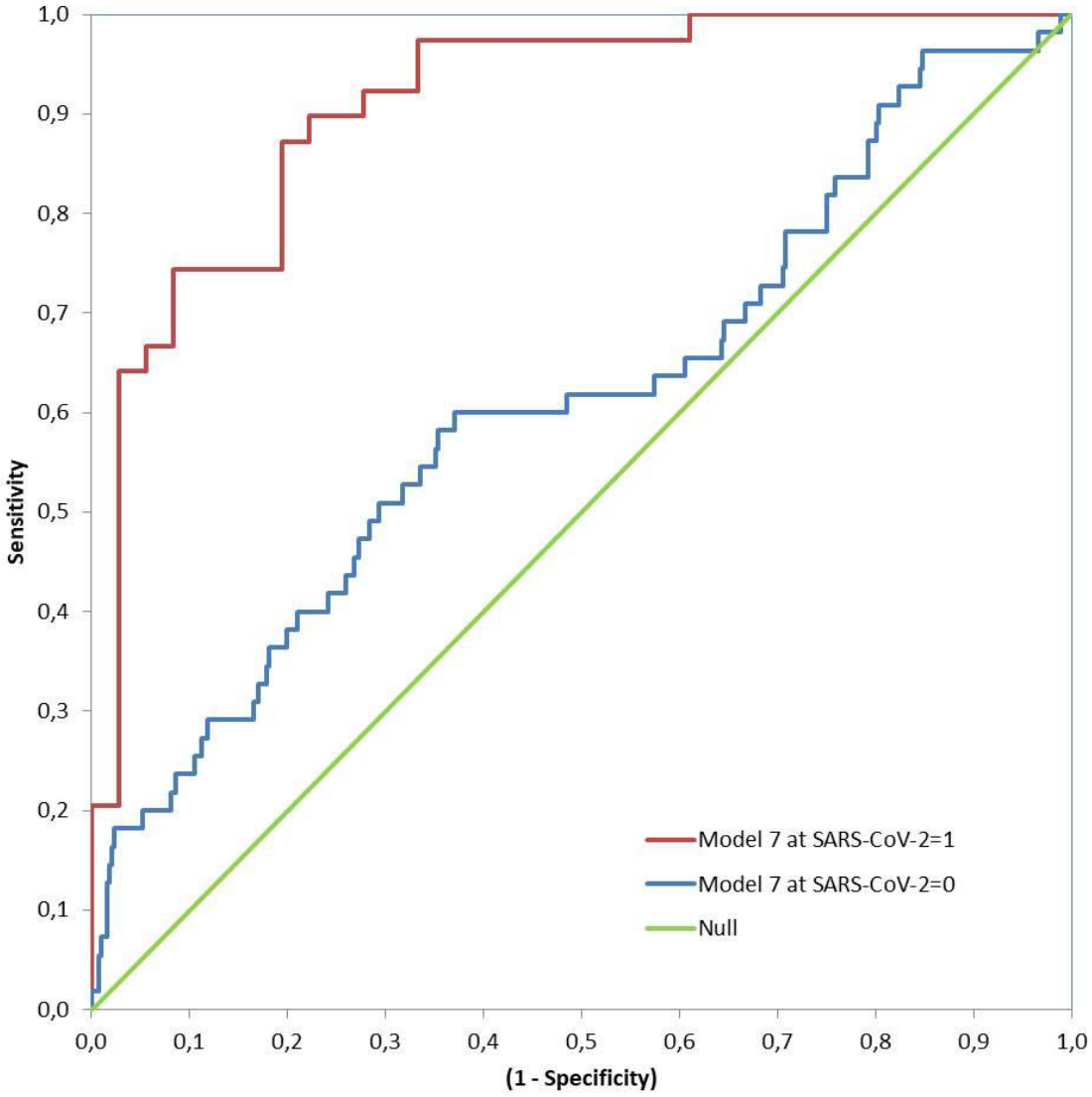

Figure 5 – ROC (Receiver Operating Characteristic) curves of model 7 to predict hospitalization at both strata (with and without exposure to SARS-CoV-2)

Legend: Null – Area of the null hypothesis model is 0,5; Model 7 – Logistic regression with Eosinophils, Monocytes, and Neutrophils controlled for age quantile

medRxiv preprint doi: https://doi.org/10.1101/2020.10.12.20211557; this version posted January 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

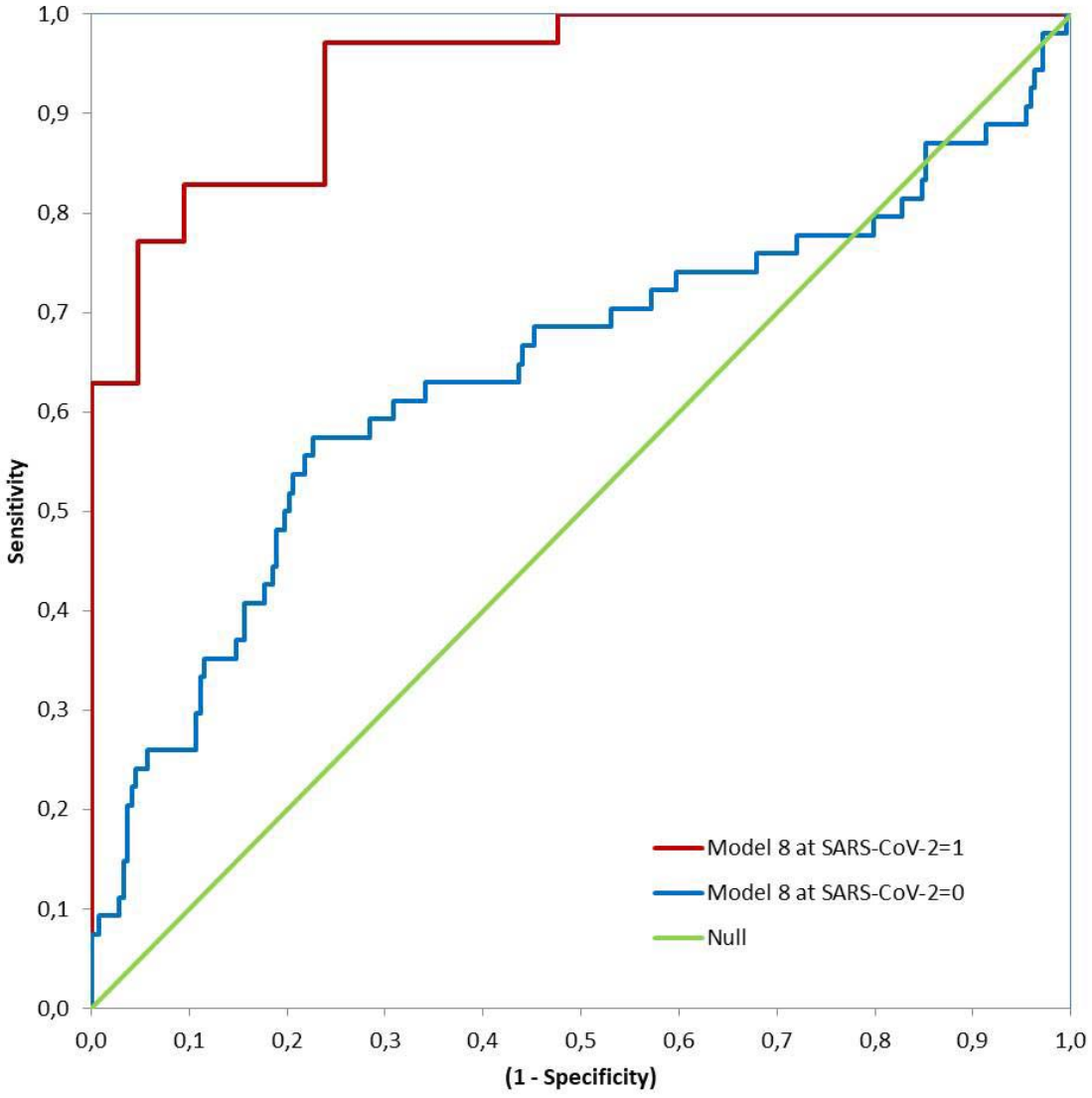

Figure 6 – ROC (Receiver Operating Characteristic) curves of model 8 to predict hospitalization at both strata (with and without exposure to SARS-CoV-2)

Legend: Null – Area of the null hypothesis model is 0,5; Model 8 – Logistic regression with Monocytes and Neutrophils controlled for Creatinine and age quantile