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 Tecnológica Federal do Paraná (UTFPR) in Ponta Grossa, Brazil
2. Researcher enrolled in the postgraduate program 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): 3872

No. of references: 35

No. of tables: Three (3)

No. of figures: Two (2)

**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

## 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 (WHO) 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, comorbidities, multi-comorbidities, and immunosuppression [3,4]. The enlarging list includes diabetes, cardiac disease, chronic lung disease, cerebrovascular disease, chronic kidney disease, cancer, liver disease, obesity, hypertension, dyspnea, fatigue, and anorexia [1,5,6].

Early identification of severe cases allows the optimization of emergency care support [1] and the improvement of 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.

Interest in developing predictive models of COVID-19 outcomes are widespread [7,8]. A review of 50 prognostic models concluded that they are at high risk of bias [8]. As most studies are focused on reporting statistical findings, our concern is with lack of minimum causal criteria to evaluate fragmented findings and to identify potential useful associations that are effectively related to COVID-19 inflammation.

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 [9]. 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 [10] to investigate associations between two sets of outcomes (hospitalization and blood tests) that are related to a common cause (moderate to severe COVID-19 inflammation) considering covariates.

The strength of the association depends on the specificity and sensitivity of the COVID-19 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 is identified and it is also stable and representative of the target population, then these blood tests may be useful as proxies in COVID-19 surveillance protocols and screening interventions.

***Theoretical framework for analyzing associations with causal criteria***

A common use of directed acyclic graph (DAG) in epidemiologic research is to identify sources of bias that may introduce spurious correlations [11,12]. A hypothetical DAG model with latent variable was conceived to evaluate various types of covariates on the focal association, figure 1. The causal path starts with 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 hypothetical covariates that

contribute directly to COVID-19 inflammation were considered risk factors (RF={$RF_1$,…,$RF_L$}, mutual causation relationships [$RF_i$→MSIC←$RF_j$]). Covariates that affect both outcomes are identified as Both-Outcome-Covariate (BOC={$BOC_1$,…,$BOC_m$}) and when affect one outcome as Single-Outcome-Covariate (SOC={$SOC_1$,…,$SOC_n$}). These covariates are not exhaustive but to generate causal graph criteria for handling confounding factors with the d-separation and d-connection concepts [10].

The d-separation concept attempts to separate (make independent) two focal sets of variables by blocking the causal ancestors and by avoiding statistical control for mutual causal descendants [10,13]. Differently, to preserve the association between descendants of MSIC, 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 unwanted dependencies) have to be d-separated (conditionally independent).

Causal relationships in the DAGs are defined with the concept of the *do*(.) operator [10,14]. The association caused by COVID-19 inflammation can be understood as a comparison of the conditional probabilities of hospitalization (H) given a set of blood tests (B) under exposure intervention (*do*(SARS-CoV-2)=1) and without exposure intervention (*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 exposed to SARS-CoV-2. And P[H|B=*b'*,*do*(SARS-CoV-2=0)] if everyone in the population had not been exposed to.

The interventions with *do*(.) generate modified DAGs (or single-world interventions graphs [9]) that allows the analysis of the covariates:

- The *do*(SARS-CoV-2=0) eliminates all arrows directed towards SARS-CoV-2 and to MSIC, because MSIC is assumed to be non-existent without exposure (figure 2). 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 (RF5, BOC1, BOC2).

- Similarly, the modified graph of *do*(SARS-CoV-2=1) is equal to the former graph and adds single arrows from RF1 and RF2 to MSIC; and converts RF3, RF4A, RF4B, and RF5 to fork types with arrows directed to MSIC.

As most covariates are either unmeasured or unknown, their absence can be evaluated following the intuition of the back-door criteria [10,11]:

(a) Covariates with arrow into the causal node (MSIC) are risk factors that may increase the focal association between outcomes by increasing the effect of MSIC; and their absence tends to weaken the focal association.

(b) Single arrow covariates and unbalanced fork covariates into one outcome (H or B) may distort the association and their absence introduces errors in the focal association, reducing the discriminative ability.

(c) Fork covariates into both outcomes (H and B) may add spurious relations (through the back-door) into the focal association, and their absence may inadvertently increase the focal association.

Type (c) introduces non-causally related relations into the focal association and the influence of covariates RF5, BOC1, BOC2 can be estimated with the modified model without exposure (figure 2). A strong association of the outcomes (without exposure) can be due to these covariates and suggest additional efforts to control for them. Another possibility is to exclude the noisy exams that have strong spurious associations.

***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 exposed and the whole population not exposed [9,14]) 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*,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. Similarly, P[H|B=*b'*,SARS-CoV-2=0] can be obtained with another model (same variables but different coefficients). It is implicit that there is the conditioning by a proper set of covariates at each intervention.

The concordance statistic (C-statistic) of a logistic regression model is a standard measure of its predictive accuracy and is calculated as the Area Under of the receiver operating characteristic Curve (AUC) [9,15]. A simple way to compare the discriminative ability of (3) and (4) is to calculate the difference of the AUC at each stratum. A difference of 0.0 means no association with COVID-19 and 0,5 means perfect focal association of the outcomes and perfect differentiation among strata.

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

The assessment of the magnitude of the naïve estimation bias requires further refinements with potential outcomes and in selecting relevant covariates to render the modeling effort analytically tractable to evaluate specific configurations. As a minimum, 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 to assess if there is any difference between them [9], 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, considering the magnitude of the coefficient odds ratio and their statistical significance without and with bootstrap procedure (resampling);
- 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 [16]; 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 observational database in which, at least partially, we could apply the framework and generate candidate prognostic models. Hospital Israelita Albert Einstein (HIAE), in São Paulo – Brazil, made public a database (HIAE_dataset)[17] 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 a mean of zero and a unitary standard deviation. As this research is based on a public and anonymized dataset, it was not revised by any institutional board or ethics commission. The logistic regression models were analyzed with the aid of IBM SPSS version 22.0 and the causal map with DAGitty.net version 3.0.

**RESULTS**

Of the 5644 patients in the HIAE_dataset [17], 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, the following blood tests were discarded because of poor performance of the univariate model when SARS-CoV-2=1: Basophils, Hematocrit, Hemoglobin, Leukocytes, Mean platelet volume, Mean corpuscular hemoglobin (MCH), Mean corpuscular hemoglobin concentration (MCHC), Mean corpuscular volume (MCV), Platelets, Potassium, Red blood Cells, Red blood cell distribution width (RDW), Serum Glucose, Sodium, and Urea (Table 1).

The remaining blood tests are creatinine, C-Reactive Protein (CRP), eosinophils, lymphocytes, monocytes, and neutrophils (Table 1). Only creatinine is not related with the immune system directly and it will be evaluated initially as a risk factor. Of the 5644 patients, eosinophils were recorded for 602 patients, lymphocytes for 602, monocytes for 601, neutrophils for 513, CRP for 506, and creatinine for 424. In dealing with missing cases, all observations with the required data were included (available-case analysis).

CRP is a biomarker of various types of inflammation [18,19]. At SARS-CoV-2=1, the model with CRP and age has good discriminative ability with AUC of ,872 (95% confidence interval (CI), lower bound (LB)=,783; upper bound (UB)=,961). But at SARS-CoV-2=0, AUC is also substantial ,774 (95% CI, LB=,713; UB=,836) with significant overlap between strata at 95% CI. CRP is a predictor of hospitalization in general, but the substantial AUC value at the negative stratum suggests that the focal association due to COVID-19 is contaminated with other non-related associations. Models with CRP demonstrated sensitivity to resampling within the HIAE_dataset [17], the significance of the coefficient moved from ,005 to ,144 (from ,140 to ,148 in other simulations). 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 (LB=,826; UB=,977) and ,755 (LB=,684; UB=,827) in the positive and negative strata, and CRP_reactive, monocytes, neutrophils, and age generated AUC of ,920 (LB=,853; UB=,987) and ,753 (LB=,678; UB=,827), respectively. High levels of AUC at the negative stratum mean that CRP is a response with significant associations due to other causes than COVID-19. Differently from other

prognostic studies [20,21,22,23,24,25] (none used data at negative stratum), CRP was excluded as candidate.

The Neutrophils to Lymphocytes Ratio (NLR) is considered as a possible indicator of severity [21,24,26,27] of COVID-19, but the NLR could not be evaluated with HIAE_dataset [17] as the variables were standardized (division by zero) and were analyzed separately. Lymphocytes presented inconsistent behavior in models with two blood tests. Models with only lymphocytes (with and without age quantile) indicated lymphopenia when SARS-CoV-2=1, as expected [28,29]. The model with lymphocytes, neutrophils and age reversed the sign of the lymphocytes coefficient (SARS-CoV-2=1), possibly, due to collinearity between these blood tests (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. As models with combinations of neutrophils were slightly better than with lymphocyte, lymphocyte was dropped from analysis.

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

Considered individually, eosinophils, monocytes, and neutrophils generated models that have good discriminative performance to estimate the probability of hospitalization (models 1, 2, 3 with AUC>,810 at positive stratum). The combination of these blood tests generates models (4, 5, 6, 7) with better discriminative ability (AUC>,856 at SARS-CoV-2=1). None of these blood tests presented AUC superior to ,680 at SARS-CoV-2=0. All models with two or more blood tests presented a difference of discriminative ability higher than Δ>,220.

Two patterns of associations are 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 focus on models 6, 7, and 8, but models with at least two blood tests (4 to 8) are potential candidate associations.

Models 6 and 8 have significant blood test coefficients at p<.05 (with and without bootstrapping), but model 6 has an intermediate performance in the difference of discriminative ability between strata. Model 7 can also be seen as an extension of model 6 by adding eosinophils into the model. Considering creatinine as a risk factor (as a marker of the renal function), model 8 is the overall best model with significant coefficients at p<,05 and the highest difference of discriminative ability between strata (Δ=,268). This inclusion eliminated the influence of eosinophils from the model and can be considered as an improvement from model 6 (monocytes and neutrophils with age) by adding creatinine.

When the coefficients of model 6 (table 2) are converted to conditional probabilities we find that with monocytes and neutrophils at average, hospitalization probability is >50% for age quantile >11. At average age (quantile 9) and -1 SD (one standard deviation below average) of monocytes (or +1 SD neutrophils) result in hospitalization probability >50%. At the age quantile 15 and -1/2 SD of monocytes (or +1/2 SD of neutrophils) result in hospitalization probability of 86%. Model 7 has similar predictions with monocytes and neutrophils (of model 6) with the addition of eosinophils. When age, monocytes and neutrophils are at average, there is a hospitalization probability of 51,1% with eosinophils at -1 SD; and 90,2% when age quantile is 15.

Model 8 with creatinine has different responses than models 6 and 7. 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). In fact, only below average levels of creatinine lower hospitalization probabilities. Monocytes and neutrophils are also steeper than models 6 and 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%.

Main model biases may be due to contamination with noisy associations and missing cases selection. The AUC at SARS-CoV-2=0 is a simplified measure of the magnitude of the spurious association bias in both outcomes and all models presented relevant noisy associations (AUC from ,588 up to ,679, but not as high as CRP with ,774). 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 be less pronounced than spurious association bias.

**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. It is not intended to predict mortality nor severe to critical cases that require ICU. The risk of overfitting was minimized by not accepting isolated variables in “ad hoc mined” models, and by applying causal criteria to evaluate associations. The AUC evaluation at the negative SARS-CoV-2 stratum to estimate the influence of unwanted covariates 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 (as an operational procedure of the null values in measures of associations for continuous multivariable prognostic models). The selected models are candidates only, the dataset [17] on which they are based cannot be representative beyond the patient health profiles of this hospital. HIAE is a reference hospital in Brazil with practices, standards, and hospitalization criteria that attends a high social-economic segment (mostly living in São Paulo). The observational sample refers to the initial phase of the pandemics in Brazil and the patterns may change with medicine prescriptions and other changes of SARS-CoV-2.

In comparison to other prediction studies, we identified a few focused on the transition from moderate to severe cases of COVID-19 [20,21,22,23,24,25,26,27]. Most studies recommend NLR and CRP as a predictor. None 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" ROC at SARS-CoV-2=0, so that variables with more peculiar responses to COVID-19 were included. CRP is a general marker and not a peculiar response to COVID-19. 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) and so the protocol should be different. To include it in a model, one should control for all other causes of CRP reactive.

We have not included lymphocyte count in the models and have not evaluated the neutrophil to lymphocyte ratio (NLR) as predictor. Lymphocytes and neutrophils are strongly related in this dataset. The similarity of correlations at both strata of SARS-CoV-2 suggests that NLR can also be a noisy association with hospitalization too. If other studies validate that CRP is a noisy predictor (and also possibly NLR), the remaining associations will have less specificity and sensitivity, but at least, they will not generate unreliable COVID-19 predictions.

We evaluated age and creatinine as risk factors. Controlling for age quantile improved the AUC of all models at the positive stratum of SARS-CoV-2. There are other risk factors that can be evaluated with this framework, but not with HIAE_dataset [17], and they could lead to different patterns or enhance a few ones. 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 [30] and whether creatinine (and other renal markers) may be associated with COVID-19 renal inflammatory response [31]. 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 "inflammation" is latent in the DAG, one cannot test key conditional

independencies from this framework (DAGs must be hypothesis driven). Additionally, the usefulness of characteristic associations only due to COVID-19 (when existent) is that they can help in the identification and estimation of risk factors.

As we drop noisy predictors, we are effectively dealing with hypothesis about the physiopathology of COVID-19 inflammation. Even though not as frequent as the mentions of neutrophils, there are studies on the complex role of eosinophils [30,32] and monocytes [33,34] in COVID-19 inflammation indicating eosinopenia in severe cases and monocytopenia in some phase of the cytokine storm and other COVID-19 pathologies [35].

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.

All models can be reproduced by downloading the dataset [17]. More important, we believe that most hospitals (and COVID-19 care centers) in regions affected by the pandemics can apply the framework to generate similar models appropriate to the target population in which they are inserted by making systematic efforts to collect blood tests and potential risk factors at admission together with the SARS-CoV-2 RT-PCR, and other clinical data. Therefore, by making these databases public (anonymized and with standardized data), they will allow future external validation in larger target populations and meta-analysis efforts.

**Acknowledgements:**

We are grateful to Antônio Magno Lima Espeschit and Sônia Mara de Andrade who contributed with suggestions to this research.

**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 [17].

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 Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001”.

**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 – **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

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

10 – **Pearl J**. Causality: models, reasoning, and inference. 2nd ed. Cambridge: Cambridge University Press, 2009.

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

12 – **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

13 – **Hayduk L, et al**. Pearl's d-separation: One more step into causal thinking. *Structural Equation Modeling: A Multidisciplinary Journal*. 2009, 10. 289-311. doi: https://doi.org/10.1207/S15328007SEM1002_8

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

15 – **Hosmer DW, Lemeshow S, Sturdivant RX**. Applied Logistic Regression. 3rd ed. Hoboken: John Willey & Sons, 2013.

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

17 – **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

18 – **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.

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

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 – **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

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 – **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

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*.

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

30 – **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

31 – **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

32 – **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

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

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

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

| | SARS-CoV-2=1 | | | | | | SARS-CoV-2=0 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | B | p | OR | OR 95% C.I. Lower | OR 95% C.I. Upper | N | B | p | OR | OR 95% C.I. Lower | OR 95% C.I. 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

Table 2 – Potential candidate logistic regression models for predicting hospitalization with blood tests and age quantile

| | | SARS-CoV-2 = 1 | | | | | 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: is the coefficient of the variable; p is the significance value of the coefficient; OR is the odds ratio of B (exp(B)); C.I.: confidence interval

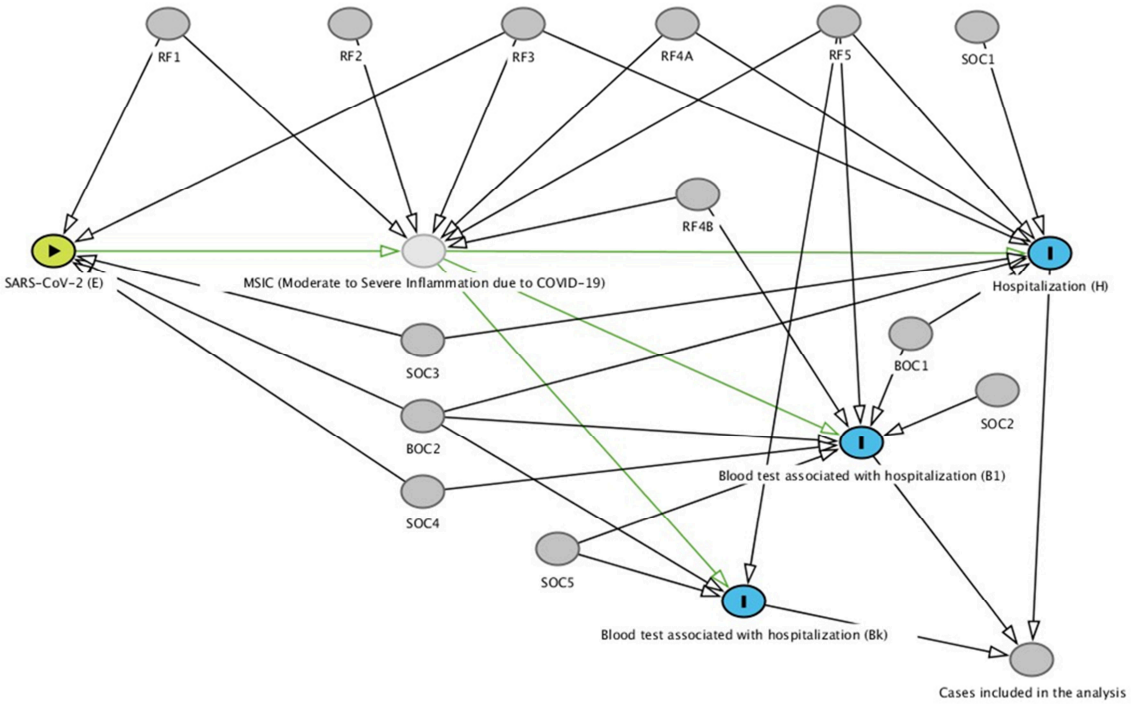
Table 3 – Discriminative ability of potential candidate models for predicting hospitalization from blood tests

| | | | Model | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| Variables included in the model: | | zEosinophils | ● | | | ● | ● | | ● | |
| | | zMonocytes | | ● | | ● | | ● | ● | ● |
| | | zNeutrophils | | | ● | | ● | ● | ● | ● |
| | | Age quantile (0 - 19) | ● | ● | ● | ● | ● | ● | ● | ● |
| | | Creatinine | | | | | | | | ● |
| SARS-CoV-2=1 | AUC | | ,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 = ,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 |
| SARS-CoV-2=0 | AUC | | ,627 | ,588 | ,669 | ,607 | ,678 | ,668 | ,679 | ,672 |
| | Standard Error | | ,037 | ,038 | ,041 | ,038 | ,042 | ,041 | ,042 | ,041 |
| | Asymptotic Significance | | ,000 | ,010 | ,000 | ,002 | ,000 | ,000 | ,000 | ,000 |
| | AUC 95% CI Asymptotic | lower bound | ,554 | ,514 | ,589 | ,533 | ,595 | ,587 | ,596 | ,592 |
| | | upper bound | ,700 | ,663 | ,750 | ,680 | ,761 | ,750 | ,761 | ,752 |
| | Classification table (cut value = ,5) | percentage correct H=0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | | percentage correct H=1 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 7,3 |
| | | Overall percentage | 83,4 | 83,4 | 87,2 | 83,4 | 87,2 | 87,2 | 87,2 | 82,9 |
| | 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) | | | **0,211** | **0,222** | **0,192** | **0,250** | **0,221** | **0,228** | **0,231** | **0,268** |
| Overall discriminative ability order | | | **7** | **5** | **8** | **2** | **6** | **4** | **3** | **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)

Figure 1 – Hypothetical directed acyclic diagram of a COVID-19 inflammation causal path with risk factors and covariates of two types (single outcome and both outcomes)



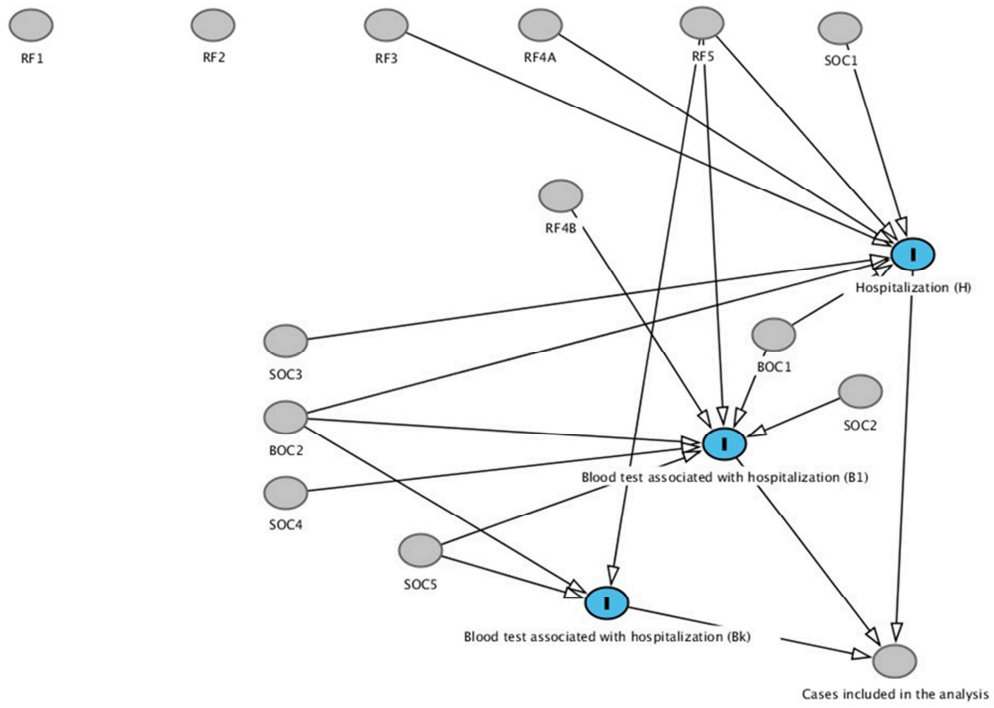

Legend:

Exposure = SARS-CoV-2 (E) (acute respiratory syndrome coronavirus 2)

Outcomes = H: hospitalization (H={regular ward, semi-intensive care,intensive care unit}); B: blood tests ($B=\{B_1,\ldots,B_K\}$)

Covariates = RF: risk factor ($RF=\{RF_1,\ldots,RF_{4A}, RF_{4B},RF_5\}$); SOC: single outcome covariate ($SOC=\{SOC_1,\ldots,SOC_5\}$); BOC: both outcomes covariate ($BOC=\{BOC_1,BOC_2\}$);

Figure 2 – Modified causal path with the operator *do*(SARS-CoV-2=0) to analyze the influence of covariates at the focal outcomes (H and B) without exposure




Legend:

Exposure = SARS-CoV-2 (E) (acute respiratory syndrome coronavirus 2)

Outcomes = H: hospitalization (H={regular ward, semi-intensive care,intensive care unit}); B: blood tests ($B=\{B_1,\ldots,B_K\}$)

Covariates = RF: risk factor ($RF=\{RF_1,\ldots,RF_{4A}, RF_{4B},RF_5\}$); SOC: single outcome covariate ($SOC=\{SOC_1,\ldots,SOC_5\}$); BOC: both outcomes covariate ($BOC=\{BOC_1,BOC_2\}$);