Unsupervised Machine Learning to Identify Convalescent COVID-19 Phenotypes

Sarah Adamo
Department of Electrical Engineering and Information Technology
University of Naples "Federico II"
Naples, Italy
sarah.adamo@unina.it

Carlo Ricciardi
Department of Electrical Engineering and Information Technology
University of Naples "Federico II"
Naples, Italy
carlo.ricciardi@unina.it

Pasquale Ambrosino
Cardiac Rehabilitation Unit
Istituti Clinici Scientifici Maugeri IRCCS
Telese Terme, Italy
pasquale.ambrosino@icsmaugeri.it

Mauro Maniscalco
Pulmonary Rehabilitation Unit
Istituti Clinici Scientifici Maugeri IRCCS
Telese Terme, Italy
mauro.maniscalco@icsmaugeri.it

Arcangelo Biancardi
Bioengineering Unit
Istituti Clinici Scientifici Maugeri IRCCS
Telese Terme, Italy
arcangelo.biancardi@icsmaugeri.it

Giuseppe Cesarelli
Department of Chemical, Materials and Production Engineering
University of Naples "Federico II"
Naples, Italy;
giuseppe.cesarelli@unina.it

Leandro Donisi
Department of Chemical, Materials and Production Engineering
University of Naples "Federico II"
Naples, Italy;
leandro.donisi@unina.it

Giovanni D'Addio
Bioengineering Unit
Istituti Clinici Scientifici Maugeri IRCCS
Telese Terme, Italy
gianni.daddio@icsmaugeri.it

Abstract— After the acute disease, post-COVID-19 patients may present several and persistent symptoms, known as the new paradigm of "post-acute COVID-19 syndrome". This necessitates a multidisciplinary rehabilitation that has been proposed but whose effectiveness is still to be assessed. In this study, convalescent COVID-19 patients undergoing pulmonary rehabilitation (PR) after reporting long-term symptoms were consecutively enrolled. Then, they were grouped by laboratory parameters at admission through an unsupervised Machine Learning (ML) approach. We aimed to identify potential indicators that could discriminate several phenotypes leading to a different responsiveness to the rehabilitation program. A k-means clustering method was performed; then, statistical analysis was employed to compare clinical and hematochemical parameters of the obtained clusters. The dataset consisted of 78 patients (84.8% males, mean age 60.72 years). The optimal number for clustering was k=2 with a silhouette coefficient of 0.85, and D-Dimer resulted the most discriminating parameter, thus confirming its role as a marker of inflammation. The phenotypes exhibited statistically significant differences in terms of age (p=0.007), packs of cigarettes per year (p=0.003), uricemia (p=0.010), PCR (p=0.026), D-Dimer (p<0.001), red blood cells (p=0.005), hemoglobin (p=0.039), hematocrit (p=0.026), PaO_2 (p=0.006), SpO_2 (p=0.011). Overall, our findings suggest the effectiveness of ML in identifying personalized prevention, interventional and rehabilitation strategies.

Keywords—COVID-19, machine learning, rehabilitation, exercise, disability, outcome

I. INTRODUCTION

The coronavirus disease 2019 (COVID-19) is a syndrome characterized by several clinical manifestations that could have different complications, ranging from mild to severe symptoms [1]. Convalescent COVID-19 patients may present long-term effects with persistent symptoms, thus introducing the new paradigm of a "post-acute COVID-19 syndrome" [2]. Consequently, the need for an early and multidisciplinary rehabilitation has been proposed [3-6], even though the effectiveness of this approach in the post-acute care setting is still to be assessed [7].

In the last years, Machine Learning (ML) has been widely used to investigate several diseases, since it has been showed to be an effective method to explore clinical parameters, predict outcomes and, consequently, identify useful indexes [8-10]. Recently, it proved its usefulness in several specialties: cardiology, neurology, rehabilitation engineering and in analyzing data, signals and images [11-15].

ML can be divided into supervised and unsupervised learning. In this paper, we are focused on the second one, which is useful for finding hidden patterns among unlabeled data based on distance criteria, while supervised learning is employed for classification and regression modeling [16]. Recently, unsupervised ML has been used for studying COVID-19 disease in order to find hidden patterns and information about most relevant factors and comorbidities related to a most severe form of the disease [17-19]. Supervised ML and classification algorithms have been implemented to predict COVID-19 prognosis and outcomes basing on hematochemical parameters [20]. Nevertheless, no one combined the usage of clustering methods with blood tests for analyzing convalescent COVID-19 patients.

The aim of this study is to identify and analyze different phenotypes of post-COVID-19 patients undergoing rehabilitation through a clustering method by evaluating hematochemical parameters. This preliminary study could be useful to identify potential indicators that could play a significative role in the improvement of health conditions in patients affected by the above-mentioned disease.

II. MATERIALS AND METHOD

A. Study Population

The cohort was composed of 78 patients (84.8% males, mean age 60.72 years) admitted to a post-COVID-19 rehabilitation program after being discharged from COVID-19 acute care ward and after reporting long-term symptoms. Consecutive patients referring to the Pulmonary Rehabilitation Unit of the Istituti Clinici Scientifici Maugeri (ICS) Maugeri Spa SB, IRCCS of Telese Terme, Benevento, Italy, were screened for enrollment. Inclusion criteria were: recent history of COVID-19, confirmed by positive nasopharyngeal swab test for SARS-CoV-2; negativization of at least two swab tests within the previous two months. Patients aged <18 years or unwilling to enter the study were excluded.

B. Study procedures

After informed consent signature, the main demographic and clinical characteristics were systematically collected in all included patients and stored in an electronic database. After de-identification, all collected data were extracted for analyses.

In brief, a venous blood sample was collected from all patients in the fasted state to measure blood count and the main hematochemical parameters. A blood gas analyser (ABL 825® FLEX BGA, Radiometer Medical Aps, Copenhagen, Denmark) was used to measure arterial oxygen (PaO_2) and carbon dioxide tension ($PaCO_2$).

C. ML and Statistical Analyses

Preliminary assessment through Matlab software (*v. 2021b*) and unsupervised ML through KNIME Analytics platform (*v. 4.5.0*) were performed. KNIME is an already well-known platform for developing workflows which allow users to implement ML analyses and manipulations of data, and it has already been used for several biomedical studies [21-25].

A k-means clustering algorithm was implemented to evaluate if different phenotypes could be distinguished basing on clinical parameters of post-COVID-19 patients. K-means is a very simple and iterative unsupervised learning algorithm that assigns *n* similar data to *k* clusters to identify hidden patterns, by minimizing the in-cluster sum of square. Each cluster is represented by a centroid [26, 27]. Since the correct number of clusters is not known a priori, the optimal *k* was determined evaluating clusters through silhouette criterion, a score that calculates the goodness of clustering technique depending on how groups are well separated and clearly distinguished, and its value ranges from -1 to 1 [28, 29]. The silhouette was calculated through Euclidean distance. Then, parallel coordinates plot was used to visualize multidimensional data mapped in a 2-dimensional set [30, 31].

After obtaining clustered data, a univariate statistical analysis was performed in IBM Spss (*v. 27*) to compare the obtained groups. The normality distribution of the data was assessed with the Kolmogorov-Smirnov test. The Levene's test was performed to assess the homoscedasticity of variances between groups for normally distributed data, and then a t-test for independent samples was performed; otherwise, a Mann-Whitney was employed.

III. RESULTS

A. Study population

The study population consisted of 78 convalescent COVID-19 patients (84.8% males, mean age 60.72 years). In Table 1, the baseline demographic and clinical characteristics, including those pertaining to the acute phase of COVID-19, have been reported.

TABLE I.   BASELINE DEMOGRAPHIC AND CLINICAL FEATURES OF POST-ACUTE COVID-19 PATIENTS

Patients (n)	78
Age, years	60.72 ± 10.69
Female, N (%)	14 (15)
Smokers, N (%)	8 (10)
BMIª, kg/m²	28.65 ± 5.35
Hospitalization length, days	27.53 ± 9.04
High flow oxygen, N (%)	24 (31)
Mechanical ventilation, N (%)	17 (22)
Hypertension, N (%)	35 (45)
Hypercholesterolemia, N (%)	7 (9)
Hypertriglyceridemia, N (%)	4 (5)
Diabetes, N (%)	11 (14)
Heart failure, N (%)	8 (10)
Atrial fibrillation, N (%)	5 (6)
History of stroke/TIAᵇ, N (%)	3 (4)

a. BMI: Body Mass index; b. TIA: Transient Ischemic Attack

B. ML results

Table 2 shows the silhouette coefficients related to the number of clusters from 2 to 4, where the optimal result was obtained for *k*=2.

TABLE II.   SILHOUETTE COEFFICIENTS FOR NUMBER OF CLUSTERS

	k=2	*k=3*	*k=4*
Silhouette Coefficient	**0.85**	0.83	0.61

Parallel coordinates plot highlighted that the D-Dimer parameter well discriminated the two clusters, as shown in Figure 1, where D-Dimer was represented by feature 16.

Figure 2 shows PaO_2 data clustering based on D-Dimer values, with PaO_2 representing an index for blood oxygenation.
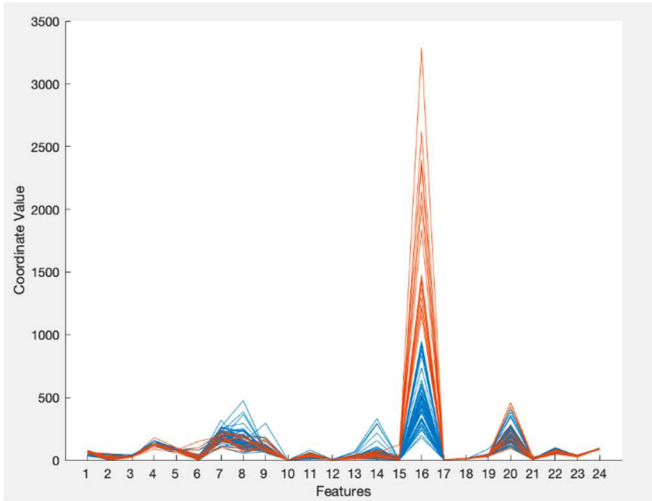
Fig. 1. Parallel coordinates, where the clusters are represented with different colours.
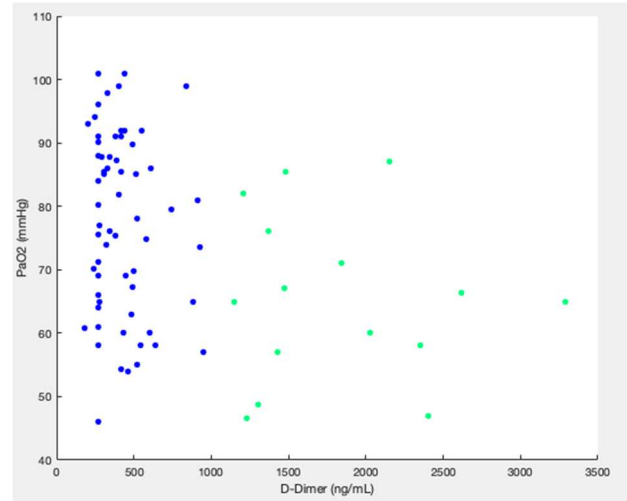

Fig. 2. Clustering for arterial oxygen tension (PaO$_2$) depending on D-Dimer.

C. Statistical Results

Table 3 shows clinical and laboratory parameters of the two phenotypes obtained through k-means clustering. Results pointed out that patients belonging to cluster 0, with worst health conditions in admission, were older (p=0.007) and smoked more cigarettes per year (p=0.003). Patients of cluster 1 showed a lower value of PCR (p=0.026) and higher values of red blood cells (p=0.005), hemoglobin (p=0.039), hematocrit (p=0.026), PaO$_2$ (p=0.006) and SpO$_2$ (p=0.011). Furthermore, the most discriminating parameter between two groups was the D-Dimer: patients of cluster 0 exhibited a much higher value than cluster 1 with a significance at 0.001.

TABLE III.   COMPARING HEMATOCHEMICAL FEATURES BETWEEN THE TWO CLUSTERS

Features (units of measurement)	Cluster 0 (n=15)	Cluster 1 (n=63)	p-value
Age, years	67.47 ± 7.00	59.21 ± 10.90	**0.007**
Female, n (%)	3 (20)	11 (16)	Ns
BMIa, m^2/kg	26.52 ± 3.82	29.22 ± 5.57	Ns
Hospitalization, days	19.13 ± 14.15	16.29 ± 14.75	Ns
Cigarettes, packs per year	35.80 ± 43.03	10.56 ± 19.37	**0.003**
TC^b, mg/dl	174.53 ± 41.68	188.14 ± 41.52	Ns
TGs^c, mg/dL	132.07 ± 52.30	164.17 ± 75.16	Ns
Glycemia, mg/dL	108.60 ± 36.75	91.97 ± 33.51	Ns
Creatinine, mg/dL	0.77 ± 0.18	0.81 ± 0.14	Ns
Urea, mg/dL	42.60 ± 11.51	37.94 ± 11.97	Ns
Uricemia, mg/dL	4.03 ± 1.54	5.23 ± 1.59	**0.010**
AST^d, UI/L	23.07 ± 11.70	21.70 ± 10.90	Ns
ALT^e, UI/L	44.13 ± 19.94	58.89 ± 65.34	Ns
PCR^f, mg/dL	18.68 ± 31.16	6.18 ± 7.60	**0.026**
D-Dimer, ng/mL	1821.33 ± 632.67	419.84 ± 186.14	**<0.001**
Red blood cells, 10^6/mL	4.11 ± 0.60	4.57 ± 0.53	**0.005**
Hemoglobin, g/dL	12.04 ± 2.02	13.03 ± 1.54	**0.039**
Hematocrit, %	36.64 ± 5.44	40.48 ± 7.60	**0.026**
Platelet, 10^3/mL	242.20 ± 119.318	204.03 ± 66.56	Ns
Leukocyte, 10^3/mL	9.16 ± 5.33	7.84 ± 2.56	Ns
PaO$_2$^{g}, mmHg	65.47 ± 13.13	77.62 ± 14.09	**0.006**
PaCO$_2$^{h}, mmHg	34.84 ± 3.97	36.32 ± 3.34	Ns
SpO$_2$^{i}, %	92.53 ± 4.17	95.11 ± 3.07	**0.011**

[a] BMI: body mass index; [b] TC: total cholesterol; [c] TG: Triglycerides; [d] AST: aspartate aminotransferase; [e] ALT: alanine aminotransferase; [f] CRP: C-reactive protein; [g] PaO$_2$: arterial oxygen tension; [h] PaCO$_2$: arterial carbon dioxide tension; [i] SpO$_2$: oxygen saturation. Ns: not significant.

IV. Discussion and Conlusion

Our clustering model effectively identified two groups of convalescent COVID-19 patients according to functional and laboratory parameters. In line with literature evidence, patients in cluster 0 showed a number of clinical and demographic features suggesting a worse clinical status and an inflammatory profile [32]. In detail, higher PCR values were documented in this subset of patients., which was also characterized by an older age. Accordingly, age has been identified as one of the most important risk factors for severe-to-critical COVID-19 [33]. Moreover, cluster 0 was characterized by lower PaO_2 as well as SpO_2 values, which is consistent with the lower hemoglobin levels and a lower number of red blood cells. Most importantly, D-dimer was the parameter showing the highest discriminating ability between the two clusters. The role of D-dimer as a marker of inflammation and coagulation is widely accepted [34]. Recently, it has been demonstrated that COVID-19 severity is associated with several laboratory modifications, including a consistent increase in D-dimer levels as well as other coagulation markers (*e.g.*, fibrinogen, prothrombin time) [35]. Such blood modifications reflect the procoagulant state of COVID-19 patients, confirmed by the high risk of venous and arterial thrombosis both in the acute phase and during convalescence [36, 37].

Regarding the employment of ML, it has been used in this research to identify two phenotypes of post-COVID-19 patients, but this approach could be employed also in other medical pathologies where the distinction among phenotypes has not been already defined. This study could be extended including parameters both in admission and in dismission to the rehabilitation program. Therefore, a supervised ML approach could be employed to validate the identified indexes in the current study and to evaluate further potential rehabilitation outcomes in convalescent COVID-19 patients.

Overall, our findings suggest the effectiveness of ML for patient stratification in the post-acute care setting (*e.g.*, community, hospital, rehabilitation), thus potentially contributing to the identification of personalized prevention, interventional and rehabilitation strategies for COVID-19.

References

[1] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet, 395(10223), 497-506..

[2] Amdal, C. D., Pe, M., Falk, R. S., Piccinin, C., Bottomley, A., Arraras, J. I., ... & Bjordal, K. (2021). Health-related quality of life issues, including symptoms, in patients with active COVID-19 or post COVID-19; a systematic literature review. Quality of Life Research, 30(12), 3367-3381.

[3] Ambrosino, P., Fuschillo, S., Papa, A., Di Minno, M. N. D., & Maniscalco, M. (2020). Exergaming as a supportive tool for home-based rehabilitation in the COVID-19 pandemic era. Games for health journal, 9(5), 311-313.

[4] Gloeckl, R., Leitl, D., Jarosch, I., Schneeberger, T., Nell, C., Stenzel, N., ... & Koczulla, A. R. (2021). Benefits of pulmonary rehabilitation in COVID-19: a prospective observational cohort study. ERJ open research, 7(2).

[5] Buckley, B. J., Harrison, S. L., Fazio-Eynullayeva, E., Underhill, P., Jones, I. D., Williams, N., & Lip, G. Y. (2021). Exercise rehabilitation associates with lower mortality and hospitalisation in cardiovascular disease patients with COVID-19. European Journal of Preventive Cardiology.

[6] Spruit, M. A., Holland, A. E., Singh, S. J., Tonia, T., Wilson, K. C., & Troosters, T. (2020). COVID-19: interim guidance on rehabilitation in the hospital and post-hospital phase from a European Respiratory Society-and American Thoracic Society-coordinated international task force. European Respiratory Journal, 56(6).

[7] Demeco, A., Marotta, N., Barletta, M., Pino, I., Marinaro, C., Petraroli, A., ... & Ammendolia, A. (2020). Rehabilitation of patients post-COVID-19 infection: a literature review. Journal of International Medical Research, 48(8), 0300060520948382.

[8] Inan, O. T., Baran Pouyan, M., Javaid, A. Q., Dowling, S., Etemadi, M., Dorier, A., ... & Klein, L. (2018). Novel wearable seismocardiography and machine learning algorithms can assess clinical status of heart failure patients. Circulation: Heart Failure, 11(1), e004313.

[9] Lin, W. Y., Chen, C. H., Tseng, Y. J., Tsai, Y. T., Chang, C. Y., Wang, H. Y., & Chen, C. K. (2018). Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. International journal of medical informatics, 111, 159-164.

[10] Xie, Y., Jiang, B., Gong, E., Li, Y., Zhu, G., Michel, P., ... & Zaharchuk, G. (2019). Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. American Journal of Roentgenology, 212(1), 44-51.

[11] Donisi, Ricciardi, C., Cesarelli, G., Coccia, A., Amitrano, F., Adamo, S., & D'Addio, G. (2022). Bidimensional and Tridimensional Poincaré Maps in Cardiology: A Multiclass Machine Learning Study. Electronics, 11(3), 448. https://doi.org/10.3390/electronics11030448

[12] Ricciardi, C., Amboni, M., De Santis, C., Ricciardelli, G., Improta, G., Iuppariello, L., ... & Cesarelli, M. (2019, September). Classifying different stages of Parkinson's disease through random forests. In Mediterranean Conference on Medical and Biological Engineering and Computing (pp. 1155-1162). Springer, Cham. https://doi.org/10.1007/978-3-030-31635-8_140

[13] Jambukia, S. H., Dabhi, V. K., & Prajapati, H. B. (2015, March). Classification of ECG signals using machine learning techniques: A survey. In 2015 International Conference on Advances in Computer Engineering and Applications (pp. 714-721). IEEE.

[14] Yaman, E., & Subasi, A. (2019). Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification. BioMed research international, 2019.

[15] Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine learning for medical imaging. Radiographics, 37(2), 505-515.

[16] Mitchell, T. M. Machine learning. 1997. Burr Ridge, IL: McGraw Hill, 1997.

[17] Hassan, B. A., Rashid, T. A., & Hamarashid, H. K. (2021). A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star. Computers in biology and medicine, 138, 104866.

[18] Geva, A., Patel, M. M., Newhams, M. M., Young, C. C., Son, M. B. F., Kong, M., ... & Overcoming COVID-19 Investigators. (2021). Data-driven clustering identifies features distinguishing multisystem inflammatory syndrome from acute COVID-19 in children and adolescents. EClinicalMedicine, 40, 101112.

[19] King, B., Barve, S., Ford, A., & Jha, R. (2020, August). Unsupervised clustering of COVID-19 chest X-ray images with a self-organizing feature map. In 2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS) (pp. 395-398). IEEE.

[20] Altini, N., Brunetti, A., Mazzoleni, S., Moncelli, F., Zagaria, I., Prencipe, B., ... & Bevilacqua, V. (2021). Predictive Machine Learning Models and Survival Analysis for COVID-19 Prognosis Based on Hematochemical Parameters. Sensors, 21(24), 8503.

[21] Scrutinio, D., Ricciardi, C., Donisi, L., Losavio, E., Battista, P., Guida, P., ... & D'Addio, G. (2020). Machine learning to predict mortality after rehabilitation among patients with severe stroke. Scientific reports, 10(1), 1-10. 10:20127.

[22] Improta, G., Ricciardi, C., Cesarelli, G., D'Addio, G., Bifulco, P., & Cesarelli, M. (2020). Machine learning models for the prediction of acuity and variability of eye-positioning using features extracted from oculography. Health and Technology, 10(4), 961-968. https://doi.org/10.1007/s12553-020-00449-y

[23] Ricciardi, C., Improta, G., Amato, F., Cesarelli, G., & Romano, M. (2020). Classifying the type of delivery from cardiotocographic signals: A machine learning approach. Computer Methods and Programs in Biomedicine, 196, 105712. https://doi.org/10.1016/j.cmpb.2020.105712

[24] Donisi, L., Ricciardi, C., Cesarelli, G., Pagano, G., Amitrano, F., & D'Addio, G. (2020, July). Machine Learning applied on Poincaré Analyisis to discriminate different cardiac issues. In 2020 11th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO) (pp. 1-2). IEEE. doi: 10.1109/ESGCO49734.2020.9158144

[25] Ricciardi, C., Donisi, L., Cesarelli, G., Pagano, G., Coccia, A., & D'Addio, G. (2020, July). Feasibility of machine learning applied to poincaré plot analysis on patients with CHF. In 2020 11th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO) (pp. 1-2). IEEE. doi: 10.1109/ESGCO49734.2020.9158152

[26] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern recognition, 36(2), 451-461.

[27] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics), 28(1), 100-108.

[28] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

[29] Ogbuabor, G., & Ugwoke, F. N. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. AIRCC's International Journal of Computer Science and Information Technology, 10(2), 27-37.

[30] Inselberg, A., & Dimsdale, B. (1990, October). Parallel coordinates: a tool for visualizing multi-dimensional geometry. In Proceedings of the First IEEE Conference on Visualization: Visualization90 (pp. 361-378). IEEE.

[31] Inselberg, A., & Dimsdale, B. (2009). Parallel coordinates. Human-Machine Interactive Systems, 199-233.

[32] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., ... & Tan, W. (2020). A novel coronavirus from patients with pneumonia in China, 2019. New England journal of medicine.

[33] Guan, W. J., Ni, Z. Y., Hu, Y., Liang, W. H., Ou, C. Q., He, J. X., ... & Zhong, N. S. (2020). Clinical characteristics of coronavirus disease 2019 in China. New England journal of medicine, 382(18), 1708-1720.

[34] Di Minno, M. N. D., Calcaterra, I., Papa, A., Lupoli, R., Di Minno, A., Maniscalco, M., & Ambrosino, P. (2021). Diagnostic accuracy of D-Dimer testing for recurrent venous thromboembolism: A systematic review with meta-analysis.: VTE recurrence and D-dimer. European Journal of Internal Medicine, 89, 39-47.

[35] Di Minno, Calcaterra, I., Lupoli, R., Storino, A., Spedicato, G. A., Maniscalco, M., Di Minno, A., & Ambrosino, P. (2020). Hemostatic Changes in Patients with COVID-19: A Meta-Analysis with Meta-Regressions. Journal of Clinical Medicine, 9(7), 2244. https://doi.org/10.3390/jcm9072244

[36] Ambrosino, P., Papa, A., Maniscalco, M., & Di Minno, M. N. D. (2021). COVID-19 and functional disability: current insights and rehabilitation strategies. Postgraduate Medical Journal, 97(1149), 469-470.

[37] Liu, P. P., Blet, A., Smyth, D., & Li, H. (2020). The science underlying COVID-19: implications for the cardiovascular system. Circulation, 142(1), 68-78.