Jundishapur J Health Sci. 2023 April; 15(2):e136913. https://doi.org/10.5812/jjhs-136913.

Published online 2023 July 22. **Methods Article**

Datasets Created from Routine Laboratory Parameters for Use in the Diagnosis, Prognosis, and Mortality of COVID-19

Mehmet Tahir HUYUT [1, *]

[1] Erzincan Binali Yıldırım University, Erzincan, Turkey

[*] *Corresponding author*: Department of Biostatistics and Medical Informatics, Faculty of Medicine, Erzincan Binali Yıldırım University, Erzincan, Turkey. Email: mehmettahirhuyut@gmail.com

Received 2023 April 18; **Revised** 2023 June 24; **Accepted** 2023 July 03.

Abstract

It is important to diagnose coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 at an early stage and to monitor severely infected patients in order to reduce the lethality of the disease. In addition, there is a need for alternative methods with lower costs and faster results to determine the severity of the disease. In this context, routine blood values can be used to determine the diagnosis/prognosis and mortality of COVID-19. In this study, three optimized datasets were prepared to determine the features that affect the diagnosis, prognosis, and mortality of COVID-19. These datasets can be used by researchers to determine the diagnosis and severity of COVID-19 with various classifier machine learning models and artificial intelligence methods. It is hoped that studies on these datasets will reduce the negative pressures on the health system and provide important clinical guidance for decision-makers in the diagnosis and prognosis of COVID-19.

Keywords: COVID-19, Diagnosis, Prognosis, Mortality, Biochemical and Hematological Biomarkers, Routine Blood Values, Feature Selection, Artificial Intelligence, Machine Learning, Neural Network

1. Background

The scientific community has worked hard to reduce the impact and scope of the current coronavirus disease 2019 (COVID-19) outbreak (1). The early diagnosis of the disease and evaluation of its development is extremely important for the timely implementation of medical protocols (2-4). In this context, the datasets were prepared to determine both the diagnosis/prognosis/mortality of COVID-19 with only routine blood value (RBV) data and to reveal which features are important accordingly. In many previous studies on the diagnosis/prognosis and mortality of COVID-19, the clinical importance of RBV data was noted (5-13).

Routine blood parameter values are at the forefront of reliable, fast, and economical methods in determining the diagnosis, prognosis, and mortality of diseases (14, 15). Routine blood parameter values in these datasets provide critical information for the diagnosis and determination of the severity of various diseases (16). Especially for severe and fatal cases of various diseases, RBV characteristics provide important information about early predictive factors (9, 11, 12, 17). Researchers can use these datasets with different classification models to determine the diagnosis/prognosis and mortality of COVID-19. In addition, researchers can reveal the relationship structures between the RBV features in these datasets and the diagnosis/prognosis/mortality of COVID-19 with various artificial intelligence (AI) models.

It is crucial to detect the diagnosis and prognosis of COVID-19 induced by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the early period in order to reduce the lethality of the disease. It is known that diagnostic tests for the detection of COVID-19 require special equipment and facilities, and it is necessary to wait at least 4 - 5 hours for the result. In addition, advanced examinations, such as computed tomography, are required to determine the severity of the disease (11, 12). Reducing the cost in the fight against this epidemic and not delaying the intervention with early diagnosis are important for patients and health systems (14-18). For this purpose, it has been reported that diagnoses and prognoses based on RBV, which are less costly and faster to assist clinical procedures, might be an alternative source for COVID-19 (10, 13, 19, 20).

In this context, the datasets were prepared to

determine both the diagnosis/prognosis/mortality of COVID-19 with only RBV data and to reveal which features are important accordingly. These datasets can be used by researchers in determining the diagnosis and severity of COVID-19 with various classifier machine learning (ML) models and AI methods and be compared to the success of the methods in the present original research article.

2. Methods

To detect and predict the prognosis of COVID-19 without using advanced equipment and methods, three datasets were created by scanning the retrospective records within April and December 2021 from the information system of Erzincan Binali Yıldırım University Mengücek Gazi Training and Research Hospital, Erzincan, Turkey. During the dates covered, SARS-CoV-2 was diagnosed in the studied hospital only by real-time reverse transcriptase polymerase chain reaction using nasopharyngeal or oropharyngeal swabs. The patients in these datasets are of Turkish and Kurdish origins. The RBV data at first admission were recorded to prevent various complications. Table 1 shows the characteristics of the patients included in these datasets. These features include biochemical, hematological, and immunological RBVs.

In the raw data obtained by scanning the digital records from the EBYU-MG patient system during the specified dates, 68 different RBVs measured at the time of application of patients were obtained. In the datasets, 68 RBV parameters were examined, and RBVs (properties) measured from at least 80% of patients were calibrated and used. Additionally, in the raw data records, diagnosis (e.g., COVID-19, hypertension, diabetes, and chronic obstructive pulmonary disease), treatment unit (i.e., intensive care unit [ICU] or non-ICU), and demographic characteristics (e.g., age, gender, smoking, and alcohol use) were available. Comorbidity data were not available for most patients as it was a retrospective dataset (with a lack of comorbidity data in more than 80% of patients). In the analyzed data, categorical data were coded; repeated measurements and missing data were filled with the mean of the relevant parameter, and quantitative data were normalized. The RBV and age data were on a quantitative scale; diagnostic data were on a multinomial scale, and treatment unit and gender were on a binomial scale. Then, individuals over the age of 18 years were filtered from the patients diagnosed with COVID-19. Among all patients, those diagnosed with COVID-19 were filtered out. The information of approximately 66,000,000 patients was recorded, and the entire recording period took 20 hours. The datasets were converted to ".sav" (IBM SPSS Statistics format).

For the measurement of biochemical values, Beckman Coulter Olympus AU2700 Plus Chemistry Analyzer (Beckman Coulter, Tokyo, Japan) device was used, and analyses were performed with spectrophotometric tests. Hematological values were measured from the cell blood count using the Sysmex XN-1000 Hematology System (Sysmex Corporation, Japan).

Serum prothrombin time, activated partial prothrombin time, and fibrinogen values, which are immunological values, were analyzed using a Ceveron-Alpha digital coagulation device (Diapharma Group Inc., West Chester, Canada). The erythrocyte sedimentation rate was measured by the photometric capillary flow kinetic analysis using the TEST 1 BCL instrument (Alifax, Polverara, Italy). Ferritin levels were measured with a chemiluminescence immunoassay device (Centaur XP, Siemens Healthcare, Germany). C-reactive protein levels were evaluated by the nephelometric method using the BNTM II instrument (Siemens, Munich, Germany). Procalcitonin, D-dimer, and troponin levels were analyzed from whole blood using AQT90 flex Radiometer VR (Bronshoj, Denmark).

3. Results

The first dataset was named COVID-19_RBV1. This dataset includes 51 features of 2000 COVID-19 positive and 2000 COVID-19 negative patients. In this dataset, these data can be used to diagnose COVID-19, as the class labels were created as patient and intact.

The second dataset was named COVID-19_RBV2. This dataset includes 51 features of 279 and 1721 ICU and non-ICU COVID-19 patients, respectively. The ICU and non-ICU patients were defined as severely and mildly infected, respectively. Since the class labels as ICU and non-ICU were created in this dataset, these data can be used to identify important features in the prognosis of the disease.

The third dataset was named COVID-19_RBV3. For the third dataset, the exclusion information of COVID-19 patients within April and December 2021 was examined. The RBV records of patients who died from COVID-19 and survived were searched, and information on 38 RBVs was obtained. This dataset includes information on 233 patients who died from COVID-19 and 2364 patients who survived on the specified dates. In this dataset, these data can be used to determine the mortality of the disease, as the class labels were created as surviving and deceased from COVID-19.

The dataset contains three following files:

Appendix 1 in Supplementary File: For the first dataset, those who tested positive for COVID-19 were labeled as patients, and those who tested negative were labeled as

healthy (control group). Calibrated 51 RBV features were used for the first dataset. In the COVID-19_RBV1 dataset (i.e., the first dataset), COVID-19-infected patients and healthy individuals were coded as 1 and 0, respectively (COVID-19 positive: 1, COVID-19 negative: 0; first column: Diagnosis; other columns: RBV features).

Appendix 2 Supplementary File: For the second dataset, COVID-19 patients were labeled as ICU and non-ICU according to treatment units. In the second data, ICU and non-ICU patients were defined as severely and mildly infected, respectively. Calibrated 51 RBV features were used for the second dataset. In the COVID-19_RBV2 dataset (i.e., the second dataset), severely and mildly COVID-19 infected patients were coded as 2 and 1, respectively (severely COVID-19 group: 2, mildly COVID-19 group: 1; first column: Treatment units/service [ICU or non-ICU]; other columns: RBV features).

Appendix 3 Supplementary File: For the third dataset, the exclusion information of COVID-19 patients within April and December 2021 was examined. The RBV data of patients who died from COVID-19 and survived were calibrated, and 38 RBV data were used in the third dataset. In the COVID-19_RBV3 dataset (i.e., the third dataset), patients who died (non-survived COVID-19) were coded as 1, and living patients (survived COVID-19) were coded as 0 (first column: containing the patient's output information [survived or non-survived]; other columns: RBV features).

Of the 2000 COVID-19 patients in the first dataset, 894 (44.7%) and 1106 (55.3%) patients were female and male, respectively. The mean age values of female and male patients were 53.15 ± 17.06 and 56.42 ± 20.84 years, respectively. In the second dataset, 174 (62.4%) and 81 (37.6%) of 279 ICU COVID-19 patients were male (mean age: 77.51 ± 8.33 years) and female (mean age: 70.36 ± 10.92 years), respectively. Of the 1721 non-ICU COVID-19 patients, 826 (48.0%) and 895 (52.0%) patients were female (mean age: 51.36 ± 17.52 years) and male (mean age: 52.46 ± 19.36 years), respectively. The third dataset contains the RBV data of 2597 patients during treatment. Of these 2597 patients, 233 cases (9.0%) died, and 2364 cases (91.0%) survived. Of the patients who lost their lives, 143 (61.3%) and 90 (38.7%) cases were male and female, respectively. The mean age values of the survived and deceased patients were 55 ± 14.62 and 76 ± 12.13 years, respectively.

4. Discussion

The COVID-19_RBV1 dataset provides an excellent opportunity for methods to be used to diagnose COVID-19. The COVID-19_RBV2 dataset offers tremendous opportunities for methods to be used to determine the prognosis of COVID-19. The COVID-19_RBV3 dataset is an excellent resource for determining the mortality/severity of COVID-19.

The datasets present a tremendous opportunity for researchers who would like to work with AI and classifier ML models in the diagnosis/prognosis/mortality of COVID-19 because most of the RBV features used in studies conducted for this purpose in the literature are available in the present datasets. All the data are optimized, preprocessed, and ready to use. It is hoped that studies on these datasets will reduce the negative pressures on the health system and provide important clinical guidance for decision-makers in the diagnosis and prognosis of COVID-19.

Acknowledgments

The authors would like to express their gratitude to Erzincan Binali Yıldırım University Mengücek Gazi Training and Research Hospital, Erzincan, Turkey, for support and assistance in providing the data.

Footnotes

Authors' Contribution: Conceptualization: M. T. H.; Methodology: M. T. H.; Validation: M. T. H.; Formal analysis: M. T. H.; Investigation: M. T. H.; Resources: M. T. H.; Data curation: M. T. H.; Writing-original draft preparation: M. T. H.; Writing-review and editing: M. T. H.; Supervision: M. T. H.

Conflict of Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval: The study protocol was approved by the Institutional Ethics Review Board of Erzincan Binali Yıldırım University after being approved by the Ministry of Health of the Republic of Turkey in accordance with the Declaration of the Helsinki World Medical Association (ethics committee decision No.: 2021/02-07).

Funding/Support: This work did not receive any funding support.

References

1. Huyut MT, Soygüder S. The multi-relationship structure between some symptoms and features seen during the new coronavirus 19 infection and the levels of post-covid anxiety and depression. *Eastern J Medicine*. 2022;**27**(1):1–10. https://doi.org/10.5505/ejm.2022.35336.
2. Mertoglu C, Huyut MT, Arslan Y, Ceylan Y, Coban TA. How do routine laboratory tests change in coronavirus disease 2019? *Scand J Clin Lab Invest*. 2021;**81**(1):24–33. [PubMed ID: 33342313]. https://doi.org/10.1080/00365513.2020.1855470.

3. Huyut MT, Ilkbahar F. The effectiveness of blood routine parameters and some biomarkers as a potential diagnostic tool in the diagnosis and prognosis of Covid-19 disease. *Int Immunopharmacol*. 2021;**98**:107838. [PubMed ID: 34303274]. [PubMed Central ID: PMC8169318]. https://doi.org/10.1016/j.intimp.2021.107838.
4. Huyut MT, Ustundag H. Prediction of diagnosis and prognosis of COVID-19 disease by blood gas parameters using decision trees machine learning model: A retrospective observational study. *Med Gas Res*. 2022;**12**(2):60–6. [PubMed ID: 34677154]. [PubMed Central ID: PMC8562394]. https://doi.org/10.4103/2045-9912.326002.
5. Mertoglu C, Huyut MT, Olmez H, Tosun M, Kantarci M, Coban TA. COVID-19 is more dangerous for older people and its severity is increasing: A case-control study. *Med Gas Res*. 2022;**12**(2):51–4. [PubMed ID: 34677152]. [PubMed Central ID: PMC8562399]. https://doi.org/10.4103/2045-9912.325992.
6. Huyut MT, Huyut Z. Forecasting of Oxidant/Antioxidant levels of COVID-19 patients by using Expert models with biomarkers used in the Diagnosis/Prognosis of COVID-19. *Int Immunopharmacol*. 2021;**100**:108127. [PubMed ID: 34536746]. [PubMed Central ID: PMC8426260]. https://doi.org/10.1016/j.intimp.2021.108127.
7. Amgalan A, Othman M. Hemostatic laboratory derangements in COVID-19 with a focus on platelet count. *Platelets*. 2020;**31**(6):740–5. [PubMed ID: 32456506]. https://doi.org/10.1080/09537104.2020.1768523.
8. Kukar M, Guncar G, Vovko T, Podnar S, Cernelc P, Brvar M, et al. COVID-19 diagnosis by routine blood tests using machine learning. *Sci Rep*. 2021;**11**(1):10738. [PubMed ID: 34031483]. [PubMed Central ID: PMC8144373]. https://doi.org/10.1038/s41598-021-90265-9.
9. Huyut MT. Automatic detection of severely and mildly infected COVID-19 patients with supervised machine learning models. *Ing Rech Biomed*. 2023;**44**(1):100725. [PubMed ID: 35673548]. [PubMed Central ID: PMC9158375]. https://doi.org/10.1016/j.irbm.2022.05.006.
10. Yang HS, Hou Y, Vasovic LV, Steel PAD, Chadburn A, Racine-Brzostek SE, et al. Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning. *Clin Chem*. 2020;**66**(11):1396–404. [PubMed ID: 32821907]. [PubMed Central ID: PMC7499540]. https://doi.org/10.1093/clinchem/hvaa200.
11. Huyut MT, Velichko A. Diagnosis and prognosis of COVID-19 disease using routine blood values and lognnet neural network. *Sensors (Basel)*. 2022;**22**(13). [PubMed ID: 35808317]. [PubMed Central ID: PMC9269123]. https://doi.org/10.3390/s22134820.
12. Velichko A, Huyut MT, Belyaev M, Izotov Y, Korzun D. Machine learning sensors for diagnosis of COVID-19 disease using routine blood values for internet of things application. *Sensors (Basel)*. 2022;**22**(20). [PubMed ID: 36298235]. [PubMed Central ID: PMC9610709]. https://doi.org/10.3390/s22207886.
13. Jiang S, Huang Q, Xie W, Lv C, Quan X. The association between severe COVID-19 and low platelet count: evidence from 31 observational studies involving 7613 participants. *British J Haemato*. 2020;**190**(1). https://doi.org/10.1111/bjh.16817.
14. Huyut MT, Velichko A, Belyaev M. Detection of risk predictors of COVID-19 mortality with classifier machine learning models operated with routine laboratory biomarkers. *Applied Sciences*. 2022;**12**(23). https://doi.org/10.3390/app122312180.
15. Huyut MT, Huyut Z. Effect of ferritin, INR, and D-dimer immunological parameters levels as predictors of COVID-19 mortality: A strong prediction with the decision trees. *Heliyon*. 2023;**9**(3). e14015. [PubMed ID: 36919085]. [PubMed Central ID: PMC9985543]. https://doi.org/10.1016/j.heliyon.2023.e14015.
16. Tahir Huyut M, Huyut Z, Ilkbahar F, Mertoglu C. What is the impact and efficacy of routine immunological, biochemical and hematological biomarkers as predictors of COVID-19 mortality? *Int Immunopharmacol*. 2022;**105**:108542. [PubMed ID: 35063753]. [PubMed Central ID: PMC8761578]. https://doi.org/10.1016/j.intimp.2022.108542.
17. Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, et al. Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. *Int Immunopharmacol*. 2020;**86**:106705. [PubMed ID: 32652499]. [PubMed Central ID: PMC7296324]. https://doi.org/10.1016/j.intimp.2020.106705.
18. Huyut MT, Kocaturk İ. The effect of some symptoms and features during the infection period on the level of anxiety and depression of adults after recovery from COVID-19. *Current Psychiatry Res Rev*. 2022;**18**(2):151–63. https://doi.org/10.2174/2666082218666220325105504.
19. Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput Struct Biotechnol J*. 2020;**18**:784–90. [PubMed ID: 32280433]. [PubMed Central ID: PMC7118541]. https://doi.org/10.1016/j.csbj.2020.03.025.
20. Cabitza F, Campagner A, Ferrari D, Di Resta C, Ceriotti D, Sabetta E, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clin Chem Lab Med*. 2020;**59**(2):421–31. [PubMed ID: 33079698]. https://doi.org/10.1515/cclm-2020-1294.

Table 1. Routine Blood Parameter Properties Found in Datasets

Routine Blood Value Parameters	Unit of Measurement
Biochemical parameters	
Alanine aminotransaminase	U/L
Aspartate aminotransferase	U/L
Alkaline phosphatase	μmol/s.L
Albumin	μmol /L
Amylase	μmol/s.L
Creatine kinase myocardial band	U/L
Direct bilirubin	μmol/L
Glucose	mmol/L
Creatinine	μmol/L
Creatinine kinase	U/L
Lactate dehydrogenase	U/L
Low-density lipoprotein	mg/dL
Total bilirubin	μmol/L
Total protein	g/L
Potassium	mmol/L
Uric acid	mmol/L
Urea	mmol/L
Triglyceride	mmol/L
Sodium	mmol/L
Calcium	mg/dL
Chlorine	mmol/L
Cholesterol	mmol/L
Gamma-glutamyl transferase	μmol/s.L
High-density lipoprotein cholesterol	mmol/L
Estimated glomerular filtration rate	no
Hematological parameters	
Eosinophils count	10^9/L
Hematocrit	%
Hemoglobin	g/L
Lymphocytes count	10^9/L
Monocytes count	10^9/L
Basophil count	10^9/L
Neutrophils count	10^9/L
Red blood cells	10^{12}/L
Red cell distribution width	%
White blood cells	10^9/L
Platelet count	10^9/L
Mean corpuscular hemoglobin	pg

Mean corpuscular hemoglobin concentration	g/dL
Mean corpuscular volume	fL
Mean platelet volume	fL
Platelet distribution width	fL
Inflammatory, cardiac, and coagulation parameters	
D-dimer	μg/L
C-reactive protein	mg/L
Ferritin	μg/L
Fibrinogen	g/L
International normalized ratio	No
Prothrombin time	Sec
Procalcitonin	μg/L
Erythrocyte sedimentation rate	mm/hr
Troponin	ng/L
Activated partial prothrombin time	Sec