Integrated Citizen Centered Digital Health and Social Care
A. Värri et al. (Eds.)

doi:10.3233/SHTI200706

Reuse of Clinical COVID-19 Patient Data: Pre-Processing for Future Classification

Elena LAZAROVA[a,1], Sara MORA[a], Antonio DI BIAGIO[b],
Antonio VENA[b] and Mauro GIACOMINI[a]
[a] *Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Italy*
[b] *Division of Infectious Diseases, IRCCS Ospedale Policlinico San Martino, Genoa, Italy*

Abstract. One of the most important challenges in the scenario of COVID-19 is to design and develop decision support systems that can help medical staff to identify a cohort of patients that is more likely to have worse clinical evolution. To achieve this objective it is necessary to work on collected data, pre-process them in order to obtain a consistent dataset and then extract the most relevant features with advanced statistical methods like principal component analysis. As preliminary results of this research, very influential features that emerged are the presence of cardiac and liver illnesses and the levels of some inflammatory parameters at the moment of diagnosis.

Keywords. COVID-19, feature extraction, principal component analysis, imputation of data, pseudo-anonymous data

1. Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the novel coronavirus that provoked the pandemic of COVID-19. It was first reported in Hubei (China) in December 2019, but it quickly spread across the globe. In Italy, at July 15,2020, the total number of reported positive cases were 243.506 including 34.997 deaths. One of the most important challenges at the moment is to analyse what happened to investigate guidelines and appropriate instruments to face the disease in a more prepared way in the event that a new local pandemic episode occurs. In particular, it is important to work on previously collected data from COVID-19 patients in order to design and develop Decision Support Systems (DSS) that can help medical staff to identify a cohort of patients that is more likely to have worse clinical evolution. To achieve this objective, it is useful to define which clinical and laboratory parameters influence the outcome most, for example the death of a patient or admission to the Intensive Care Unit (ICU). In this way, advanced statistical procedures and Machine Learning (ML) techniques can be applied to identify, extract and analyse significant features for COVID-19 management. This paper first describes the applied pre-processing operations done on a sample of data collected from COVID-19 patients. The dataset was obtained through an already existing platform for the automatic collection of

[1] Corresponding Author, Elena Lazarova, Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Italy; elena.lazarova@dibris.unige.it

data from HIV infected patients, used for multicentric clinical studies since 2011 [1,2]. The second part reports the comparison between the considered advanced statistical algorithms to determine relevant features. Some preliminary results are reported and discussed.

2. Methods

2.1. Data collection

The complete list of parameters included in the protocol of data collection was approved by the Liguria ethics committee and for data storage and usage for research purposes, after a process of pseudo-anonymization. Patients also signed the informed consent. The system used to collect pseudo-anonymous clinical and laboratory data of COVID-19 patients was derived directly from the Liguria HIV Network. It is based on a Service Oriented Architecture (SOA) to ensure the interoperability between the hospital information system and the Liguria HIV Network [3] with the appropriate privacy and security level [4]. As the system was designed and developed to collect useful data in the HIV (but also HCV and TB) field, it was necessary to update the structure and to insert new parameters specific for COVID-19, for example the ones related to arterial blood gas test. In order to also collect the anamnestic information included in the approved protocol, specific sections were inserted into the platform so that medical staff could insert types of data that are not originally digital. Data included present and previous illnesses of the patient; home therapies; reported information about the symptoms and parameters measured at the moment of the hospitalization. Data on study participants, indicated as COVID-19 hospitalized patients, were collected from 22th February to 15th June, 2020. For each patient, data were collected starting from COVID-19 diagnosis until in-hospital death or discharge. The attributes used in the analysis reported in this manuscript are 71 in total, the input parameters included: gender, age, previous underlying diseases (for example those included in the Charlson Comorbidity Index) and treatments, laboratory findings at baseline (tests done 48hours prior to and following the first nasopharyngeal swab positive for SARS-CoV-2). The targets were death of the patient, ICU admission and discharge of the patient without any of the previous events. In the study a total of 912 patients were selected from a more numerous group (about 1200 patients), including criteria were: conclusion of the hospitalization so that information about patient death or discharge was present and definitive, complete insertion of almost all anamnestic information, including the treatments.

2.2. Advanced Statistical Procedure

The aim of this section is to briefly present the advanced statistical procedure that the authors considered appropriate to discriminate between features to single out the most influencing ones. The software used to pre-process and analyse data was MatLab (version 2018a). The authors proposed two different approaches with the common objective of reducing the high dimensionality of data before using classification methods.

Principal Component Analysis (PCA) is one of the most popular dimensionality reduction procedures. It computes the identification of a smaller number of uncorrelated variables from a larger dataset and its outputs are a transformed dataset with weights of individual instances and the weights of principal components. It is used in predictive models and exploratory data analysis [5,6].

2.3. Missing data management

Missing data in medical research is a common problem because, in general, real data contains several missing values. There are different types of "missingness" that can occur and this may influence how the researchers should analyse the data that they have collected.

Missing completely at random (MCAR): Patients with complete data cannot be distinguished from others with complete data. When data are MCAR, the missing values can be thought of as a random sub-sample of the actual values.

Missing at random (MAR): Patients with incomplete data differ from patients with complete data, but the pattern of "missingness" is traceable or predictable from other variables in the dataset, rather than being due to the specific variable on which the data are missing.

Not missing at random (NMAR): Missing values do depend on unobserved ones.

There are several methods for handling missing data [7,8].

Listwise deletion (or complete case analysis): If a row of the dataset has missing data for any of the parameters, then simply exclude that record from the analysis. It is the easiest way to deal with missing data and it requires minimal computing, but it probably excludes a great fraction of the entire dataset.

Imputation methods: Attempt to estimate the values of the missing data and 'fillin' or impute new values. Once this has been achieved the analysis can proceed as if the dataset were 'complete'.

During this research, the second option, imputation methods, was chosen to deal with missing data. Dataset parameters were divided into two groups: binary variables and continuous ones; while population was stratified by sex (M and F) and age (under 50 years, between 50 and 59, between 60 and 69, between 70 and 79, over 80 years) into 10 groups. Then a specific function was created to fill missing values for each patient for each specific parameter with the mode (binary variable) / mean (continuous variable) of the not null values of the group it belonged to.

3. Results

3.1. Study population

This section briefly presents the characteristics of the study's sample, it consisted of 912 patients, 546 males (60%) and 366 females (40%) with a combined mean age of 69 (SD = 16) years. The mean value of the Charlson Comorbidity Index adjusted by age is 4 (± 3) [9,10]. After a preliminary analysis we decided to only consider features that had a percentage of missing data less than 25%, so we excluded: albumin, absolute number of CD4 and CD8 T cells at baseline; systolic and diastolic blood pressure, FiO2 and PO2 at hospitalization.

3.2. PCA

The PCA was used to identify the most robust and representative features within the considered and previously mentioned group. In order to better underline how each variable contributes to the principal components, we decided to analyse the features normalized modules in the space identified by the first two principal components. Table 1 shows the first quartile of the features list calculated on the module of the first two components in the PCA space ordered by PCA weights.

Table 1. 1st features quartile division according to PCA weights.

Feature	Width	Feature	Width
Lactate Dehydrogenase	0,366	Alanine Transaminase	0,232
Azotemia	0,297	Prothrombin time PT%	0,230
Charlson Comorbidity Index (CCI)	0,296	Peripheral vascular disease	0,223
Age	0,294	Cerebrovascular disease	0,213
Aspartate Aminotransferase	0,293	Chronic antiplatelet home therapy	0,208
Ferritin	0,274	Cerebrovascular pathology	0,207
Heart failure	0,242	Coronary pathology	0,204
Congestive heart failure	0,242	Total bilirubin	0,200
Troponin I	0,241	SO_2 ART / peripheral SO_2	0,198

4. Discussion

The preliminary results of this research basically show that influential features in COVID-19 patients include: Charlson Comorbidity Index; increased troponin levels and the levels of some inflammatory parameters at the moment of diagnosis.

Regarding CCI, our study supports recent findings showing that previous history of cardiovascular disease, cerebrovascular disease, liver disease or acute kidney injury (e.g. all features included in the CCI) are the most important determinants for developing severe COVID-19 [11]. Because of these factors are even more important for outcome than the virulence of Sars-CoV-2 strains [12], we believe that they should always be considered as determinant features for decision support systems. As for the prognostic value of troponin, it is important to mention that COVID-19 remains associated with high risk for developing cardiovascular complication [13], so that it is essential to identify high-risk patients who may benefit from early aggressive treatment strategies. Previous studies have been focused on the prognostic impact of troponin levels in patients with COVID-19 [14]. Troponin elevation has been found to be associated with an increased risk of myocardial injury and death for COVID-19 patients. Lastly, confirming the association between some inflammatory parameters (e.g. ferritin) and disease severity [15], our study supports the involvement of a cytokine storm in the clinical outcome of the patients [16]. The implication of the host immune response in the disease process among COVID-19 patients suggests a potential role of antiinflammatory drugs as adjunctive therapy. Therefore, our preliminary analysis can be also used to define a subset of parameters to be rapidly considered to enhance the safety in terms of treatment for COVID-19 patients. However, follow-up studies evaluating the role of anti-inflammatory drugs in well-defined sub-groups are warranted.

5. Conclusion

This manuscript's aim is to present preliminary results of the analysis conducted on a dataset related to COVID-19 patients, that are quite aligned with current medical knowledge. We believe that a pre-processing of this type is adequate for the correct preparation of further and more accurate classification models based on machine learning to help medical staff in the therapeutic decisions related to the infection. Moreover, we can assess that a moderate level of missing data, if correctly addressed in the pre-processing phase, cannot prevent a correct classification in situations like the presented one.

References

[1] Fraccaro P, Pupella V, Gazzarata R, et al. The ligurian human immunodeficiency virus clinical network: a web tool to manage patients with human immunodeficiency virus in primary care and multicenter clinical trials. Med 2 0. 2013 Aug;2(2):e5.

[2] Giannini B, Riccardi N, Cenderello G, Di Biagio A, Dentone C, Giacomini M. From Liguria HIV Web to Liguria Infectious Diseases Network: How a Digital Platform Improved Doctors' Work and Patients' Care. AIDS Res Hum Retroviruses. 2018 Mar;34(3):239-240.

[3] Gazzarata R, Giannini B, Giacomini M. A SOA-Based Platform to Support Clinical Data Sharing. J Healthc Eng. 2017 May;Article ID:2190679.

[4] Gazzarata G, Gazzarata R, Giacomini M. A standardized SOA based solution to guarantee the secure access to EHR. Procedia Computer Science. 2015;64:1124-1129.

[5] Lee SA. Coronavirus Anxiety Scale: A brief mental health screener for COVID-19 related anxiety. Death Stud. 2020 Apr;44(7):393-401.

[6] Adiwijaya WU, Lisnawati E, Aditsania A, Kusumo DS. Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification. Journal of Computer Science 2018;14(11):1521-1530.

[7] Bennett DA. How can I deal with missing data in my study?. Australian and New Zealand journal of public health. 2001;25(5):464-469.

[8] Soley-Bori M. Dealing with missing data: Key assumptions and methods for applied analysis. Boston University. 2013;23.

[9] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373-383.

[10] Glasheen WP, Cordier T, Gumpina R, Haugh G, Davis J, Renda A. Charlson Comorbidity Index: ICD-9 Update and ICD-10 Translation. Am Health Drug Benefits. 2019;12(4):188-197.

[11] Chen N, Zhou M, Dong X, Qu J, Gong F, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet. 2020;395(10223):507-513.

[12] Xiaobo Y, Yuan Y, Jiqian X, Huaqing S, Jia'an X, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. The Lancet Respiratory Medicine. 2020;8(5):475-481.

[13] Bansal M. Cardiovascular disease and COVID-19. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 2020;14(3):247-250.

[14] Imazio M, Klingel K, Kindermann I, et al. COVID-19 pandemic and troponin: indirect myocardial injury, myocardial inflammation or myocarditis?. Heart. 2020;106:1127-1131.

[15] Wu C, Chen X, Cai Y, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. JAMA Intern Med. 2020;180(7):934–943.

[16] Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. Intensive care medicine. 2020;46(5):846-848.