Clinical risk factors of ICU & fatal COVID-19 cases in Brazil

Juliana B. Mattos[1], Eraylson G. Silva[1], Paulo S.G. de Mattos Neto[1], Renato Vimieiro[2]

[1] Centro de Informática, Universidade Federal de Pernambuco, Recife-PE, Brasil
{jbm4,egs,psgmn}@cin.ufpe.br
[2] Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil
rvimieiro@dcc.ufmg.br

Abstract. The Coronavirus disease 2019 (COVID-19) was first detected in China in December 2019. In a few months, the disease got pandemic proportions, overloading health systems all around the world. Risk factors related to the progression and outcome of the disease are still unclear. Moreover, clinical aspects of patients can differ between societies, and other demographic elements may impact survival responses. A better characterisation of local manifestation of COVID-19 is crucial to a better general understanding of the disease, and thus to improve treatment decisions and health systems' management. In this article, we performed an initial analysis of clinical factors related to admission in ICU or death of SARS-CoV-2 confirmed Brazilian patients, based on 1,138,690 medical records from the Brazilian government. To our knowledge, this study is the first to assess clinical risk factors for disease progression in Brazil. We provide a concise data set of medical registers related to COVID-19 in the whole Brazilian territory, and we describe the baseline comorbidities and symptoms observed in the data collection. Then, we assess the correlation between the manifestation of symptoms/comorbidities and the patients' survival response through Kaplan-Meier survival estimates. The results here reported are mostly in accordance with findings reported in previous works.

CCS Concepts: • **Information systems** → **Data mining**; • **Mathematics of computing** → **Survival analysis**; • **Applied computing** → **Consumer health**.

Keywords: Clinical Analysis, COVID-19, Risk factors, SARS-CoV-2, Survival Analysis

1. INTRODUCTION

The Coronavirus disease 2019 (COVID-19), caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [He et al. 2020], was first detected in late December 2019, in the city of Wuhan, China. Since then, it has spread quickly nationwide, taking pandemic proportions. By July 2020, it has affected more than 200 countries, with almost 13M confirmed cases and approximately 570k confirmed deaths[1]. The high spread rate of SARS-CoV-2 has overloaded health systems all around the world, and governments have struggled to manage the disease progression. In Brazil, the first confirmed cases were reported by the end of February 2020. In less than five months, it has spread all over the country, registering over 1.8M confirmed cases and 70k deaths[2].

The infection caused by SARS-CoV-2 affects mainly the respiratory capacity of patients, and present a higher mortality rate in older ages. Clinical studies [Song et al. 2020; Huang et al. 2020; Richardson et al. 2020] conducted in China and USA report the main symptoms and comorbidities observed in COVID-19 patients. Besides, the studies also indicate that the chance of mortality increases in the presence of coexisting medical conditions. Although some studies [Zhou et al. 2020; Zheng et al. 2020; Schnake-Mahl et al. 2020] strive to describe risk factors for the progression of COVID-19, the risk characterisation for distin-

[1]World Health Organization (WHO) website for monitoring COVID-19 progression: https://covid19.who.int/
[2]Brazilian Government official website for monitoring COVID-19 progression: https://covid.saude.gov.br/

guishing patients with higher chances of complications is still unclear. A clearer understanding of elements that reflect better or worse on patients' survival can directly support treatment decisions and improve management of the health system. In addition, [Dowd et al. 2020] highlight the role of demography for understanding differences in cross-country fatality and the impact of the pandemic on different populations. In this sense, [Nepomuceno et al. 2020] highlight that different distributions of chronic diseases – and other demographic aspects – across diverse populations may reflect differently on fatality risk. Therefore, besides the importance of a broad comprehension of clinical characteristics related to SARS-CoV-2 progression, it is also essential to tailor such analysis to local aspects of the disease.

In this article, we aim at identifying risk factors that interfere in the intensive care unit (UCI) use or death of COVID-19 patients in Brazil. We retrieved data on 1,138,690 Brazilian SARS-CoV-2 confirmed patients until July 14th, 2020, from Brazilian government open database. We conducted extensive data processing to generate a concise data set with 47 boolean clinical features related to the most observed symptoms and comorbidities. Then, we assessed the survival influence of baseline comorbidities based on the Kaplan-Meier survival curves. We compare the results with the ones reported by [Richardson et al. 2020] in an analysis of US patients. We also provide characteristics of most observed symptoms. The remainder of this document is organised as follows. Section 2 provides a brief review on Survival Analysis. Section 3 describes the data collection used for the analysis and the processing method for generating the final data set. In Section 4, we present the experimental setting and achieved results. Finally, in Section 5, we draw some conclusions and present some directions to extend this study.

2. SURVIVAL ANALYSIS

We define Survival Analysis [Kleinbaum 1998] as the study of the *time* for an *event* occurrence. The *event* can be any experience of interest to the study, e.g. the use of ICU or death of COVID-19 patients. The *survival time* is the period between a defined date for the beginning of the study observation and a date for event occurrence (or until end of study observation), e.g. from the onset of illness until hospitalisation or death. We say a subject survived if it has not experienced the event of interest in the Survival Analysis. Therefore, the term *survival* can be understood as time to event, and, from now on, we will refer to it and its variants with relation to the survival event. We call *censoring* the existence of subjects for which the event date was not observed and, therefore, is unknown. *Right-censoring* is when the event date is past the time observed during the study. This censoring usually occurs because a subject was lost to observation or because the study ended, both before event occurrence. *Left-censoring* (and *interval-censoring*) happens when a subject suffers the event during the study period, but the exact date is unknown (not observed). The survival data is composed by survival times T and the *survival status* δ that indicates whether a subject is censored ($\delta = 0$) or has suffered the event ($\delta = 1$).

The *survival function* $S(t) = P(T_i > t)$ is commonly used for modelling survival data, and it indicates the probability that a subject under study survives (i.e. do not suffers the event) up to a specified future time $t \in \tau$, being $\tau = \{t_1, \ldots, t_j, \ldots, t_k | t \in T\}$ the set of unique ordered survival times. It can be estimated by the non-parametric Kaplan-Meier (KM) survival estimate [Kaplan and Meier 1958], given by Equation 1, where $\hat{S}(t_{j-1})$ is the probability of surviving up to t_{j-1}, r_j is the number of subjects that survived up to just before t_j, and d_j the number of events that happened at the time interval $[t_j, t_{j+1})$; for $t_0 = 0$, $S(t_0) = 1$.

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \left(1 - \frac{d_j}{r_j}\right) \tag{1}$$

We call *survival curve* the plot of the KM estimates against time. The *median survival time* statistic refers to survivability, and it indicates the time when half of the subjects under study are expected to be alive. It is defined as the time $t_m \in \tau$ for which the probability of surviving is equal to 50%, i.e. the time $t_m \mid \hat{S}(t_m) = 0.5$. It is essential to notice that the median survival time is a measure of centrality related to survivorship and not to the survival times τ, and it indicates the time for which the probability of surviving is central (50%) rather than the central time in τ.

The *logrank* [Peto et al. 1977] statistical test is widely used for comparing two survival functions. It uses the number of observed events O and the number of events E that are expected to happen (Equation 2) to test the null hypothesis that there is no overall difference between survival functions. The test follows the Chi-squared distribution, which is used for the P-value estimation. The logrank $X^2 \sim \chi_1^2$ for comparing the survival estimates of two groups – G_1 and G_2 – is given in Equation 3, where $O^{G_1}(O^{G_2})$ is the number of observed events in $G_1(G_2)$, and $E^{G_1}(E^{G_2})$ the number of expected events in $G_1(G_2)$.

$$E^G = \sum_{\forall t_j} \frac{r_j^G}{r_j} \times d_j \tag{2}$$

$$X^2 = \frac{(O^{G_1} - E^{G_1})^2}{E^{G_1}} + \frac{(O^{G_2} - E^{G_2})^2}{E^{G_2}} \tag{3}$$

3. DATA COLLECTION

The data collection used in this article refers to the Brazilian epidemiological database of the severe acute respiratory syndrome (Sídrome Respiratória Aguda Grave - SRAG). Such data is public and available on the government website[3], that also provides information on data origin and collection procedure. The data is collected through manual charts and manually inserted on an electronic system, presenting a large number of errors and inconsistencies. Alternatively, we retrieved the SRAG data from the *Observatório COVID-19 BR*[4], which is an independent initiative from Brazilian researches to provide updated SRAG data through a web crawler that browsers data releases from the Brazilian government. The data provided has initial processing; however, there is still a large number of inconsistencies and formatting problems. In this article, we used a sample of SRAG data, through the *SRAGHospitalizado* files available on *Observatório COVID-19 BR* repository[5]. The sample of files we used refer to the dates 2020-05-04, 2020-05-18, 2020-06-01, 2020-06-09, 2020-06-16, 2020-06-17, 2020-06-23, 2020-06-30, 2020-07-07 and 2020-07-14. Such files sum 3.5GB of data. The files referring to the following dates were not applicable due to poor formatting: 2020-06-02, 2020-06-03, 2020-06-04, 2020-06-05, 2020-06-06, 2020-06-07 and 2020-06-08. Those data files are not daily, but instead, they are accumulative over a period.

The SRAG database refers to all cases of the severe acute respiratory syndrome. Therefore, we filtered only SARS-CoV-2 confirmed cases through a logical disjunction of the following circumstances: (1) the final classification of the medical register reported as *covid19*; (2) *positive* RT-PCR test result for SARS-CoV-2; and (3) *negative* result of non-molecular biology test methods for influenza or other viruses. The selection resulted in a total of 1,138,690 patients' registers.

The original SRAG database has 21 boolean fields related to symptoms and comorbidities. Also, the database has two text fields that describe other symptoms and comorbidities not encompassed in the boolean fields. We conducted regular expression search on the textual fields to generate clinical features related to the most frequent patterns in the text. The generated features assumed True value in the occurrence of its patterns, and value equal False in the not appearance. Registers with no text information were represented as missing values for all features under generation. Then, we performed logical disjunction between the original boolean clinical fields and similar generated features in order to unify all clinical information.

Finally, we adjusted typing errors related to date fields. Date registers with year superior to 2020 were altered to 2020, and records with value not related to dates were mapped to missing values. The considered date fields are described as follows: (a) date of case notification; (b) the patient-reported time related to the onset of illness; (c) date of hospitalisation; (d) date of admission in the intensive care unit (ICU); (e) date of

[3] Brazilian government website for SRAG surveillance: `https://opendatasus.saude.gov.br/dataset/bd-srag-2020`

[4] *Observatório COVID-19 BR* website: `https://covid19br.github.io/`

[5] *Observatório COVID-19 BR* SRAG repository: `https://github.com/covid19br/central\_covid/tree/master/dados/SIVEP-Gripe`

death or end of hospitalisation; and (f) date of register closure. The code for retrieving and processing the SRAG database, as well as all additional information for feature generation, and also the final data set are available on this article's website[6].

4. RESULTS AND DISCUSSION

4.1 Experimental setup of survival data set

In this article, we analyse clinical factors related to critical patient prognosis. Therefore, we define the survival *event* as death or admission in ICU. The feature related to the *survival status* δ assumes True value for cases that experienced at least one of the two instances encompassed in the event definition, and False otherwise (*censoring*). Such a feature was defined based on the logical disjunction of the following two information: (1) admission to ICU as positive; and (2) patient evolution reported as death.

For the *survival time*, we are interested in the period from onset of illness to death or admission in ICU. Due to a large number of missing values and inconsistencies on the dates' registration, the initial and final study dates were defined as follows. For the initial date, we considered both the time of case notification and the date reported as the onset of illness. We chose the smaller one since it belongs to the period between December 1st, 2019, and July 14th, 2020 (date of SRAG database last considered updating). Analogously, for the end of survival study, we considered the smallest between the following three dates, since it belongs to the period between the initial time and July 14th, 2020: (i) date of admission to ICU; (ii) date of death/end of hospitalisation; and (iii) date of register closure.

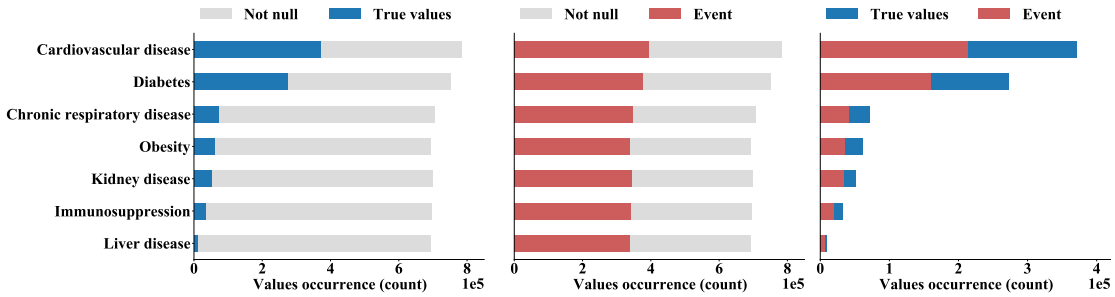
Before generating the *survival time* as the time in days between the two dates mentioned above, we handled missing values for the final date in two steps: (1) for all censored cases, we inputted the date of last database updating: 2020-07-14; and (2) for event-true cases, we used the date of hospitalisation (when existent and following the aforementioned final date requirements) and set those cases to censoring, i.e. set *survival status* to False. Then, we removed the event-true cases with no final date and no date for hospitalisation from the final survival data set. This because input a final date posterior to the (unknown) event occurrence date would imply *left-censoring*, which is not encompassed by the Survival Analysis methods applied in this article. The final data set is composed of 1,138,475 records related to COVID-19 confirmed patients in Brazil.

4.2 Results Analysis

In this article, we assess the survival influence of seven baseline comorbidities analysed by [Richardson et al. 2020] on COVID-19 patients in US: cardiovascular disease including hypertension (370878, 32.6%), chronic respiratory disease (71308, 6.3%), immunosuppression (33166, 2.9%), kidney disease (51210, 4.5%), liver disease (9919, 0.9%), obesity (61289, 5.4%) and diabetes (273031, 24.0%). Figure 1 shows the distribution of those features on the data set. The event incidence among those features is in average 49.4%. However, when analysing patients presenting such comorbidities, the incidence of death and admission to ICU increases to an average of 59.5%.

Figure 2 presents the survival curves for each comorbidity individually. It is possible to observe that patients presenting those medical conditions also present a faster decrease in the probability of surviving (not suffering the event). Such a faster decrease in survivability can also be observed based on the median survival time, i.e. the time for which the probability of surviving is central (50%). When compared to the population's median survival time of 20 days, patients presenting kidney disease are the worst case, with a median survival of 11 days. They are followed by the ones presenting liver disease and obesity (median of 12 days), and chronic respiratory disease, diabetes, cardiovascular disease and immunosuppression (median of 14 days). In general, for patients presenting any analysed comorbidity, the survival half-life is a week

[6]This article's website: `https://github.com/jbmattos/KDMiLe2020_Risk-Factors-COVID19-Brazil`



(a) Feature occurrence distribution

(b) Event distribution by feature

(c) Event distribution by feature occurrence

Fig. 1: Comorbidities distribution on data set: (a) comorbidity occurrence among non-null information; (b) event occurrence among non-null information; and (c) event occurrence among comorbidity occurrence

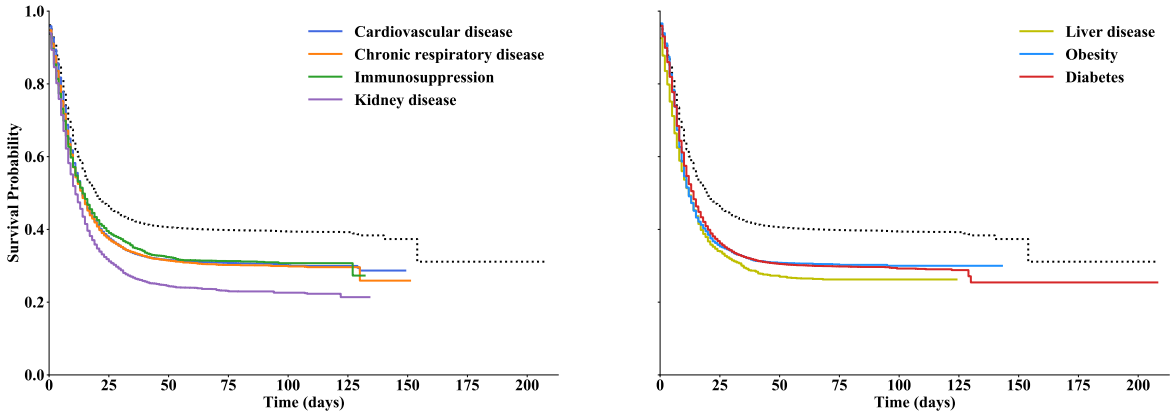


Fig. 2: Survival curves of analysed baseline comorbidities; the dotted line represents the survival curve based on all data set cases

Table I: P-values of Logrank statistical test performed on the comorbidities survival functions, by pairs. "Population" represents the survival function based all patients of the data set. Bold values represent the cases that accepted the null hypothesis for a significance level of 0.05

	Cardiovascular disease	Chronic respiratory disease	Immunosuppression	Kidney disease	Liver disease	Obesity	Diabetes
Population	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cardiovascular disease	-	5.78e-08	1.69e-02	0.00	0.00	0.00	1.68e-13
Chronic respiratory disease	-	-	**2.60e-01**	0.00	0.00	1.55e-04	**3.37e-01**
Immunosuppression	-	-	-	0.00	0.00	1.66e-04	**4.37e-01**
Kidney disease	-	-	-	-	8.28e-04	0.00	0.00
Liver disease	-	-	-	-	-	1.35e-13	0.00
Obesity	-	-	-	-	-	-	1.71e-09

shorter. In addition to a steeper decay in survivability, patients presenting such medical conditions also present a worse prognosis (i.e. lower survival probability) when compared to all patients in the data set (population). Table I present the P-values of the logrank test comparing the survival functions of population and each comorbidity group of patients, by pairs. For a level of significance of 5%, the survival response of patients presenting chronic respiratory disease, or immunosuppression, or diabetes can be considered similar. For all other combinations, we refuse the null hypothesis that there is no overall difference between the survival functions. This rejection also implies that the event distribution is different across the groups, including the comparison with the population.

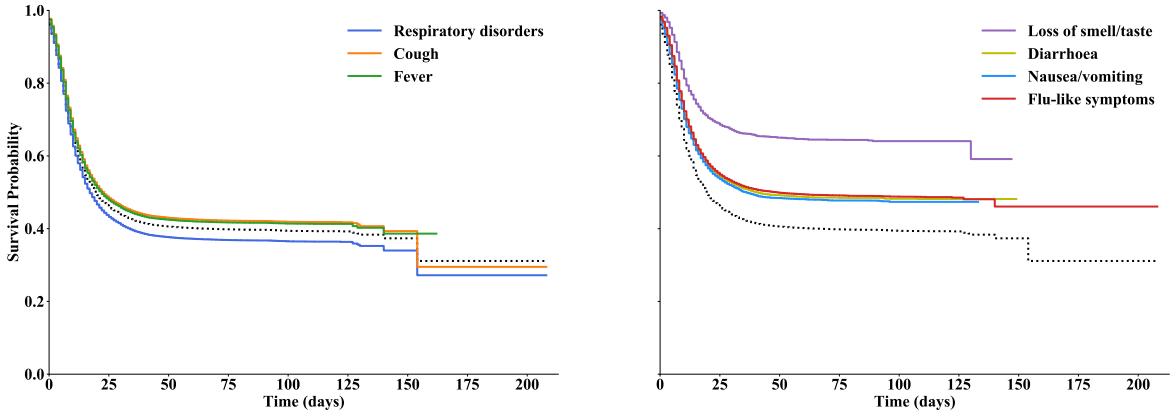


Fig. 3: Survival curves of analysed baseline symptoms; the dotted line represents the survival curve based on all data set cases

Additionally, an initial analysis of observed symptoms revelled the following most frequent clinical manifestations: respiratory disorders including shortness of breath, dyspnea, oxygen saturation lower than 95% and cyanosis (949005, 83.4%); cough (851002, 74.7%); and fever (809483, 71.1%). Less frequent registered symptoms include flu-like symptoms (438690, 38.5%) as headache, sore throat, fatigue, myalgia/arthralgia, nasal congestion, runny nose and sneeze; diarrhoea (153017, 13.4%); nausea/vomiting (97053, 8.5%); and loss of smell/taste (52061, 4.6%). Figure 3 presents the survival curves of the patient groups manifesting each of those symptoms. We refuted the logrank null hypothesis for all groups, and thus all survival curves are statistically different. It is possible to observe that although respiratory disorders, cough and fever are persistent symptoms and strong markers for COVID-19, they are not good markers for patients' prognosis when compared to the population. In contrast, the presence of loss of smell/taste, diarrhoea, nausea/vomiting and flu-like symptoms are associated with a better prognosis, i.e. patients presenting such symptoms also present higher survivability.

4.3 Discussion and Limitations

In this article, we analysed the association between seven baseline comorbidities and survival of COVID-19 patients: cardiovascular disease, chronic respiratory disease, immunosuppression, kidney disease, liver disease, obesity and diabetes. All comorbidities presented are related to an increased chance of admission to ICU or death. The relation between such medical conditions and the worse prognosis is also reported in other studies. [Zheng et al. 2020] report a higher proportion of diabetes, cardiovascular disease, respiratory disease and hypertension among patient who developed critical illness or died, when comparing to non-critical patients. In comparison to the comorbidities distribution presented by the US patients analysed by [Richardson et al. 2020], Brazilian patients exhibited a significantly lower incidence of the described medical conditions. Such difference, however, may occur because the data is not standardised, and the clinical features were extracted through search in manually inputted texts. Such procedure results in significant potential loss of information, and the final sample of patients may not be representative of the population.

When analysing baseline symptoms of COVID-19 Brazilian patients, the most observed were respiratory disorders, cough, fever, loss of smell/taste, diarrhoea, nausea/vomiting and flu-like symptoms. Similar findings were reported for patients in the US and China [Zhou et al. 2020; Richardson et al. 2020]. [Zheng et al. 2020] observed that the proportion of fever, headache and myalgia/arthralgia was lower in critical/mortal patients when comparing to non-critical patients. Our results also show an association between those symptoms and groups of patients with better prognosis. The authors also report a higher proportion (without statistical significance) of cough, sputum production, fatigue, diarrhoea and nausea/vomiting in critical/mortal patients. In contrast, our results show those symptoms strongly related to groups of patients less likely to admission in ICU and death. Lastly, the authors also report that the proportion of shortness

of breath/dyspnea was statistically significant higher in critical/mortal patients. This reporting is in accordance with our findings that show that the group of patients presenting respiratory disorders also present worse survival when comparing to the population. However, such results' differences may be due to the clinical data collection procedure, which is highly manual and results in a considerable loss of information.

Finally, the collection of standardised data poses a significant limitation to the analysis of clinical risk factors of COVID-19 in Brazil. The retrieved clinical information of the reported symptoms and comorbidities refers to an average of 26.5% of the patients in the processed data set. Anyhow, the results reported in this article are an initial analysis. Besides individual symptoms and comorbidities, the association between them – and associations between clinical factors and other demographic aspects – may interfere in disease progression and outcome. Further investigation is, therefore, crucial to better characterisation of risk factors of COVID-19.

5. CONCLUSIONS

In this article, we performed an initial analysis of clinical factors related to admission in ICU or death of SARS-CoV-2 confirmed patients, based on a concise database generated from the Brazilian epidemiological open-data of the severe acute respiratory syndrome. The findings regarding the comorbidities and symptoms observed on the data are in accordance with the disease's main clinical markers already reported on the literature and other information vehicles. When assessing the influence of such clinical aspects on the disease's prognosis, our findings suggest that Brazil's patients may present different comorbidities distribution and experience symptoms differently from results reported on other countries' patients. However, there is a need for deeper analysis to assure the representativity of the analysed patients' sample and to better delineate the methodological drawing and limitations of this study. Besides, further investigation is necessary to assess the combined influence of risk factors. Additionally, a control group is needed to assure the validity of conclusions regarding the clinical influences over COVID-19 survival response. Therefore, the next steps to expand this study are as follows. The first step is to expand the initial exploratory data analysis here presented with more precise and detailed methodological drawing, and more in-depth results analysis when comparing to the recent literature. The second step is to expand the initially processed dataset of symptoms and comorbidities with further medical and demographic information, to discover combined clinical and demographic factors of COVID-19 associated with disease progression and outcome. The additional medical data includes administrated medicines, results from chest x-rays, use of mechanical ventilators, and viral respiratory co-infection. The demographic aspects may consist of age, ethnicity, municipality, and other census information available by the Brazilian Institute of Geography and Statistics (IBGE) – such as social indicators and statistics; municipality profile and economy; family income; education and professional qualification; and health system profile and economy. And, finally, to perform risk analysis based on Exceptional Model Mining to discover combined factors related to subgroups of COVID-19 patients that present distinct – exceptional – survival response.

ACKNOWLEDGMENT

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and by the National Council for Scientific and Technological Development – CNPq

REFERENCES

DOWD, J. B., ANDRIANO, L., BRAZEL, D. M., ROTONDI, V., BLOCK, P., DING, X., LIU, Y., AND MILLS, M. C. Demographic science aids in understanding the spread and fatality rates of covid-19. *Proceedings of the National Academy of Sciences* 117 (18): 9696–9698, 2020.

HE, F., DENG, Y., AND LI, W. Coronavirus disease 2019: What we know? *Journal of medical virology* 92 (7): 719–725, 2020.

HUANG, C., WANG, Y., LI, X., REN, L., ZHAO, J., HU, Y., ZHANG, L., FAN, G., XU, J., GU, X., ET AL. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet* 395 (10223): 497–506, 2020.

KAPLAN, E. L. AND MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53 (282): 457–481, 1958.

KLEINBAUM, D. G. Survival analysis, a self-learning text. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 40 (1): 107–108, 1998.

NEPOMUCENO, M. R., ACOSTA, E., ALBUREZ-GUTIERREZ, D., ABURTO, J. M., GAGNON, A., AND TURRA, C. M. Besides population age structure, health and other demographic factors can contribute to understanding the covid-19 burden. *Proceedings of the National Academy of Sciences* 117 (25): 13881–13883, 2020.

PETO, R., PIKE, M., ARMITAGE, P., BRESLOW, N. E., COX, D., HOWARD, S., MANTEL, N., MCPHERSON, K., PETO, J., AND SMITH, P. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. *British Journal of Cancer* 35 (1): 1, 1977.

RICHARDSON, S., HIRSCH, J. S., NARASIMHAN, M., CRAWFORD, J. M., MCGINN, T., DAVIDSON, K. W., BARNABY, D. P., BECKER, L. B., CHELICO, J. D., COHEN, S. L., ET AL. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area. *Jama*, 2020.

SCHNAKE-MAHL, A. S., CARTY, M. G., SIERRA, G., AND AJAYI, T. Identifying patients with increased risk of severe covid-19 complications: Building an actionable rules-based model for care teams. *NEJM Catalyst Innovations in Care Delivery* 1 (3), 2020.

SONG, F., SHI, N., SHAN, F., ZHANG, Z., SHEN, J., LU, H., LING, Y., JIANG, Y., AND SHI, Y. Emerging 2019 novel coronavirus (2019-ncov) pneumonia. *Radiology* 295 (1): 210–217, 2020.

ZHENG, Z., PENG, F., XU, B., ZHAO, J., LIU, H., PENG, J., LI, Q., JIANG, C., ZHOU, Y., LIU, S., ET AL. Risk factors of critical & mortal covid-19 cases: A systematic literature review and meta-analysis. *Journal of Infection*, 2020.

ZHOU, F., YU, T., DU, R., FAN, G., LIU, Y., LIU, Z., XIANG, J., WANG, Y., SONG, B., GU, X., ET AL. Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The lancet*, 2020.