Journal Pre-proof

Machine Learning Approaches in COVID-19 Survival Analysis and Discharge Time Likelihood Prediction using Clinical Data

Mohammadreza Nemati, Jamal Ansary, Nazafarin Nemati

PII: S2666-3899(20)30094-5

DOI: https://doi.org/10.1016/j.patter.2020.100074

Reference: PATTER 100074

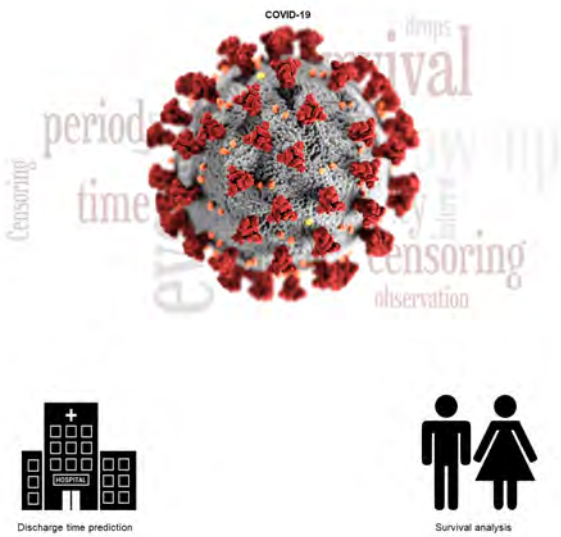
To appear in: *Patterns*

Received Date: 12 April 2020

Revised Date: 23 June 2020

Accepted Date: 1 July 2020

Please cite this article as: Nemati M, Ansary J, Nemati N, Machine Learning Approaches in COVID-19 Survival Analysis and Discharge Time Likelihood Prediction using Clinical Data, *Patterns* (2020), doi: https://doi.org/10.1016/j.patter.2020.100074.



Title

Machine Learning Approaches in COVID-19 Survival Analysis and Discharge Time Likelihood Prediction using Clinical Data

Authors and Affiliation

Mohammadreza Nemati[1, 4, *], Jamal Ansary[2], Nazafarin Nemati[3]

[1]University of Toledo, Electrical Engineering and Computer Science, Toledo, Ohio, US

[2]University of Toledo, Mechanical, Industrial and Manufacturing Engineering, Toledo, Ohio, US

[3]Foothill Collage, Biological and Health Sciences, Los Altos, California, US

[4]Lead Contact: mnemati@rockets.utoledo.edu

[*]Correspondence

Keywords

COVID-19, Survival Analysis, Machine Learning, Statistical Analysis

Summary

As a highly contagious respiratory disease, COVID-19 has yielded high mortality rates since its emergence in December of 2019. As the number of COVID-19 cases soars in epicenters, health officials are warning about the possibility of the designated treatment centers being overwhelmed by coronavirus patients. In this study, several computational techniques are implemented to analyze the survival characteristics of 1182 patients. The computational results agree with the outcome reported in early clinical reports released for a group of patients from China that confirmed a higher mortality rate in men compared to women and in older age groups. The discharge time prediction of COVID-19 patients was also evaluated using different machine learning and statistical analysis methods. The results indicate that the Gradient Boosting survival model outperforms other models for patient survival prediction in this study. This research study is aimed to help health officials make more educated decisions during the outbreak.

Introduction

In December of 2019, a soaring number of unusual pneumonia cases was reported in Wuhan, China. The cause of this outbreak was soon determined to be a novel coronavirus, referred to as COVID-19[1]. On March 11, 2020, the World Health Organization (WHO) recognized COVID-19 as a global pandemic with significantly high infection and mortality rates compared to its predecessors, including SARS and MERS[2]. As of March 24, 2020, the virus has spread to more than 170 countries, with more than 422,613 confirmed cases and 18,891 death toll[3]. The initial reports indicated that the mortality rate varies among the states due to differences in demography, age distribution, and health infrastructure. China reported an overall 2.3% mortality rate among COVID-19 patients. However, a significantly higher mortality rate (14.8%) was reported for senior patients (80 years or older)[4]. In Italy, where more than 23% of residents are 65 or older[5] , the overall mortality rate has been about 5%, while the statistics showed a rate of

around 20% for senior patients[6]. Across the world, epicenters of the coronavirus outbreak are beginning to confront rapid surges in confirmed cases that may overwhelm health care hospitals and medical personnel. Precise mathematical models capable of predicting the duration of recovery and discharge time can provide valuable information for health officials to design proper strategies to reduce the death toll. It has been shown by early studies that statistical analysis can be applied to COVID-19 problems to build predictive models that can assess risk factors and mortality rates[7-9]. In this paper, we will use survival analysis techniques including statistical analysis and machine learning approaches to predict patient survival times and to examine the effect of basic risk factors on hospital discharge time probabilities. What distinguishes survival analysis from the typical machine learning algorithms is that some parts of the training data may be partially observed – censored samples. There are numerous cases in this study that the date of event of interest, patient discharge time, is not available. Instead of employing typical predictive models that cannot make use of these cases, we will utilize well-suited methods capable of carrying out analysis on censored cases which yields more reliable outcomes by preventing massive data shrinkage. These methods are introduced in the experimental procedure section.

Results and Discussion

Discharge Time Prediction Accuracy

In this section, we report and compare the results of techniques used in discharge time prediction. Taking the performance metric into consideration, due to the existence of censored samples, the typical area under receiving operating characteristic curve (AUC) is not used to evaluate the performance of survival analysis models. Instead, model performances in discharge time prediction are compared by a metric namely Concordance index (C-index). C-index is a standard metric to assess the predictions of algorithms in survival analysis by calculating the percentage of concordant pairs among all feasible evaluation pairs[10]. C-index does not consider the difference value between predicted and actual survival times, but it compares only the ranking times of events of interest in all possible pairs. For example, if patient A's actual event happens before patient B, and then the predicted event time for A is before B, no matter how long before B as long as it happens prior to that, this pair is considered as a concordant pair.

According to the results shown in Table 1, the IPCRidge has the least accuracy among the six algorithms evaluated in this work. This algorithm is expected to perform randomly if the assumptions of this algorithm are violated. Since in this study, the distribution of the survival data is not known and the censoring status is not independent of the features, the IPCRidge performs randomly. CoxPH and Coxnet models show similar results as the data does not contain numerous features even after transforming categorical features to numerical. The Coxnet model excels when dealing with high dimensional datasets where the feature selection due to the potential correlation is crucial. According to the results, regular SVM outperforms SVM with radial basis kernel function. The observations suggest that there is no significant non-linear reliance in the dataset, and the relationship between the features and the survival time can be approximated by linear functions. Therefore, the non-linear function is not required in this case to perform the regression while using the SVM. When one gets to the boosting methods, the results indicate that the Stagewise GB algorithm is not only more accurate compared to the other boosting method, but also it beats other algorithms in discharge time prediction in terms of accuracy. The benefit of the ensemble method is to take advantage of a collection of decision trees (DT) instead of one predictor, so it tends to yield the best results in this study. When It comes to other methods, deep learning is a powerful method that can be applied in various fields and similar to the aforementioned

methods, it can be extended to handle censored data[11-13]. This method requires a larger number of samples as well as the boosting methods, so their performance can be fairly compared in future studies.

Hospital Discharge Rate

By leveraging the properties of reversed Kaplan-Meier (KM) estimator, probabilities of patient discharge time from the hospital can be estimated. According to Figure 1A, the probability of getting recovered and discharged in male hospitalized patients in the first 15 days beginning from showing the symptoms, is higher than in females. Although this probability is higher for the first 15 days, after day 15 up until nearly 40 days of showing the symptoms, the probability of recovery in women is slightly higher. This suggests a higher average probability of discharge from hospital for females, longer recovery times, and a higher average morbidity rate for males compared to females. According to the hazard ratio (HR) in association with sex that was obtained from the Cox regression, females have approximately 5 percent more chance on average than males to discharge from hospital. In addition, as reported by Pan et al.[14], the average discharge time probability of 21 male and female patients from being hospitalized to get discharged is almost 1 after 17 days. However, based on Figure 1B, this probability is approximately 1 after 27 days. The difference between data used and the number of patients can account for the discrepancy between these reported results. Nonetheless, since a far greater number of cases are studied in this paper, the result's variability must be lower.

Age is one of the risk factors that is used in this study to predict survival times and is of great interest to be studied to determine the impact of it on patient survivals. First, second, and third quartiles of age ranges are utilized to categorize age into four subgroups. Figures 1C and 1D shows the effect of age on hospital discharge rates. It is evident that there exist clear boundaries between these age groups. According to this figure, lower hospital discharge rates are associated with older age groups. Beside the Kaplan-Meier results, the coefficient of Cox regression regarding the age suggest that by increasing the age for 1 unit (year), the probability of discharging from the hospital decreases approximately 3 percent.

Finally, a separate analysis is conducted for cases older than the age median, 46, to examine the effect of sex in older patients. The results indicate that the probability of recovery in females after 35 days is equal to 0.86 and 0.83 after 37 days for the males. This suggests that older females have slightly higher survival rates. These results are also in agreement with the initial outcomes of clinical researches concerning the influence of sex[15, 16].

Conclusion

The clinical data from 1182 COVID-19 patients is used in this paper to measure the prediction accuracy of the discharge time of hospitalized patients by implementing different survival analysis models. Firstly, the results indicate that Stagewise GB delivers the most accurate discharge time prediction compared to the other algorithms while using only age and sex as model features. It is worthwhile to note that since predictions are based on age and sex as model features, this study provides a baseline criterion for future studies once more detailed clinical data is available. Secondly, the Kaplan-Meier and Cox regression method results suggest that sex and age of the hospitalized patients have a direct effect on their recovery time. Findings indicate that being male or being in older age groups is associated with lower hospital discharge probabilities. This study provides a baseline for recovery time prediction for future research studies. Upon the accessibility of other risk factors such as patient pre-conditions, there will be an opportunity to measure the impact of them on patient survival.

Experimental Procedures

Resource Availability

Lead Contact:

Mohammadreza Nemati is the lead contact of this study and can be reached through e-mail: mnemati@rockets.utoledo.edu

Material Availability:

The machine learning and statistical models used in this study can be obtained via this GitHub repository:
https://github.com/Mnemati/Machine-Learning-Approaches-in-COVID-19-Survival-Analysis

Data and Code Availability

All the raw data used in this project are obtained from an open access COVID-19 epidemiological data website (https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30119-5/fulltext). The code is available at GitHub (https://github.com/Mnemati/Machine-Learning-Approaches-in-COVID-19-Survival-Analysis).

Method Details

Data Description and Preparation

An open-access dataset is used in this research study. This dataset was collected by a group of researchers from different universities and research labs[17]. According to the dataset descriptions, data is mostly extracted from national health reports and online resources, released mainly by state/local health officials and hospitals of different countries. The epidemiological information includes various features about the surveyed cases, including case ID, age, gender, the onset date of symptoms, date of hospitalization, infection confirmation date, death or discharge time, death or discharge status, symptoms, chronic disease history, travel history, and location. Several filtering processes are applied to prepare the data for training and statistical analysis. Incomplete cases with missing data points are first removed from the dataset. Among available fields in the dataset, only a limited number of features, including age, sex, available dates, and outcome (death or discharge), are kept in the dataset. Finally, due to format inconsistency in some fields such as age and outcome parameters, the filtered dataset is reformatted. One of the stumbling blocks of survival analysis is to calculate the days from the beginning of the study to the event date (discharge) or the last available date (censoring days). Although the beginning date is available for all the cases, censoring days are calculated by subtracting the last available date from the beginning date for each case. The next processing step is to restructure the data to make it compatible with survival analysis methods. The first component in the structured data is the status of the case (censored or uncensored). For cases with discharged outcomes, the status is considered uncensored (True), while for occurrences with no available outcome information, the status is set to censored (False). According to the number of censored and uncensored cases, there is no severe imbalance identified between these two classes, and thus there is no further need to handle the class imbalance[18]. The second component is the event or censoring days. In the final stage, age and sex are added as the predictor variables, so a more detailed description of them is vital. One of the main objectives of this study is to evaluate the impact of different age and sex categories on patient survivals. As illustrated in Figure 2, age is stratified into 4 categories by determining its quartiles. Regarding the sex, it is a categorical variable, so a dummy encoding is applied to it to transform it to a numerical value for further analysis. Nonetheless, one of the limitations of this study is that this dataset does not contain other potential risk factors such as blood type, body mass index (BMI), or barely contains patient pre-

conditions. Therefore, we are not able to add them beside age and sex. According to the Figure 3, after the data processing step is finished, survival algorithms can perform the time-to-event analysis.

Survival Analysis Methods

Survival analysis is a well-established technique in statistics used to predict time to the event of interest during a specific observed time interval. Survival analysis is widely used in economy[19] and healthcare[20] for numerous applications. Prediction of death time after cancer treatment and predication of time between the first heart attack and the second attack are some of the survival analysis examples in healthcare domain[21-23]. In this paper, the event of interest is the time when a patient is discharged from the hospital.

Survival analysis is a form of regression by which a continuous variable is to be predicted. However, the main difference between this type of regression and the conventional regression techniques is that unlike ordinary regression, the training data for survival analysis is partially observed. In other words, the exact time of the event is unknown. This type of sample is called censored. Some of the censorship conditions are depicted in Figure 4. As it can be seen, patient D does not experience the event during the observation period. A case might leave the study, such as patient A. Another situation is when the patient's status cannot be determined due to lost or incomplete records (patient B). Due to different possible censorship circumstances, standard predictive models are not applicable to this problem. In this study, it is assumed that the start date is identical to the symptom onset or hospitalization date. However, the discharge time information is unknown for many samples. For these cases, the last follow-up date is considered as the censoring time.

Different statistical and machine learning methods have been developed by survival analysis researchers to address prediction problems in various fields. Based Figure 5, among the statistical practices, Kaplan-Meier (KM) estimator, Cox Proportional Hazard (CoxPH), Coxnet, and Accelerated Time Failure are selected for evaluation. Additionally, several machine learning approaches, including, Stagewise Gradient Boosting, Componentwise Gradient Boosting, and Support Vector Machines are used. These techniques are discussed and compared in the following.

Kaplan-Meier Estimator

As discussed earlier, massive amounts of data can be censored to generate partial information. In some applications, however, it is ideal to avoid reducing the sample size. Kaplan-Meier (KM) estimator, also known as a product limit estimator, is a powerful non-parametric method capable of computing survival. The incidence probabilities of an event are first calculated at a specific time. These consecutive probabilities are then multiplied to achieve the final survival estimation[24]. Despite its benefits, KM estimator has some limitations. For instance, KM is not an appropriate estimator to account for the effects of a variety of covariates on survival simultaneously. Also, unlike regular healthcare problems in which the event of interest is typically the occurrence of a failure such as the next heart attack or kidney graft loss, in this work, the event of interest is the time that patient recovers. So, a modified version of the KM estimator, known as the reverse of KM estimator, is implemented.

Cox Proportional Hazard

Unlike KM estimator that cannot handle multiple features at the same time, CoxPH enables the simultaneous processing of numerous features. CoxPH is a widely used, linear, and semi-parametric technique that estimates the effect of each survival variable on the entire cohort. According to Wang et al.[25], CoxPH relies on assumptions and restrictions that limit its applications. The features are assumed to have an exponential impact on the outcome. Also, it is assumed that different individuals have identical hazard functions. More importantly, since the baseline hazard function $h_0(t)$ remains unspecified, it is not a well-suited model in some real-world problems.

Coxnet

One of the shortcomings of CoxPH is its vulnerability to overfitting in high dimensional, massive sample size datasets. Due to this issue, not only the training time can be considerably high, but also CoxPH is likely to memorize the training samples. Also, CoxPH is not effective when there is multicollinearity in the dataset. To address these shortcomings, a regularized version of CoxPH called Coxnet model is evaluated[26]. The modification is achieved by adding different penalties.

- L1 regularization: It adds an L1 penalty. L1 can lead to sparse models where the model has a few coefficients. Lasso regression uses L1 regularization.
- L2 regularization: It adds an L2 penalty. Despite the previous one, it does not yield a sparse model. Ridge regression uses this method.
- Elastic net: A combination of the two previous models yields to the elastic net model. A classic regressor model with an elastic net penalty is called Coxnet model. Since sufficient clinical data is not yet available for COVID-19, and a limited number of covariates are used, it is not necessary to apply dimensionality reduction methods[27, 28]. Therefore, it is expected for CoxPH and Coxet to yield similar results in this study.

Accelerated Failure Time Model

Although previous models are robust regression techniques, other types of regression models are available that might yield useful information for interpretation purposes. Accelerated Failure Time model (IPCRidge) lies in the category of parametric and linear models with a different form of regression. In this model, samples are weighted by the inverse probability of censoring, and the censoring status remains independent of covariates[29].

Stagewise Gradient Boosting

Stagewise Gradient Boosting (GB) is an ensemble, boosting machine learning technique. This algorithm integrates weak learners, into a weighted sum where it builds a powerful and specialized learner[25]. In fact, base learners do not perform independently, and each successive tree gives extra weight to the points that were incorrectly predicted by earlier predictions. each tree is trained on an ever-more specialized sub sample of the training set[30]. This algorithm uses an ensemble of learners to determine how the hazard function changes in regard with the features and has been proven to be effective for real clinical datasets in many cases[31]. Therefore, extending this algorithm to handle censored datapoint and survival analysis is of great interest and according to previous researches, Cox model and Decision Trees have been used to develop survival GB[25, 31]. Despite the power of this algorithm, since it uses numerous learners, time complexity for tuning the hyperparameters specifically in high dimensional

settings can be quite high. Moreover, in the contexts where the number of samples is not enough, this algorithm can potentially have other shortcomings, such as low prediction accuracy on the test set.

Componentwise Gradient Boosting

Unlike Stagewise GB, Componentwise GB algorithm aims at estimating the coefficients either by updating one component of β or by fitting the gradient with the help of all covariates in each step. The algorithm calculates the gradient of the log-partial likelihood and then fits this gradient to the input matrix by a so-called base procedure such as least squares estimation[32]. Like the previous boosting method, the training time can be considerable while dealing with massive datasets.

Support Vector Machine

Support Vector Machine (SVM) is a standard supervised machine learning algorithm which is widely used for regression and classification and has wide applications in healthcare problems such as predicting organ (e.g. liver) disease[33]. Prior researches have extended the properties of this algorithm to enable handling censored data in survival analysis[34-36]. By applying an updated asymmetric form of the penalty function, survival SVM can take advantage of regular SVM's abilities in handling high dimensional data while adapting it for censored and uncensored samples. By using a Kernel function and transforming the data into higher dimensions, the margins between different classes could be maximized in the case of non-linearity. Both linear and kernel SVMs are used to handle survival analysis problems. In this work, more efficient versions of SVM called Fast SVM and Fast Kernel SVM are implemented[36].

Acknowledgements
The Author wish to thank Freepik.com and Medical Product Outsourcing. The cover and Figure 3 have been designed using these two resources, respectively.

Author Contribution
All authors conceived of the study and reviewed the manuscript. M.N. ran the models and supervised the group activity. M.N, J.A., and N.N. were involved in validating the outcomes, data visualization, writing the original draft, review, and editing.

Declaration of interests
The authors declare no competing interest.

References

1. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y. and Xing, X., (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.
2. Mahase, E., (2020). Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate.
3. Organization WH. (2020). Coronavirus disease 2019 (COVID-19): situation report, 61.
4. Wu, Z. and McGoogan, J.M., (2020). Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Jama*, *323*(13), pp.1239-1242.

5. Dowd, J.B., Andriano, L., Brazel, D.M., Rotondi, V., Block, P., Ding, X., Liu, Y. and Mills, M.C., (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences*, *117*(18), pp.9696-9698.
6. Livingston, E. and Bucher, K., (2020). Coronavirus disease 2019 (COVID-19) in Italy. *Jama*, *323*(14), pp.1335-1335.
7. Ji, J.S., Liu, Y., Liu, R., Zha, Y., Chang, X., Zhang, L., Zhang, Y., Zeng, J., Dong, T., Xu, X. and Zhou, L., (2020). Survival analysis of hospital length of stay of novel coronavirus (COVID-19) pneumonia patients in Sichuan, China. *medRxiv*.
8. Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., Shi, J., Zhou, M., Wu, B., Yang, Z. and Zhang, C., (2020). Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *Journal of Allergy and Clinical Immunology*.
9. Du, R.H., Liang, L.R., Yang, C.Q., Wang, W., Cao, T.Z., Li, M., Guo, G.Y., Du, J., Zheng, C.L., Zhu, Q. and Hu, M., (2020). Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *European Respiratory Journal*, *55*(5).
10. Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B. and Wei, L.J., (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, *30*(10), pp.1105-1117..
11. Fotso, S., (2018). Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*.
12. Farhangi, A., Bian, J., Wang, J. and Guo, Z., (2019), December. Work-in-Progress: A Deep Learning Strategy for I/O Scheduling in Storage Systems. In *2019 IEEE Real-Time Systems Symposium (RTSS)* (pp. 568-571). IEEE.
13. Fotso, S., (2019). PySurvival: Open source package for survival analysis modeling.
14. Pan, F., Ye, T., Sun, P., Gui, S., Liang, B., Li, L., Zheng, D., Wang, J., Hesketh, R.L., Yang, L. and Zheng, C., (2020). Time course of lung changes on chest CT during recovery from 2019 novel coronavirus (COVID-19) pneumonia. *Radiology*, p.200370.
15. Ruan, Q., Yang, K., Wang, W., Jiang, L. and Song, J., (2020). Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive care medicine*, *46*(5), pp.846-848.
16. Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X. and Guan, L., (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet*.
17. Xu, B., Gutierrez, B., Mekaru, S., Sewalk, K., Goodwin, L., Loskill, A., Cohn, E.L., Hswen, Y., Hill, S.C., Cobo, M.M. and Zarebski, A.E., (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific data*, *7*(1), pp.1-6.
18. Zoghi Z, Serpen G. (2020). UNSW-NB15 Computer Security Dataset: Analysis through Visualization. Proceedings of International Conference on New Computer Science and Engineering Trends (NCSET2020)
19. Ji, W., Wang, X. and Zhang, D., (2016), October. A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th acm international on conference on information and knowledge management* (pp. 1373-1382).
20. Reddy CK, Li Y., (2015). A Review of Clinical Prediction Models. Healthcare data analytics. 36:343-78.
21. Pölsterl, S., Gupta, P., Wang, L., Conjeti, S., Katouzian, A. and Navab, N., (2016). Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. *F1000Research*, *5*.
22. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, pp.8-17.

23. Abbasi-Kesbi, R., Memarzadeh-Tehran, H. and Deen, M.J., (2017). Technique to estimate human reaction time based on visual perception. *Healthcare technology letters*, *4*(2), pp.73-77.
24. Goel, M.K., Khanna, P. and Kishore, J., (2010). Understanding survival analysis: Kaplan-Meier estimate. *International journal of Ayurveda research*, *1*(4), p.274.
25. Wang, P., Li, Y. and Reddy, C.K., (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, *51*(6), pp.1-36.
26. Mittal, S., Madigan, D., Burd, R.S. and Suchard, M.A., (2014). High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics*, *15*(2), pp.207-221.
27. Lee, J.A. and Verleysen, M., (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.
28. Esmaeilbeig, Z. and Ghaemmaghami, S., (2018), August. Compressed Video Watermarking for Authentication and Reconstruction of the Audio Part. In *2018 15th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC)* (pp. 1-6). IEEE.
29. Kalbfleisch, J.D. and Prentice, R.L., (2011). *The statistical analysis of failure time data* (Vol. 360). John Wiley & Sons.
30. Schapire, R.E., Freund, Y., Bartlett, P. and Lee, W.S., (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, *26*(5), pp.1651-1686.
31. Chen, Y., Jia, Z., Mercola, D. and Xie, X., (2013). A gradient boosting algorithm for survival analysis via direct optimization of concordance index. *Computational and mathematical methods in medicine*, *2013*.
32. Bühlmann, P. and Yu, B., (2003). Boosting with the L 2 loss: regression and classification. *Journal of the American Statistical Association*, *98*(462), pp.324-339.
33. Fathi, M., Nemati, M., Mohammadi, S.M. and Abbasi-Kesbi, R., (2020). A MACHINE LEARNING APPROACH BASED ON SVM FOR CLASSIFICATION OF LIVER DISEASES. *Biomedical Engineering: Applications, Basis and Communications*, p.2050018..
34. Khan, F.M. and Zubek, V.B., (2008), December. Support vector regression for censored data (SVRc): a novel tool for survival analysis. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 863-868). IEEE.
35. Pölsterl, S., Navab, N. and Katouzian, A., (2016). An efficient training algorithm for kernel survival support vector machines. *arXiv preprint arXiv:1611.07054*.
36. Pölsterl, S., Navab, N. and Katouzian, A., (2015), September. Fast training of support vector machines for survival analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 243-259). Springer, Cham.

Titles and legends

Titles

Figure 1: Probability Estimation of Discharge Time in Different Age, and Sex Groups
Figure 2: Age Variation of 1182 Patients
Figure 3: Data Processing Steps
Figure 4: Demonstration of Data Censorship Status
Figure 5: Survival Analysis Algorithms

Legends

Figure 1: (A) Discharge time probability estimation of sex groups after showing the symptoms.
(B) Discharge time probability estimation of sex groups after hospitalization.
(C) Discharge time probability estimation of 2 categories of age groups.

(D) Discharge time probability estimation of 4 categories of age groups.

Figure 2: Patients are categorized into 4 different age groups. First, second and third quartiles are 34, 46, and 60, respectively.

Figure 3: (A) Data collection and filtering
(B) Data processing steps required for analysis

Figure 4: Patient A, B, D have not experienced any events until the end of study, so they are considered as censored samples, but patient C is not censored because the event has occurred and it is fully observed.

Figure 5: Survival analysis techniques applied on COVID-19 data to predict survival time and hospital discharge time probabilities

Al	IP	C	C	Stagev	Componenty	Fa	Fast kern
R	4	1	1	71	70.6(	1	61.0

Table 1 title: Prediction Accuracies of 7 Survival Analysis Algorithms

Table 1 legend: Performance (C-index) comparison of IPCRidge, Cox Proportional Hazard, Coxnet, Stagewise Gradient Boosting, Componentwise Gradient Boosting, Fast Support Vector Machine, and Fast Kernel Support Vector Machine.

A

Data collection

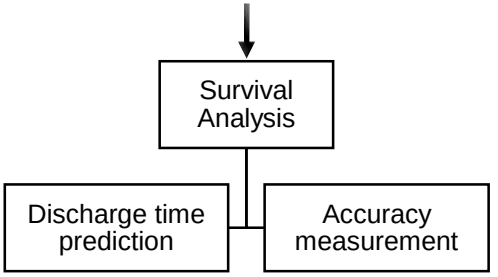
Raw data (CSV)

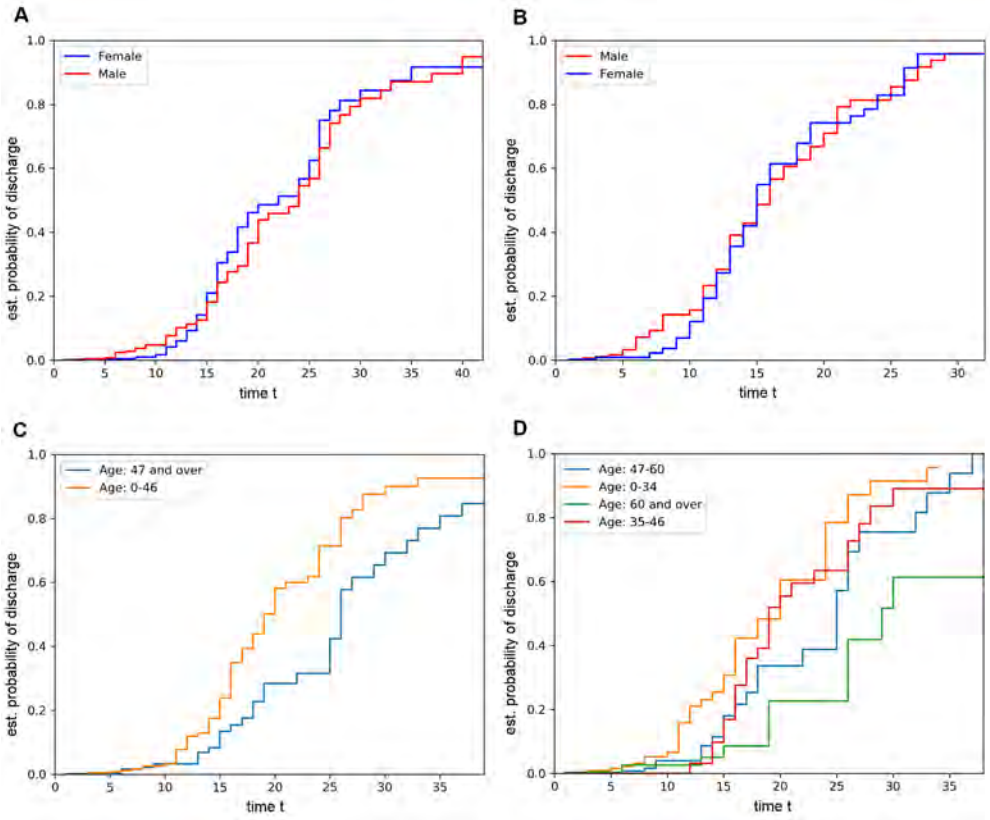
Age	Sex	Fatigue
32	F	1
44	F	1
29	M	0
39	M	0
39	F	0
46	M	0
66	M	0

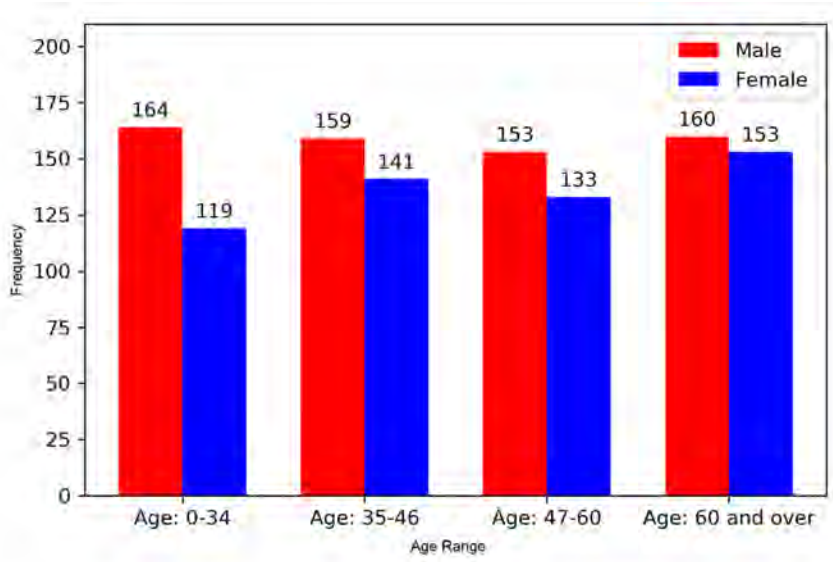
B

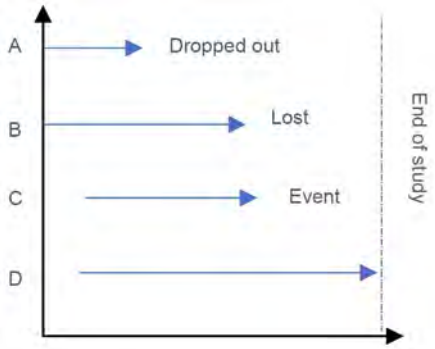
Data processing

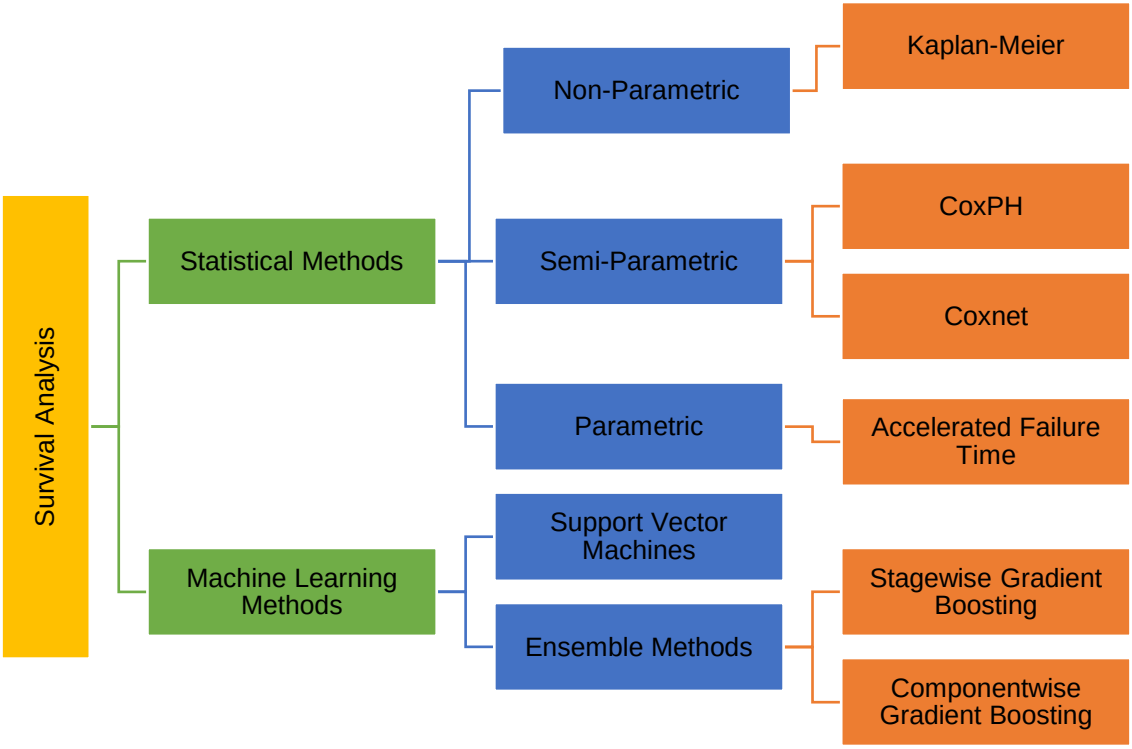
Age	Sex	Onset	Admission	Confirm	Outcome	Discharge Date	Status
30	M	1-18-2020	1-20-2020	1-22-2020	Discharge	1-29-2020	T
47	M	1-10-2020	1-21-2020	1-23-2020	Censored: Experiencing no event.		F
49	M	1-15-2020	1-20-2020	1-29-2020	Confirm – Onset is equal to censoring days		F
47	F	1-17-2020	1-17-2020	1-23-2020	Uncensored: Event occurred		F
59	F	1-19-2020	1-19-2020	1-26-2020	Discharge	2-16-2020	T
30	M	1-17-2020	1-17-2020	1-25-2020			F
39	M	1-20-2020	1-20-2020	1-23-2020	Discharge	1-27-2020	T




A
Female
Male
est. probability of discharge
time t
B
Male
Female
est. probability of discharge
time t
C
Age: 47 and over
Age: 0-46
est. probability of discharge
time t
D
Age: 47-60
Age: 0-34
Age: 60 and over
Age: 35-46
est. probability of discharge
time t






Survival Analysis
Statistical Methods
Machine Learning Methods
Non-Parametric
Semi-Parametric
Parametric
Support Vector Machines
Ensemble Methods
Kaplan-Meier
CoxPH
Coxnet
Accelerated Failure Time
Stagewise Gradient Boosting
Componentwise Gradient Boosting

In brief

COVID-19 has spread to many countries in a short period, and overwhelmed hospitals can be a direct consequence of rapidly increasing coronavirus cases. In this study, by choosing patient discharge time as the event of interest, survival analysis techniques including statistical analysis and machine learning (ML) approaches are used to build predictive models capable of predicting patient period of stay in hospital. This time is crucial because it allows decision makers to be prepared for hospital overloads.

Bigger Picture

A record-breaking pressure has been placed on healthcare systems by the COVID-19 pandemic. As a result of fast-growing requests for medical care in hospitals, with limited space and number of intensive care units (ICUs), estimation of the length of stay of patients with COVID-19 in hospitals can provide insightful information to decision-makers for efficient allocation of equipment and managing hospital overload in different countries. This work introduces statistical models and machine learning (ML)-based approaches that can be directly applied to real-world COVID-19 data to predict the patient discharge time from hospital and evaluate how the patient clinical information could have an impact on the length of stay in hospital. While considerable insights have been achieved about the patient recovery times in this paper, applications of these data-driven approaches are expected to gather substantial interest in the near future once more detailed clinical data is available.

Highlights

- 1182 hospitalized patients were studied in this research.
- Survival analysis can be applied to predict patient length of stay in the hospital.
- We used 7 machine learning (ML) and statistical analysis techniques.
- The impact of clinical covariates on survival times was studied.