Competing risk between in-hospital mortality and recovery: An application of DeepHit on COVID-19 clinical data

Lintu M.K.[1] **and Asha Kamath**[1,*]

[1]Department of Data Science, Manipal Academy of Higher Education, Manipal, Karnataka, India
[*]asha.kamath@manipal.edu

ABSTRACT

Unexplained pneumonia appeared in Wuhan was soon determined to be a novel coronavirus disease, referred to as COVID-19. On March 11, 2020, WHO characterized COVID-19 as a pandemic and the virus has been recognized as a global threat. A plethora of studies is being carried out using various statistical and mathematical models to predict the probable evolution of this pandemic. Though most of them are focusing on building predictive models to assess mortality rates and risk, concentrating on the length of hospital stay can improve decision making and treatment plans. While modeling the length of stay, possible outcomes observed are either discharge or mortality. Modeling the duration of recovery and death provides valuable information for health officials to design proper strategies to reduce the burden on the health system during the outbreak. In this study, we are exploring this competing event aspect of the survival data obtained from COVID-19 patients using the state-of-the-art model DeepHit, a discrete survival model.

Introduction

The coronavirus disease 2019(COVID-19) outbreak evolved into a pandemic that spread rapidly worldwide. The pandemic is still unfortunately under progression. The virus has affected most of the countries around the world. The severity of the disease is associated with a large number of co-morbidities. From the previous studies, it has been found out that gender, age and pre-existing chronic diseases such as cardiovascular disease, cancer, respiratory disease, etc. are associated with increased risk for adverse outcomes.[1,2] In addition to this information, it is important to focus on the length of hospital stay and other survival aspects of the COVID-19 patients. Because of the lack of clinical data, not many studies have come out with these objectives.

Survival analysis considers the problem of analyzing data where the target variable is time until the occurrence of an adverse event. Events could be anything such as death, hospitalization, hospital discharge, etc. The challenge underlying the scenario is to learn the model parameters from survival time data while handling censoring.[3] The concept of censoring comes into the picture since the survival time is unknown for some of the subjects. Lost to follow-up or culmination of the study period results in censoring.[4] Several models have been developed to analyze the lifetime data. The fundamental problem is understanding the relationship between the covariates and time-to-event.[5] The typical application of survival analysis is, analyzing the risk of death onset of a certain disease. Management of a case can be bettered if the accurate prediction of survival time is feasible.

The traditional format of survival data consists of a tuple for each of the subjects. The tuple with covariates x, time until the first event or censoring, and a label indicating the censoring status. While dealing with time-to-event data, it is common to observe more than one outcome occurring among the subjects, which we call competing events. Competing risk setting is ubiquitous in epidemiological studies and clinical trials since the subjects are likely to experience multiple possible adverse events. Since the problem is challenging to address, most of the existing models treat one event as the event of interest and others as right censoring.[6] Since this practice is inadequate, research on extensions of survival models to handle these complex scenarios is ongoing. It is important to incorporate competing risks since the occurrence of an event of interest is often obscured by other competing events. But incorporating this into classical survival models is not trivial.[7]

Most of the competing risk models are based on cause-specific cumulative incidence function(CIF). Cause-specific CIF gives the probability that the event k occurs on or before time t for a patient with covariates x. Fine-Gray model is one of the survival models which has looked into the competing risk aspect. Hence, it is considered as a benchmark in competing risk analysis.[8] As the patients are likely to suffer from multiple comorbidities, competing risk models are emerging as an important predictive tool in medicine.

In the time-to-event data of COVID-19 patients, discharge, and death being two events,[9] we framed the competing risks setup.[10,11] Survival analysis has received substantial attention in the machine learning community. Deep learning methods improve the performance of survival data.[12,13] The machine learning models for the competing risks attempts to

give empirical estimates of the true cumulative incidence functions. The rapid development of deep learning methods has enhanced the predictive performance of survival models.[14,15] DeepHit[6] is one such deep-learning technique developed to handle the competing events with good empirical performance. The upper hand of DeepHit over the Fine-Gray model and other deep-learning models in terms of discriminative power has been established in many studies.[6,15–17]

Implementation of data-driven models on rapidly increasing outbreaks or epidemics aids in providing better insights for controlling it. We explore the use of a deep competing risk model to analyze the survival characteristics of 2877 COVID-19 patients admitted to hospitals across the globe from December 10, 2019, to March 31, 2020. This kind of predictive modeling on real data can help clinicians to know the possible outcome of the patients in advance and improve the treatment policies.

In the rest of the paper, we present the data, explain the methods, and discuss the results.

Data

The raw data of 2877 admitted patients is obtained from an open-access COVID-19 epidemiological data website (https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30119-5/fulltext). Table 1 presents the data layout of three hypothetical subjects.

A group of researchers collected this epidemiological dataset from different research labs. The data was extracted from online resources and national health portals released by state/local health officials and hospitals of different countries. The dataset consists of a subject ID, date of hospital admission, gender, age, the onset date of symptoms, death or discharge date, discharge status, history of any chronic disease, symptoms, location, and travel history.

Since the hospital admission dates and discharge/death dates were available, the time to these events is calculated directly. Observations without the admission dates were removed from the data. Event 1 is considered as discharge from hospital and Event 2 is in-hospital mortality. Censoring time is obtained by subtracting the last available date from the admission date. Since the outcomes were well defined, there was no complication in defining death, discharge, and censoring. We have included only the patients who got admitted till March 30, 2020 to avoid a massive censoring percentage. For a few subjects, COVID-19 was confirmed after death. We have considered the covariates: age, gender, and chronic disease history. Refer Table 2. for outcome-wise descriptive statistics. All the percentages are calculated based on the remainder of 2877. The schematic plot for competing-risk time-to-event data for five hypothetical subjects can be seen in Figure 1.

Methods

DeepHit, a deep neural network that learn the distribution of survival times directly without making any assumption on the underlying stochastic process is used. The discriminative performance of DeepHit in the presence of multiple competing events is significant among the competing risks models. The model trains a multi-task network to learn the estimate of the joint distribution of the first hitting time and competing events.[6,17] Using DeepHit, we can predict the competing risks, discharge from the hospital (Event 1), and death prior to discharge (Event 2).

We used *fancyimpute*, a library for missing data imputation algorithms to deal with the missing covariate values. The algorithm uses available information from all the columns and imputes the missing values.

Since we are considering 2 competing events, the network consists of 4-layers. The first layer is a fully-connected layer for the shared subnetwork, followed by two fully-connected layers for each cause-specific sub-network. The output layer is a softmax layer. ReLu activations are used in all three layers. The network training is done by back-propagation via Adam optimizer with a batch size of 50 and a learning rate of 10^{-4}. The dropout probability of 0.1 and Xavier initialization was applied for all the layers. We used the *pycox* package to implement the model.[16]

For training, testing, and validation, 60%, 20%, and 20% of the data are randomly separated. For evaluation, 5-fold cross-validation is being applied.

The time-dependent concordance index(C^{td}) is used to evaluate the discrimination ability of the model.[18] According to the time-dependent concordance index, the predicted survival probability of a subject who experienced the event should be less than those who have survived longer. As the metric approaches one, better the performance of the model. The concordance solely evaluates a method's discriminative performance.

Results

The time-to-event data of 2877 COVID-19 patients were analyzed. Among these, 184(6.4%) discharges and 63(2.2%) deaths were reported. The socio-demographic descriptions of the events in the COVID-19 cases are depicted in Figures 2,3, and 4. Mortality was marginally higher among the males(Figure 2), higher among the elderly (Figure 3), and among those with pre-existing chronic diseases(Figure 4). Whereas, the discharge was marginally higher among males, higher among < 60 age group and those without pre-existing chronic disease condition.

The estimated joint distribution of survival time with two competing events is trained with DeepHit while incorporating right-censoring. The average discriminative index of the model from five runs of the algorithm is high(Table 3). Due to the small number of observations (2.2% of the entire dataset), the performance of DeepHit for Event 2 is variable compared to Event 1. This happens when a lesser number of Event 2 is being sampled in the validation and testing sets.

Factors like age, gender, and pre-existing chronic diseases could help in predicting the outcome of the patients(Figure 2), demonstrating the good discriminative power of the model.

It is possible to obtain estimated survival curves for each individual using DeepHit. The estimated survival curves of two patients(patients with IDs 103 and 104) generated by DeepHit is depicted in Figure 3.

Discussion

As several studies pointed out,[19–21] we could also observe that elders and patients with chronic diseases are more prone to die than other age groups. Moreover, the sex and age of the patients have a direct effect on their recovery time. Few studies have also reported that being old or male, the probability of hospital discharge is lower.[22,23]

Cox proportional hazards model has been used to study the mortality and recovery of COVID-19 patients.[10,11] The bias in estimating the hazard ratio of death in the presence of competing events have been investigated even in the context of COVID-19.[24] From those studies, there is a need of considering recovery and death due to COVID-19 as competing events to avoid a substantial risk of misleading results.[20,25]

Due to the lack of availability of clinical data, competing risk models on COVID-19 data are less explored to the best of our knowledge.[25,26] This is the first study to use a deep competing risk model on COVID-19 clinical data with discharge and death being two competing events.

As there were only a few clinical and prognostic factors of the infected subjects in this data, an elaborated analysis couldn't be carried out. Hence, there is a huge scope for further analysis when detailed data with other potential risk factors are available. An interesting expansion of the study includes competing risk modeling with a large number of features, development of more sophisticated deep learning algorithms to measure the impact of them on potential survival. Also, if we use a deep learning approach in real-time, it aids in better management of the cases and controls the outbreak of a pandemic.

Conclusion

During an outbreak, it is important to understand who is at risk of severe outcomes. In our study, the prediction accuracy of DeepHit on two competing events is obtained using data from 2877 patients who are hospitalized with COVID-19.

As it is important to consider all the competing events if a primary event of interest gets precluded by the others,[27] with the availability of large data with more clinically significant covariates, this problem can be addressed with new data-driven approaches.

References

1. Wu, Z. & McGoogan, J. M. 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**, 1239–1242 (2020).
2. Li, Q. *et al.* Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New Engl. J. Medicine* (2020).
3. Hosmer, D. W., Lemeshow, S., May, S. *et al.* Applied survival analysis: regression modeling of time to event data (2002).
4. Kleinbaum, D. G. & Klein, M. *Survival analysis*, vol. 3 (Springer, 2010).
5. Cox, D. R. Regression models and life-tables. *J. Royal Stat. Soc. Ser. B (Methodological)* **34**, 187–202 (1972).
6. Lee, C., Zame, W. R., Yoon, J. & van der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).
7. Nemchenko, A., Kyono, T. & Van Der Schaar, M. Siamese survival analysis with competing risks. In *International Conference on Artificial Neural Networks*, 260–269 (Springer, 2018).
8. Fine, J. P. & Gray, R. J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. statistical association* **94**, 496–509 (1999).
9. Awad, A., Bader-El-Den, M. & McNicholas, J. Patient length of stay and mortality prediction: a survey. *Heal. services management research* **30**, 105–120 (2017).

10. Oulhaj, A. *et al.* The competing risk between in-hospital mortality and recovery: A pitfall in covid-19 survival analysis research. *medRxiv* (2020).

11. Lu, M. & Ishwaran, H. Dynamic competing risk modeling covid-19 in a pandemic scenario. *emergence* (2020).

12. Al-Shedivat, M., Dubey, A. & Xing, E. P. Personalized survival prediction with contextual explanation networks. *arXiv preprint arXiv:1801.09810* (2018).

13. Faraggi, D. & Simon, R. A neural network model for survival data. *Stat. medicine* **14**, 73–82 (1995).

14. Bellot, A. & van der Schaar, M. Multitask boosting for survival analysis with competing risks. In *Advances in Neural Information Processing Systems*, 1390–1399 (2018).

15. Gupta, G., Sunder, V., Prasad, R. & Shroff, G. Cresa: A deep learning approach to competing risks, recurrent event survival analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 108–122 (Springer, 2019).

16. Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-event prediction with neural networks and cox regression. *J. Mach. Learn. Res.* **20**, 1–30 (2019).

17. Lee, C., Yoon, J. & Van Der Schaar, M. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomed. Eng.* (2019).

18. Antolini, L., Boracchi, P. & Biganzoli, E. A time-dependent discrimination index for survival data. *Stat. medicine* **24**, 3927–3944 (2005).

19. Jordan, R. E., Adab, P. & Cheng, K. Covid-19: risk factors for severe disease and death (2020).

20. Li, X. *et al.* Risk factors for severity and mortality in adult covid-19 inpatients in wuhan. *J. Allergy Clin. Immunol.* (2020).

21. Zheng, Z. *et al.* Risk factors of critical & mortal covid-19 cases: A systematic literature review and meta-analysis. *J. Infect.* (2020).

22. Nemati, M., Ansary, J. & Nemati, N. Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns* **1**, 100074 (2020).

23. Wang, P., Li, Y. & Reddy, C. K. Machine learning for survival analysis: A survey. *ACM Comput. Surv. (CSUR)* **51**, 1–36 (2019).

24. Li, L.-q. *et al.* Covid-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. *J. medical virology* **92**, 577–583 (2020).

25. Ghosh, S., Samanta, G. & Mubayi, A. Covid-19: Regression approaches of survival data in the presence of competing risks: An application to covid-19. *Lett. Biomath.* (2020).

26. Wolkewitz, M. *et al.* Statistical analysis of clinical covid-19 data: A concise overview of lessons learned, common errors and how to avoid them. *Clin. epidemiology* **12**, 925 (2020).

27. Alaa, A. M. & van der Schaar, M. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2326–2334 (Curran Associates Inc., 2017).

Data availability

Data is obtained from an open-access COVID-19 epidemiological data website (https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30119-5/fulltext).

Author contributions statement

Both the authors conceived the research and discussed all results. L.M.K developed the theoretical framework, analyzed data, and wrote the paper. A.K commented on drafts and approved the final version of the paper.

Competing interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

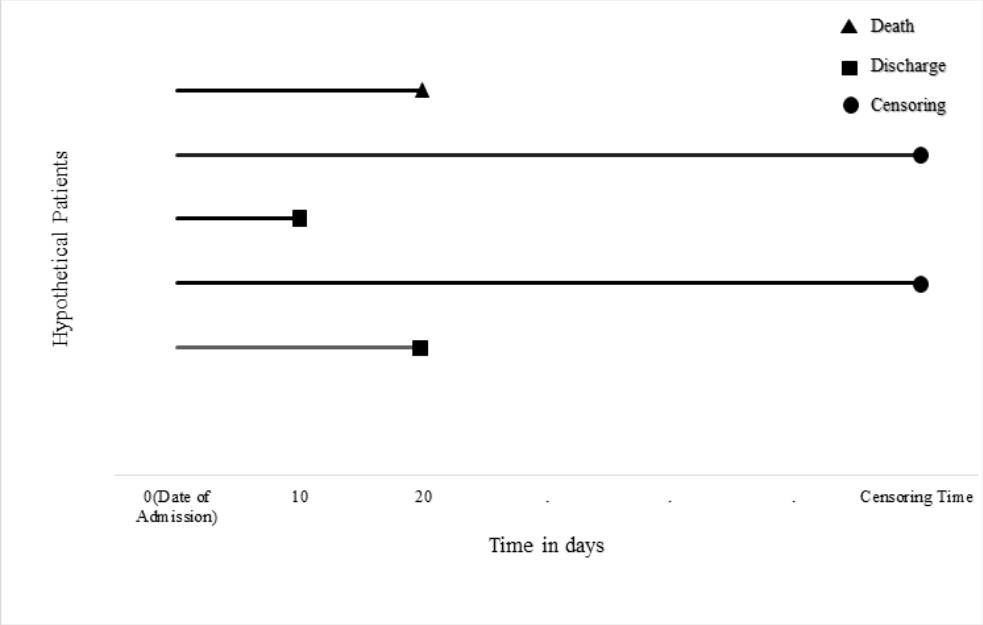

Figure 1. Schematic plot for competing-risks time-to-event data for five hypothetical subjects.

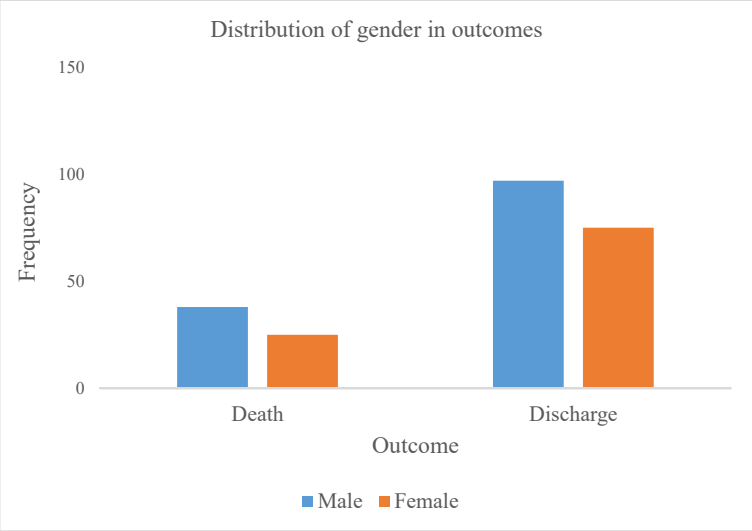

Figure 2. Frequency of the outcomes among different gender.

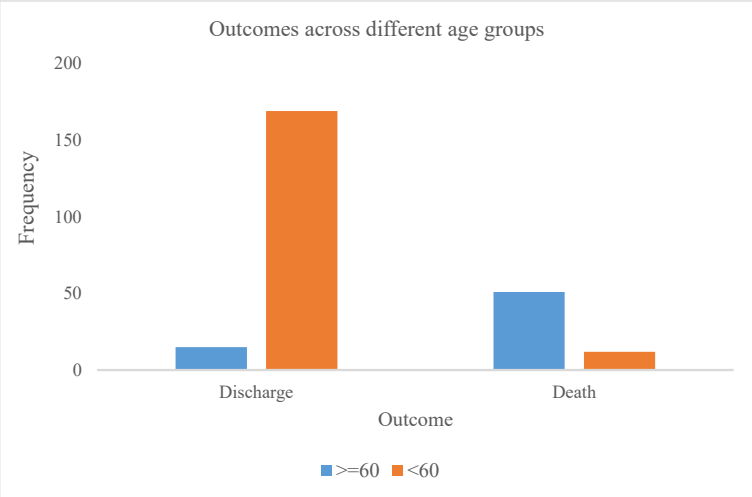

Figure 3. Frequency of discharge and death across two different age groups.

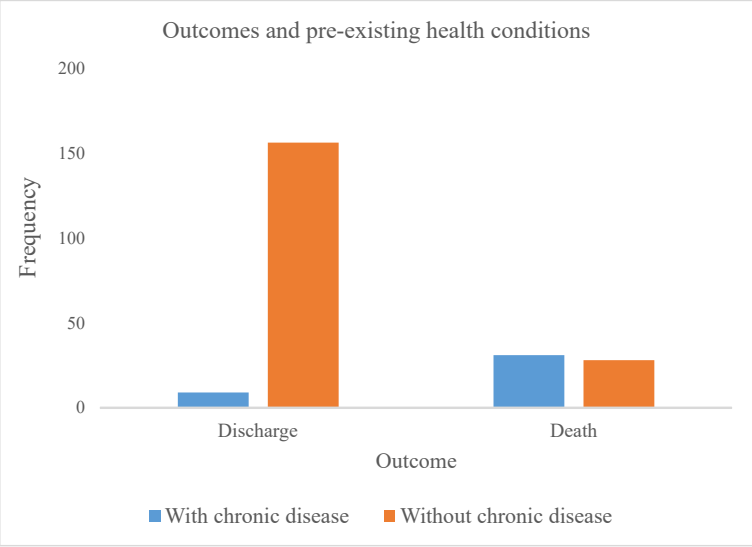

Figure 4. Frequency of the outcomes in pre-existing health conditions.

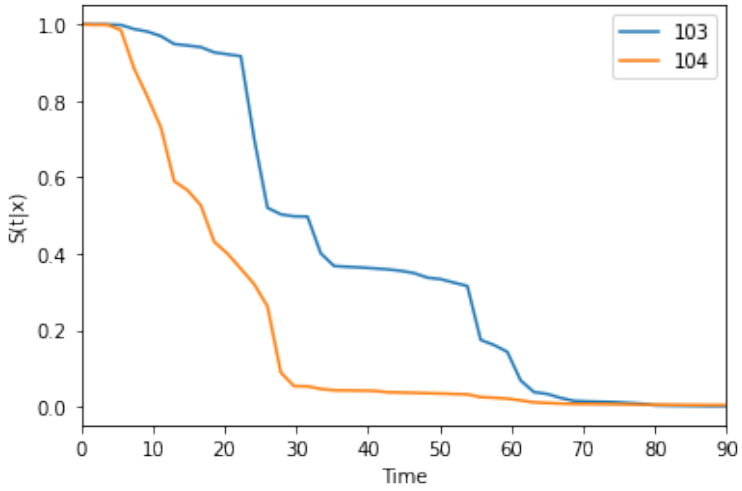

Figure 5. Estimated survival curves of two patients(103,104) generated by DeepHit. The X-axis shows the number of days from hospital admission and the Y-axis is the survival probability.

Patient ID	Age	Gender	Chronic disease	Symptoms	Date of admission	Date of outcome	Outcome
1	45	Male	TRUE	Asymptomatic	31/12/2019	12/01/2020	Death
2	25	Female	FALSE	Breathing difficulty	06/01/2020	20/01/2020	Discharge
3	63	Female	TRUE	Fever	25/02/2020	15/03/2020	Hospitalized

Table 1. Data layout of three hypothetical subjects.

Descriptive statistics	Event 1(Discharge)	Event 2(Death)
No. of subjects(%)	184(6.4)	63(2.2)
Median time in days(Range)	15(11-21)	8.5(6-12)
Median age in years(Range)	38(29.5-51)	70(65-80.8)
No. of Males(%)	97(3.4)	38(1.3)

Table 2. Event-wise summary measures of the patients.

Outcome	Mean C^{td} (95% CI)
Event 1(Discharge)	0.872(0.868,0.876)
Event 2(Death)	0.800(0.788,0.812)

Table 3. Cause-specific C^{td} index of DeepHit.