MEDI 504A – Lab 5 – Critical Appraisal of Data Science in Health studies

Paper: Machine Learning Prediction of Death in Critically Ill Patients with Coronavirus Disease 2019
By: Matthew Manson

Paper Summary: Patients who are critically ill with COVID-19 are observed to have variable mortality. Understanding a patient's risk of death using variables from the first two days of Intensive Care Unit admission could enable care teams to distribute care resources and facilitate appropriate care trajectory discussions more appropriately. Author's Churpek et al. sought to compare machine learning algorithms for predicting death and utilize such algorithms to identify important predictor variables that could be incorporated into a simple mortality prediction tool for use at the bedside. To do so, they utilized chart data from adult patients with lab confirmed COVID-19 who received care between March 4, 2020, and June 29, 2020 in one of the 68 U.S. ICU's participating in the STOP-COVID cohort study. 5075 patients were included in the analysis, which revealed eXtreme Gradient Boosting to have the highest AUC in both external and temporal validation. Important variables, including age, number of ICU beds, creatinine, lactate, arterial pH, and Pao2/Fio2 ratio, were incorporated into a simple tool for use at bedside that predicted death better than the Sequential Organ Failure Assessment score, National Early Warning Score, and CURB-65. The author's simple tool could be used to improve triage decisions and provide prognostic information for both patients and care providers.

Critical Appraisal following Dr. Ehsan Karim's Key Considerations when appraising a Machine Learning Research Article:

Author's Churpek et al. provide a generally well-done account of the work they completed to develop the simple bedside death prediction scoring system for critically ill COVID-19 patients. Below is a point-by-point commentary on each significant consideration of a well reported ML research article.

Clinical utility - The paper clearly outlines a disconnect between the number of critically ill COVID-19 patients and the resources required to care for them. Thus, they propose the development of a simple bedside mortality prediction tool that will help to direct resources more appropriately using data from the earliest days of an individual's ICU care. This is absolutely clinically relevant, given the strain that abundant critically ill COVID-19 patients puts on ICU resources.

The authours briefly justify the use of ML methods to support such tool development by pointing to sources claiming that such methods are "more flexible and accurate than traditional regression methods".

The paper clearly states its aims, imbedded in which is the outcome of interest: to compare ML methods for predicting 28-day mortality, identify important death risk factors, and develop the simple tool. Overall, this paper strongly justifies its use of proposed prediction models, clearly states its goals, and is clinically relevant.

Data source and study description – The data source and data to be used in this paper's analyses were both clearly described. Data was collected from the cohort of participants in the

STOP-COVID multicenter cohort study that enrolls adult ICU patients with COVID-19 in 68 US hospitals. Data specific to this project, including participants demographic information, comorbidities, symptoms, vital signs on ICU admission, longitudinal laboratory values and physiologic parameters, and outcome (defined in detail), were extracted manually via chart review, and thus, this is an observation study. Dates were clearly defined, and there was clarification and justification for utilizing routinely collected data from the first 48 hours of ICU admission.

Target population – The target population of this study was very clearly critically ill adult patients with COVID-19. The model was developed using data from the STOP-COVID study, which was also adequately explained, other than the lacking details about the total size of such study, and the proportion of participants used in developing the model in this study. Based on the authors' inclusion of a large and diverse range of ICU sites (detailed in appendix), it is likely that the results are generalizable to most ICUs in the US, and this is the goal.

Analytic data – Authors adequately describe eligibility criteria (Adult (>18) with laboratory-confirmed COVID-19 admitted to an ICU at a participating site between March 4, 2020, and June 29, 2020), which appear to be implemented rigorously and do capture the described target population. Data, collected manually from participants charts as described above, was preprocessed and this was described in the paper in detail: potential predictor variables were excluded from consideration if more than 50% of observations were missing, and if variable positivity was rare (defined as less than 5%). These practices are not unusual and sound.

The binary outcome (death) was clearly defined (in hospital death within 28 days of ICU admission). Patients discharged alive before day 28 were assumed to be still alive at day 28 by the authors – this seems like a fair assumption; however, the authors even went as far as validating this assumption for a subset of participants and shared these results in the appendix.

It is unclear whether clinicians were consulted to discuss the appropriateness of inclusion and exclusion criteria, but given the authors expertise, it can be somewhat safely assumed that this took place.

It is also unclear whether the protocol for this study was published a priori, however, given the context (push to get meaningful COVID related articles published for implementation into care) it's unlikely it was.

Data dimension, and split ratio – Data dimensions were somewhat clearly defined, although total size of the STOP-COVID Cohort study from which participants were drawn was not clear. Additionally, authors split data into two groups (for testing and training) at level of hospital seemingly arbitrarily (no justification for 75% 25% split). Details follow…

Total Data: Unclear

Analytic Data: 5075 patients included in study, 1846 died by day 28. Randomized split of total data at level of hospital - 75% of hospitals to be used for training, 25% to be used as primary external testing. Justification for this split not provided.

Training Data: n = 3825 patient admissions from 51 sites (75% of hospitals) between March and April 2020

Testing Data:

1. Primary "external" validation – n = 810 patient admissions from 17 sites (25% of hospitals) between March and April 2020

2. Secondary temporal validation - n = 440 patient admissions from ??? sites (unclear) between May 2020 and June 2020

Selection of hyperparameters for each model was done by 10-fold cross validation (internal validation) in the training data only, which is an uncommon practice. Additionally, typically tuning of hyperparameters is completed in its own specific dataset, and thus the trained model is at minor risk of overfitting to the training data.

The analytic data size is likely enough for many ML models explored, given how many predictor variables were considered in the ML models (~65, expanded upon below) (Steyerberg et al. 2019). Typically, ML models require between 50-200 EPV (Karim, 2021). This is especially true (on the higher end) for the data hungry methods like Neural Networks, which were considered here. In the analytic data, there is a total of 1846 events (deaths). There are ~65 variables. This means there are ~30 events per variable. This is not a terrible number, but for some of the methods tested (neural networks), more data would have been ideal.

Outcome label – the primary outcome, in-hospital death within 28 days of ICU admission was explicitly defined. This outcome is binary, and very clinically relevant – predicting likelihood death is extremely important to ensure care resources are properly directed. This outcome is a hard outcome – there is really zero uncertainty surrounding its measurement. An important note is that, as mentioned above, authors assumed that patients discharged alive before day 28 were assumed to be still alive at day 28. This assumption is conceptually fair and was validated in a small subset of data by the authors. The outcome variable is very likely of high quality.

Features – Features included in the ML models are a subset of those collected in the first 48 hours ICU admission for a typical patient – therefore, these features are very likely to be obtainable outside of the given study. Some potential features were excluded based on data missingness (as touched on above and below), and the final feature list was composed of ~65 variables (detailed in eTable3), including: age, vital signs and respiratory support on ICU admission, Fio2 among patients requiring invasive mechanical ventilation, laboratory values, and organ support.

Features utilized in the simple clinical tool were identified based on utilizing the lasso regression method to select from the list of the top 20 most important variables as identified by the XG boost model. Again, these features are typically collected in the first 48 hours of ICU admission, lending support to the potential usability of this tool. The use of the lasso regression to select the 10 variables for the simple clinical tool is unusually complex, and the authors did not provide justification for this – the authors could have utilized the top 10 variables identified by the XG boost Model without scrutiny.

Exclusion of features that had high levels of missing values is a practically strong idea – if these variables were highly missing in the Cohort study participants charts, they are also unlikely to be highly available in future patients for which these models may be used.

Although it is not explicitly stated that subject area experts were involved in the review and selection of variables, almost all authors have medical expertise and would likely have provided input.

There is no discussion of any transformation/dichotomization of variables in the base paper – looking at appendix eTable 2, it is clear little, if any, took place.

The baseline characteristics of all participants were provided, stratified by outcome, in the appendix. Baseline characteristics were well reported.

Missing data – Missing data was well acknowledged and reported (see appendix eTable 3). Missing data was utilized to eliminate potential features for the model, as described above. For those features included in the model with some missing data, the bagged trees method was used to impute the missing values (multiple imputation). This approach is sound. For the simple clinical tool development, missing values were imputed using the mode (single imputation). While this is not an ideal method of imputation, the authors justify this decision by suggesting that it will enable simple handling of missing values if using the tool at the patient bedside.

There was no attempted explanation for missing data, but it can be inferred that it is because they were not measured, as not all patients receive identical care.

Large differences in missingness can sometimes be observed between training and testing data. This is important to keep in mind when considering the results.

ML model choice – Multiple ML models were trained (individually) using the same training data, and testing was completed for all models in both testing sets on their own. The authors chose the model which performed the best (largest AUC, best at distinguishing between death positives and negatives), which happened to be the XG boost model in both the testing sets ("external", temporal), and utilized it to inform their variable selection for simple tool development. This is a sound rationale and procedure.

As discussed in the features section above, the authors utilize an atypical approach to specify the most important variables for use in the simple clinical tool (lasso regression to select the 10 variables). The simple clinical tool is then compared to the pre-established SOFA, NEWS and CURB-65 score performance, again using AUC. This is sound.

Also discussed above (data dimension, split ratio section), the analytic data size is moderately suited for the ML models tested here. The analytic data provides ~30 EPV, and most ML methods require between 50 and 200 EPV. However, the bare minimum rule for simpler regression models is 10 EPV. In an ideal world, the authors would have had some more data, especially for data hungry models like the Neural Networks.

ML model details - The performance of each ML model was well reported in the general paper (Figure 1, Results Section). Details regarding the implementation of the ML models is lacking in the general paper, however, details of the model hyperparameters searched and chosen are provided in the appendix eTable 8.

As discussed elsewhere, the models implemented were cross validated in the training data only, which is somewhat unusual. Whole data cross validation would have been the more common approach to fine tuning the model.

Optimism or overfitting – The authors utilized three approaches to try to combat optimism. First, they utilized 10-fold cross-validation (internal validation) to maximize AUC, but only in the training data (as discussed above). Second, they utilized 25% of the hospitals/data (an untouched subset) to externally validate the models. Third, they completed temporal validation (in data from the 440 admission between May and June 2020 (after the training data period)). In all cases, AUC was used as the measure of performance. The internal validation results were not reported, however, there was little discrepancy between the performance of each model when comparing the two external validations (external, temporal). To me, the model performance was reasonable, and optimism was adequately considered.

Generalizability – The authors externally validated all considered ML models two ways (using the "external" and temporal datasets). The XG boost method was selected by the authors as it performed best in both external validations.

The first external validations utilized the patient data collected at those 25% of hospitals excluded from the training dataset, during the same time frame as the training data. Although there is a lack of clarity in how this data was split/selected, this method of validation is sound and lends support to generalizability. In the second external validations, the authors tested the models' using data from outside (afterward) the testing data period. This can be considered temporal validation. Together, these two external validations in real world clinical data point to strong generalizability.

Reproducibility – No code or data is provided in this report. The authors do report the software (R, no version provided), the software package (caret, no version provided), and the model parameters utilized in this project (eTable 8). Although reproducing this study without data is impossible, the lack of data is not unusual for health research. With a data request, it is possible that these analyses could be reproduced. Computing time was not reported by the authors.

Interpretability – It is not explicitly stated that clinicians were consulted at any point throughout the reported project. However, most authors are clinically trained and have expertise in caring for pulmonary diseases. Therefore, it is likely that the clinical context was heavily considered throughout this study. Additionally, reference to other literature citing the importance of many variables (to COVID-19 death outcomes) considered in the models is present in the paper, lending credibility to the results. The final product, the simple model, is quite easily understood, straight forward to use, and makes sense in the context of literature and general understanding of COVID-19 patients. It is likely to be easily adopted by clinicians if they are interested in such a tool.

Subgroup – There is no consideration/discussion of subgroups in this article. These considerations do not appear to be relevant, given the author's goals of a simple tool that is generalizable to any adult individual with critical COVID-19 illness.

References:

Churpek, M. M., Gupta, S., Spicer, A. B., Hayek, S. S., Srivastava, A., Chan, L., Melamed, M. L., Brenner, S. K., Radbel, J., Madhani-Lovely, F., Bhatraju, P. K., Bansal, A., Green, A., Goyal, N., Shaefi, S., Parikh, C. R., Semler, M. W., & Leaf, D. E. (2021). Machine learning prediction of death in critically ill patients with coronavirus disease 2019. *Critical Care Explorations*, *3*(8). https://doi.org/10.1097/cce.0000000000000515

Karim, E. (2021, November 22). *Understanding basics and usage of machine learning in medical literature*. Chapter 2 Prediction from continuous outcome. Retrieved December 6, 2022, from https://ehsanx.github.io/into2ML/prediction-from-continuous-outcome.html#overfitting-and-optimism

Steyerberg, E. W. (2020). *Clinical prediction models: A practical approach to development, validation, and updating*. SPRINGER.