A Glimpse Into The Future: The Ethical Implications of Predicting Patient Mortality

Ava Klissouras

12/9/24

Introduction

Imagine waking up one day with the ability to predict the future. Furthermore, imagine if you could use this ability to predict whether or not someone lives or dies. How would you use this newfound skill? Would it be ethical to share your predictions with others, or would it influence the way those predicted to die—or survive—are treated within a clinical setting? Predicting the future has long been viewed as an unattainable omniscience, which can be possessed only within the realm of science fiction novels and superhero comics. However, thanks to the power of machine learning algorithms, this ability may soon exist within our human grasp—at least, in the case of mortality prediction. In *Development of a prognostic model for mortality in COVID-19 infection using machine learning*, Adam L. Booth, Elizabeth Abels, and Peter McCaffrey develop a machine learning algorithm that can predicts patients' COVID-19 mortality up to 48 hours prior in advance. However, this paper raises pressing ethical concerns regarding the moral soundness of predicting mortality. In America's bleak, overstretched, and underfunded medical landscape, should mortality prediction be integrated into clinical settings in order to optimize the allocation of scarce resources? Or could this attempt to declare a patient's fate before it runs its course do more harm than good, overshadowing the possibility of medical miracles and divesting resources away from those who might need them? I argue that despite its potential benefits, based upon principles of Deontology, the implementation of mortality prediction algorithms to influence resource allocation is ultimately unethical.

A Brief Summary of Methods

Development of a prognostic model for mortality in COVID-19 infection using machine learning (Booth et al.) presents a retrospective study, which evaluates laboratory data and mortality from 398 COVID-positive patients at the University of Texas Medical Branch. The authors retrospectively searched the hospital's Laboratory Information System and located patients' values for twenty-six serum chemistry and blood gas laboratory parameters, only including patients admitted to the hospital (to reduce the quantity of missing test results and thus lessen the need for later imputation). Then, they performed multivariate feature imputation, ran a logistic regression to determine feature importance (based on regression coefficients), and selected the most important features to include in their machine learning model. Then, the authors trained a Support Vector Machine with a radial basis kernel due to hypothesized nonlinear interactions between features. Finally, they calculated Shapley values for each of the five features in order to determine their relative influence on model predictions. However, they made a number of decisions in performing their analysis of the data that have the capacity to influence the reliability of their results, which I will subsequently describe, critique, and offer improvements for.

Analysis and Critique of Methods

In their analysis, the authors only included data from patients admitted to the hospital with positive COVID test results (meaning they excluded patients who had COVID but were not admitted). They reasoned that including only admitted patients would reduce the number of laboratory test results missing from each patient, thereby lessening the quantity of data that would require imputation later on in the study. Moreover, the authors excluded laboratory tests for which less than 25% of patients had measured values within 14 days following a positive COVID-19 test. They also excluded results captured within 48 hours of death, in order to ensure a sufficiently early identification of patients who are likely to pass.

The authors' choice to include and exclude certain data is well-reasoned and pragmatic. However, it may have inadvertently introduced bias, and limited their ability to uncover other important predictions from the data. First, the data are representative of a single institution—the University of Texas Medical Branch. Unless the patients reflect the population of the United States as a whole, in terms of gender and race diversity, baseline health, age, etc., it is difficult to generalize the results of this study to the entire population. Although access to medical data can be limited due to the need for IRB approval, future work

could combine medical records from various hospitals around the United States to determine if the trends that the authors' model predicted persist across a diverse population. Moreover, filtering on admitted patients neglects patients who fail to seek hospital care due to cost concerns. According to KFF, in 2024, one in four adults have put off seeking healthcare due to cost, and six in ten uninsured adults have gone without needed care for the same reason (Lopes et al.). By excluding patients who were not admitted to the hospital, insights generated by the authors' algorithm may exclude data from financially challenged or uninsured patients. Additionally, lab results did not explicitly capture any comorbid illnesses. Therefore, some of the patient deaths that occurred may have been exacerbated by other illnesses, which are not accounted for in the algorithm's predictions.

To address missing data, the authors performed multivariate feature imputation using Scikit-Learn's IterativeImputer method. This technique models each feature containing missing values as a function of the other features, and estimates the missing value using that model. The authors employed a Bayesian Ridge Regression as their model of choice, a decision which I will not discuss at length in this paper for the sake of brevity. However, I will discuss their inclusion of the expired versus non-expired variable as one of the inputs in the model. The authors chose to include this variable because they did not want to suppress their model's output towards the null hypothesis. This choice was likely beneficial in mitigating the bias of the model, as one study comparing the effect of imputation with and without including the outcome variable determined that imputation without the outcome yielded biased results. (Moons et. al.)

It is also worthwhile to discuss the authors' techniques for dealing with the imbalanced nature of the sample. The data exhibited a relative minority of COVID-positive mortalities. This can create performance bias, in which algorithms behave differently on majority and minority classes (Kaur et al.). The authors dealt with this imbalance by performing 1000 bootstrap samplings of their data set after imputing it. This technique is also called "bagging" (Bootstrap Aggregation). Bootstrapping involves repeatedly resampling a dataset (with replacement) in order to estimate population parameters from sample values. The authors did not justify their choice to address class imbalance using this method, and other techniques exist to mitigate the negative impact of class imbalances. For instance, one might employ under-sampling or over-sampling, in which (respectively) observations belonging to the majority class are removed, or observations belonging to the minority class are replicated, until the class proportions are equal. Another option is boosting, in which multiple models are fitted, which iteratively attempt to correct the errors present in the previous models. Due to the relatively small number of patients who had a positive outcome in the authors'

dataset, under-sampling would not have been feasible, as the sample size of the data would have become too small (N = 43). However, the remaining methods described above would have been feasible. In fact, a study comparing over-sampling, under-sampling, bagging, and boosting (Yap et al.) found that over-sampling resulted in the greatest improvement in sensitivity. A high sensitivity means that more of the individuals who will die from COVID are actually predicted to do so. Therefore, it is possible that Booth et. al.'s model may have benefited from oversampling as a technique to mitigate class imbalance, rather than bootstrapping.

The authors then shuffled the data into training and testing subsets, and used a logistic regression classifier to predict expiration status. A logistic regression takes the following form:

$$g(\hat{p}_i) = ln(\frac{\hat{p}_i}{1-\hat{p}_i}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_n x_n;\ \hat{p}_i \in [0,1]$$

where $\hat{p}_i$ is the predicted probability of mortality. The exponentiated regression coefficients provide the odds ratios associated with a one-unit change in each predictor. The authors state that in order to determine the most important laboratory test results contributing to COVID-19 mortality, they examined the regression coefficients. However, the values of regression coefficients are tied to the units of each variable, and if the units are incommensurate, it would be useless to compare their magnitudes. The authors do not state which units each predictor is measured in, so unless they are measured on the same scale, comparing their regression coefficients does not actually indicate their importance. What we *can* tell from regression coefficients is that, if the regression coefficient for Calcium is larger in magnitude than, say, Creatinine, then if we were able to change a patient's Calcium level by 1 standard deviation, this would have a greater impact on the log-odds of mortality than changing the patient's Creatinine levels by 1 standard deviation. Moreover, the authors selected features with the five regression coefficients greatest in magnitude to include in their machine learning model. This sort of post-hoc decision-making is subject to induce bias because the number of predictors the authors chose with the intention to "preserve parsimony" is entirely arbitrary. Typically, when determining which predictors to include in a model, it is better to set some sort of significance level a-priori, and then include the predictors which are deemed significant by a specified model selection technique. For instance, the authors could have performed forward, backward, or stepwise selection to select features instead, which add and/or remove predictors sequentially until no other variables improve the model's fit (based on a predetermined threshold of statistical significance).

After selecting the five most important predictors, the authors trained a Support Vector Machine (SVM) classifier. An SVM classifier draws an imaginary "line" (or, in this case, hyperplane) that maximizes the distance between the hyperplane and its closest observations (called the "margin"). In this way, SVM takes the form of the following optimization problem:

$$\omega* = \underset{\omega}{argmax}[\underset{n}{min}D(X_n)]$$

where $\omega*$ maximizes the minimum distance between any point and the hyperplane. The authors also chose to employ a radial basis kernel, (correctly) theorizing nonlinear interactions between predictors. Utilizing a kernel allows computations to be performed in lower-dimensional space, even if they are embedded in higher dimensions, decreasing computational intensity. Moreover, the choice of a radial basis kernel can mitigate the issue of data being non-linearly-separable. However, it is unclear whether the authors attempted to use a more simplistic kernel before moving on to a radial basis kernel. In many cases, choosing an unnecessarily complex kernel when a comparatively simplistic one will do can be detrimental to a model's performance.

After training and testing their model, the authors utilized Shapley Values to determine feature importance. A Shapley value can be thought of as the weighted average of the marginal impact of a feature on various models. These models each include a different subset of all possible predictors, and the average is weighted by the contrast between the size of the set of all possible predictors, and the size of the subset of features utilized. Shapley values are calculated as follows, where F is the set of all possible predictors, S is a subset of them, $f_{S\cup\{i\}}$ is the model with contribution from feature i, and f_s the model without contribution from feature i:

$$\phi_i = \sum_{s\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_s(x_s)]$$

In short, Shapley values show how important a feature is in shaping the model's prediction. CRP, lactic acid, and calcium had the largest magnitude Shapley values, thus exerting the strongest influence on the model's output. The Shapley values also revealed some more complex results regarding the influence exerted by each predictor. For example, BUN contributes more greatly to mortality predictions when CRP, lactic acid, and calcium do not deviate significantly from their normal values. Also, CRP influences model outcome only when Calcium is elevated, and albumin contributes greatly only when lactic acid is elevated. Increasingly abnormal values of CRP exert greater influence on model predictions, except

when BUN is high; this causes the effect of CRP to decrease. All in all, the prediction of mortality is much more complex than a simple additive relationship, as the Shapley values demonstrate a wide variety of interactions between predictors.

Analysis of Results

The authors use a variety of metrics to assess their model's effectiveness at predicting COVID-19 mortality. For instance, the logistic regression they used for the initial variable selection achieved 80% specificity and 77% sensitivity on the testing data set. Moreover, their SVM achieved 91% sensitivity and 91% specificity on the testing data. The SVM had a PPV (positive predictive value) of 62.5% and an NPV (negative predictive value) of 98.4%, where patient expiration is classified as a positive outcome. The higher NPV indicates that the model does a better job of classifying patients who do not pass away. This is to be expected, as mortality is the minority class in this analysis, and for a rare outcome, NPV typically exceeds PPV. The fact that PPV is not considerably high may generate some concern, but this simply means that around 42% of positive predictions are "false alarms." The models' specificity value indicates that the model correctly identifies 91% of individuals who will pass away from COVID-19, which is arguably a more useful metric to consider when examining mortality, if as many deaths as possible are to be prevented.

It is important to note that nowhere in their analysis did the authors consider demographic factors such as sex, race, and age. In many cases, stratifying analyses on such predictors can yield different relationships/results than those exhibited by the aggregate data set. For instance, Simpson's Paradox is a statistical phenomenon in which an association between two variables may appear, disappear, or reverse upon stratifying the population. As a relevant example, a 2020 study showed that the case fatality rate for COVID-19 was higher in Italy than in China; however, upon stratifying the data by age, the case fatality rate was revealed to be higher in China than in Italy (Von Kugelgen et al.). Therefore, it is possible that by stratifying on age (or other demographic variables), different relationships between predictors in this study may have been revealed.

Furthermore, although the aforementioned relationships between the Booth et al.'s five chosen predictors have the capacity to inform decision-making in a healthcare setting and contribute to our general knowledge of mortality from COVID-19, this information is only useful and applicable if it is correct, a decision which may be informed by validating the authors' conclusions with other existing literature. To reiterate, our authors determined the five most influential predictors of COVID-19 mortality to be CRP, BUN, serum calcium, serum albu-

min, and lactic acid. Among these, CRP, lactic acid, and serum calcium were deemed to produce the greatest effect on mortality based on their Shapley values. However, a similar study, which also seeks to predict COVID-19 mortality using machine learning, reports slightly different results (Moulaei et al.) The results of Moulaei et al.'s algorithm indicate that the most important predictors are dyspnea (shortness of breath), ICU admission, oxygen therapy, age, and fever. This paper did consider some of the same (or similar) predictors as Booth et al.'s algorithm, including CRP, BUN, albumin calcium (which effectively corrects patients' serum calcium results for serum albumin levels, interweaving the two variables used by Booth et al.'s model into one variable), and lactate dehydrogenase (which measures similar bodily functions to lactic acid, which was considered in Booth et al.'s model).

Thus, this disparity is not a result of both papers considering completely different variables when training and evaluating the models. Rather, it indicates that there is still much work to be done to create a Machine Learning model that consistently and accurately predicts patient mortality, and determine which predictors most reliably contribute to this prediction. If two models cannot agree on the most optimal predictors, how can medical professionals use one particular model in good faith to predict patient mortality, and thereby differentially allocate resources? In machine learning models such as those used in the aforementioned papers, whose results may not be immediately interpretable, feature importance helps to shed some light on the "black box" of an algorithm and gain insight into how it generates predictions. However, if different models rely on different features, it is difficult to discern whether such models are appropriate to use in a clinical setting, where subject matter expertise and evidence-based decision making are critically important. Perhaps in the future it will be feasible and beneficial to implement a mortality-predicting algorithm to assist medical personnel in allocating critical resources more optimally. However, due to the gaps in existing literature and a lack of evidence asserting that these benefits exist, implementing such a model in its current state might be potentially unreliable and problematic. However, the implementation of this model may be problematic not only due to scientific concerns, but also concerns of a philosophical and moral nature.

Analysis of Normative Consideration

The algorithm discussed in this paper generates a pressing ethical dilemma: how predicting a patient's mortality could impact the quality of care they receive. The authors mention that implementing their algorithm could allow critical care supplies and staff to be allocated more effectively in hospitals, thereby increasing efficiency and quality of care. However, there

exists minimal literature to support this claim, as the use of Machine Learning algorithms within a clinical setting is a relatively new concept, and as a result there exist considerable reporting gaps in terms of its performance (Kolasa et al.). Much of the existing literature regarding implementation of Machine Learning models to streamline resource allocation discusses this application in a hypothetical sense: that predicting patient mortality *has the capacity* to improve allocation of resources and quality of care. However, there exists little to no literature which empirically confirms it can actually do so.

To clarify the moral soundness of this algorithm's implementation in light of this lack of empirical evidence, I will turn to a philosophical understanding of morality. Based upon principles of Deontology, the use of the authors' machine learning model is inherently immoral. According to philosopher Immanuel Kant, an act is only moral of it satisfies both of the following formulations of the categorical imperative. The first is that the act must be universalizable without encountering a logical contradiction. The second is that it must treat humans as ends (i.e. rational beings), rather than mere means (Sandel). Based upon these formulations of the categorical imperative, the implementation of the authors' machine learning model to disparately provide care in a clinical setting is inherently immoral. Namely, it violates Kant's second formulation of the categorical imperative.

The authors claim that utilizing their algorithm to predict mortality can result in "better" allocation of resources. However, this begs the question, what is meant by "better?" Is it divesting resources away from individuals who are predicted likely to survive, in favor of helping rescue those teetering on the brink of death? Or, is it channeling resources towards those most likely to survive, with the hopes that they will benefit more from receiving intensive care than patients who are predicted to pass anyway?

It turns out, no matter which of these lenses through which we view the author's intent, Kant's second formulation of the categorical imperative is still violated, thus rendering the application of this algorithm immoral. If a greater concentration of resources is provided to patients who are predicted to pass, individuals who are predicted to survive involuntarily relinquish resources that could have been used for their care. They are therefore treated as mere means to an end. Conversely, if resources are directed towards patients predicted to survive, the individuals who are predicted to pass away are treated as means to an end, as they are involuntarily deprived of resources that could have been used to preserve their lives. They are robbed of the chance to experience a medical miracle and survive against their unfavorable odds, which is always a serendipitous possibility. It is important to note that order to violate the second formulation of the categorical imperative, the individuals who involuntarily lose resources that could have been used for their care must be treated

as *mere* means to an end, meaning they are not benefitting from the algorithm's use. This is precisely the case, as directing resources towards some individuals *always* leads to the direction of resources away from other individuals, no matter which group of individuals we choose.

Conclusion and Final Remarks

In summary, the ability to predict mortality from COVID-19 has the capacity to hold immense power. However, it is important that with this power comes informed and ethical decision making. Perhaps in the future, a reliable machine learning model with consistent results verified by existing literature will be made available. This could change America's medical landscape forever, allowing the optimal and efficient allocation of resources to those who need them most. Or perhaps it may do more harm than good, disparately denying individuals who are predicted to pass away of resources that could have allowed them to experience a medical miracle. In its current state, using such a model could be disastrous, both for the scientific and philosophical reasons elucidated above. I maintain that the use of such a model in any state is unethical, due to its potential to divert medical resources away from nonconsenting individuals, thereby using them as mere means to an end. According to Vardeman and Morris, "The real contribution [of statistics] is primarily moral, not technical." No matter how flawless an algorithm may be on paper– no matter how high a sensitivity value it flaunts, or how low an error rate—the most important factor in one's decision to implement it is whether or not this implementation would be ethical. To you, my reader, I leave this decision. If you had the opportunity to glimpse into the future, would you take it?

References

Kolasa, Katarzyna et al. “Systematic reviews of machine learning in healthcare: a literature review.” Expert review of pharmacoeconomics & outcomes research vol. 24,1 (2024): 63-115. doi:10.1080/14737167.2023.2279107

Lopes, Luna, et al. “Americans’ Challenges with Health Care Costs.” KFF, 7 May 2024, www.kff.org/health-costs/issue-brief/americans-challenges-with-health-care-costs/#:~:text=The%20cost%20of%20health%20care,care%20because%20of%20the%20cost. Accessed 30 Nov. 2024.

Kaur, Harsurinder, et al. “A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions.” ACM Computing Surveys, vol. 52, no. 4, July 2020, pp. 1–36. DOI.org (Crossref), https://doi.org/10.1145/3343440.

Yap, Bee Wah, et al. “An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets.” Proceedings of the first international conference on advanced data and information engineering (DaEng-2013). Springer Singapore, 2014.

Moons, Karel G. M., et al. “Using the Outcome for Imputation of Missing Predictor Values Was Preferred.” Journal of Clinical Epidemiology, vol. 59, no. 10, Oct. 2006, pp. 1092–101. DOI.org (Crossref), https://doi.org/10.1016/j.jclinepi.2006.01.009.

von Kugelgen, Julius et al. “Simpson’s Paradox in COVID-19 Case Fatality Rates: A Mediation Analysis of Age-Related Causal Effects.” IEEE transactions on artificial intelligence vol. 2,1 18-27. 14 Apr. 2021, doi:10.1109/TAI.2021.3073088

Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z. et al. Comparing machine learning algorithms for predicting COVID-19 mortality. BMC Med Inform Decis Mak 22, 2 (2022). https://doi.org/10.1186/s12911-021-01742-0

Sandel, M. J. (2010). *Justice: What’s the right thing to do?* Farrar, Straus and Giroux.

Vardeman, S., Morris, M. (2002). Statistics and Ethics: Some Advice for Young Statisticians. General.