Discussion of 2021-1733

MACHINE LEARNING REFINEMENT OF THE NSQIP RISK CALCULATOR: WHO SURVIVES THE “HAIL MARY” CASE?

**DR HERBERT CHEN** (Birmingham, AL): Can you provide details on the 508 operations that were associated with the highest mortality that were used to create this model?

 You focused on patients with mortality greater than 50%, and I am wondering would the same model apply to patients with a mortality predicted less than 50%?

 Lastly, you began your study by putting forth the question about when to do the "Hail Mary" operation. After these new data, when would you do the "Hail Mary" operation? What would be an acceptable risk?

**DR REID B ADAMS** (Charlottesville, VA): The outcome of this work is highly dependent on the methods you chose to use, and the choices made as part of the analysis. There are a variety of ways to do machine learning models. Can you discuss with us the rationale for your decisions, for instance, how did you choose the threshold you used and how alternative choices might affect the outcomes in building your model?

 From a practical perspective, it appears that this model is built or best used for those patients in emergency situations, or an extremist considering the predicted variables such as septic shock, high INR, ventilator dependence, thrombocytopenia, and elevated creatinine. Is this a correct interpretation of the model's usefulness, or do you see the utility in the elective preoperative population with multiple comorbidities?

Finally, can you discuss the utility of this model in practical terms? How do you envision using this as we make decisions in these complex patients? Is this tool independent in the NSQIP calculator or in addition to it? For instance, other than the situation when all 5 factors are present, how do you use it to decide when to throw the "Hail Mary" or not?

**DR SAMIR FAKHRY** (Nashville, TN): It is always easier to predict an outcome that is more common than less common. How would your model function if you looked at people with a very low probability of death? That is a discrimination issue, and I am curious whether you can address it.

 My other comment was about LIME. I am more familiar with lime as an additive to beverages, so please forgive my ignorance. Do you select the variables that are put into use, like in a logistic regression, we select them? Or does the software run unlimited variables through and decide which ones apply? How do you select the candidate variables for LIME to introduce into the input model?

**DR AMY HILDRETH** (Winston‑Salem, NC): I think this is fantastic work and I really applaud the authors on exploring this difficult topic. I have sort of a philosophical question for you. In discussing these "Hail Mary" operations of patients and their families, do you think better, more accurate numbers that predict outcomes will make these conversations easier, and how do you plan to use these numbers in discussing gold concordant care with patients and their families?

**DR OMAIDA VELAZQUEZ** (Miami, FL): I want to thank the authors for working on this, because the question of surgical futile care is extremely difficult, completely unsolved and we, as surgeons, need more tools that are going to help us avoid futile surgical care.

 How do you intend to optimize this model? It would seem to me that avoiding a false positive is a lot less important than avoiding a false negative for survival. You do not want to declare someone as nonsurvivable in a wrong way.

**DR MICHAEL ROGERS** (Tampa, FL): Dr Chen, from the details of the 508 included cases, there was a wide range of procedures that were included in the NSQIP database and, by no surprise, included cases that we deemed, or that we often think of as "Hail Mary" cases. The top 5 contributing cases to our model were craniectomy for subdural hematomas, ruptured AAA repair, total proctocolectomy, SMA thrombectomy, and extremity amputation. The remainder of the procedures followed this line of these "Hail Mary" cases, which leads me to the next question. Does this model apply only to cases with 50% mortality? Our model was trained on 50% mortality as a threshold, but certainly the methodology of this paper can be applied to the NSQIP database as a whole. We did not run that analysis on the entire database, but instead selected on the 50% mortality. Certainly, the GBM modeling and LIME can be applied to these larger datasets with mortality less than 50%.

 In discussing the model with practical terms, if this is to be used independently of NSQIP or in addition to, we certainly think of this as an addition to the NSQIP database as it exists. The NSQIP Risk Calculator, as it exists now, gives probabilities of mortality and other outcome variables, whereas our model gives insight into the feature variables that contribute to that mortality. It looks at the contributors and observes if there are any contributors that may be alterable prior to undergoing surgical intervention. And if it felt not modifiable, we would have a discussion with the patient's family or the patients themselves prior to embarking on a sort of "Hail Mary" case. I certainly think individually, the decision to do "Hail Mary" cases is probably going to be dependent on surgeon individuality, as well as the healthcare setting the surgeon is practicing in. That is probably an individualized discussion, but we hope that this model can help provide insight into that decision‑making and help in that process.

How did we select models, the variables that go into LIME? The variables were selected by the NSQIP; variables that were included in the Risk Calculator. Certainly, as a further follow‑up to this study, other variables of interest, including surgeon‑specific factors, hospital factors, socioeconomic factors, and as we saw today, racial and ethnicity factors, could certainly be built into this model to better characterize the risk to the patients.

Finally, do these outcomes better or more accurately help with decision‑making for the individual patient or the family memory? I certainly think any additional information we can give to families can be helpful when it comes to navigating these difficult decisions, especially late at night in the ICU when we are thinking about whether to proceed with operative intervention or not. Clearly, our model should undergo further optimization. An AUC of 0.65, as Dr Kuo pointed out, while respectable, certainly could be improved upon. However, providing families with additional information when we are having these discussions, I believe is a good thing.