Discussion of 2021-1729

MACHINE LEARNING APPROACH TO STRATIFYING PROGNOSIS RELATIVE TO TUMOR BURDEN AFTER RESECTION OF COLORECTAL LIVER METASTATIS: AN INTERNATIONAL COHORT ANALYSIS

**DR MARIA MAJELLA DOYLE** (St Louis, MO): The authors present a novel machine learning approach to stratifying prognosis relative to tumor burden after resection of colorectal liver metastases. In this cohort of over 1300 patients, the authors compare the prognostic accuracy of a machine learning approach to other, more commonly used scoring systems, such as the Fong score and the Genetic and Morphological Evaluation (GAME) scores. One of the problems we see in some of the already-created scoring systems is the lack of diversity of the cohort and, as a result, the replicability of the results can be challenging.

In this multicenter study, using a large, diverse, international cohort gives weight to the results. The authors report that machine learning-based tumor burden classification (ML-TB) was significantly more effective at stratifying patients relative to 5‑year overall survival in patients with lower, average, and high tumor burdens, compared with the Fong and GAME scoring systems.

While I love the idea of prognostic scoring, I struggle to determine the day‑to‑day utility in clinical practice, since many prognosticating scoring systems occur after the resection. The horse has left the barn, so to speak. How do you see the scoring system being used on a patient‑to‑patient basis? Are you changing or tailoring your adjuvant therapy for your patients based on their score? Can the ML-TB score be determined pre‑resection, and, if so, is there more utility in that to tailor neoadjuvant therapy or guide resection plans?

Your scoring system only accounts for overall survival. Can you comment on disease‑free survival (for which was the original Fong score was created), and how this may affect the stratification of patients? How does the machine learning (ML) algorithm identify the optimal cutoff values? Can you tell us the ML method? There is no data showing that the number of hepatic lesions equals 3 in the largest lesion size greater than 1.3 cm is the best cutoff value in comparison to other cutoffs examined in the manuscript.

Finally, we use the Fong and ASO scores for assessing patients for transplantation for colorectal liver metastases. While liver transplantation is a relatively rare indication for transplant, and much is unknown when it comes to criteria for liver transplantation or colorectal liver metastases, given your results, what are your thoughts on using ML-TB as a pre‑transplant prognosticator?

**DR PERRY SHEN** (Winston‑Salem, NC): There is a great deal of interest in today's world in the potential applications of machine learning, and this certainly applies to healthcare. So, I want to go back to basics and ask, what is machine learning exactly? It is a method of data analysis that automates the analytical model-building and is a branch of artificial intelligence, based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention.

It is clear that machine learning in data analysis is here to stay. The potential applications for determining patient prognosis in various disease processes are enormous. Right now, we are in the early stages of adoption. Dr Pawlik, through his current and previous work, is certainly a pioneer in this field.

The challenge is understanding what machine learning means and how it works. The concept of systems learning on their own is highly attractive, but we still are a long way from being comfortable with the various algorithms by which machine learning operates. But this study, I think, moves us closer in that direction. Dr Pawlik and his co‑investigators have developed a machine learning algorithm to determine the prognosis of patients with colorectal liver metastases undergoing operative treatment based on their tumor burden, the visible size of the largest lesion, and the number of lesions. They used a training and validation cohort, then compared the outcomes with whether prognostic scoring systems demonstrated superior stratification of patients into low- and high‑risk prognostic groups.

My first question is regarding methodology. The authors state that the machine learning tumor algorithm was compared with the Fong and the GAME scores, 2 other prognostic scoring systems. As originally described, the Fong score is comprised of 5 preoperative factors: Carcinoembryonic antigen (CEA) level greater than 200, disease‑free interval less than 12 months, a node positive primary tumor, size greater than 5 cm, and more than 1 tumor.

In Dr Fong's original description of his clinical risk score, he stratifies his 5 factors into 3 groups: 0 to 2 factor = best prognosis, 3 to 4 factors = guarded prognosis, and all 5 factors = poor prognosis. In this study, only Fong score cutoffs tumor size and lesion number were used to construct 3 prognostic groups for comparison to the machine learning tumor algorithm. Similarly, the GAME scores is made up of 6 available preoperative predictors of worst overall survival based on multivariate analysis. KRAS-mutated tumors, CEA greater than 20, node-positive primary, a tumor burden score of 3 to 8, or a score of 9 or greater, and extrahepatic disease. Using a point system based on these 6 factors, the GAME score was assigned 3 categories: 0 to 1 = low‑risk, 2 to 3 = medium‑risk, and greater than 4 = high‑risk. The current study only uses tumor burden score for comparison with the ML-TB algorithm. If all factors as originally described had been included in this analysis, do you think the results would have been similar, and can you discuss your rationale for not including them?

The second question is about the Cohen's d between the low vs average 5‑year overall survival free‑scoring system. The ML algorithm demonstrates a greater stratification between the low vs high and even the high vs average in overall survival of 5 years. But the low vs average group stratification was highest for the GAME score. My interpretation of Cohen's d is, the higher the number, the greater the sample size for effect size between groups. Does this have any implications for prognostic accuracy of the machine learning tumor algorithm in the low-risk compared with the average/medium-risk group, especially since the greatest percentage of patients was in the medium-risk group?

**DR HENRY PITT** (Philadelphia, PA): The patients in this study came from 5 institutions, including international sites. The percentage of patients who received neoadjuvant chemotherapy was only 55%, which is low for the US. Thus, my questions are whether you analyzed the data by US vs other countries, and whether the proposed stratification system is generalizable to both the US and to other countries?

**DR JORDAN CLOYD** (Columbus, OH): Dr Doyle asked about the clinic utility of a prediction score like this, and we can think of several reasons a clinical risk score would be very helpful. First, it is incredibly important to be able to provide patients with accurate, clinically relevant prognostic information to aid in patient‑centered conversations and shared decision-making.

Secondly, recent studies on the role of adjuvant therapy after resection of colorectal liver metastasis continue to be controversial, and future studies are going to need to be risk‑stratified. Having an accurate and effective risk-prediction tool is going to be necessary. Third, the number of colorectal cancer survivors is increasing around the world, and the economic and clinical impact of frequent scans and labs is overwhelming. An opportunity to perform personalized, risk-stratified surveillance will need to be studied closely in future trials.

To the question about preoperative decision making, we point out that the role of neoadjuvant therapy for resectable colorectal liver metastasis continues to be controversial as well. And so, while the current machine learning-derived tumor board score was created using patients who underwent resection, the previous tumor burden score has previously been validated as a preoperative decision‑making tool. We would expect similar utility in making decisions about neoadjuvant therapy or other preoperative decisions. Regarding overall survival vs disease‑free survival, we designed this study to maximize the difference in 5‑year overall survival between our 2 prognostic groups. Additional research will be needed to validate it for disease‑free survival, but, again, previous studies on tumor burden scores have been shown to be clinically relevant for disease‑free survival. Regarding the methodology of the machine learning program used, the stochastic hill climbing method is an optimization algorithm that is basically used to determine cutoffs that maximize the difference in any point, in this case, overall survival and the nuts and bolts of the artificial intelligence. Beyond that, we would defer to our biostatisticians.

Finally, as to the utility of a tool like this in transplant oncology, liver transplantation is increasingly of interest to select patients with unresectable colorectal liver metastases, and given the ongoing shortage of allografts and the need to identify patients who are most likely to benefit from transplantation, having a strong clinically predictive tool can be very valuable. I think our comment would be that machine learning is a promising tool, especially when applied to large, diverse data sets which hold a lot of promise for carefully selecting patients.

We appreciate Dr Shen's comments on machine learning more broadly, and agree that we are only skimming the surface on its utility in surgical oncology. We also agree with Dr. Shen that tumor burden scores, specifically the number of lesions and tumor size, were only components of previous clinical risk scores, such as the Fong score and the GAME store. So, while we will continue to need to use traditional risk factors in decision-making, we hope that the current study helps refine the cutoffs used in deriving tumor burden scores.

I think Dr Shen makes an astute observation on Cohen's d, and it is specific to our study that it was designed to maximize the difference between the low- and the high‑risk cohorts. It may not be as sensitive for detecting differences between low- and average- or average- and high-risk, and that may be something that requires further follow‑up.

Finally, to address Dr Pitt's question about the differences in US vs international cohorts based on the percentage of neoadjuvant therapy received, we have not performed those kinds of subset analyses, but certainly can do them in the future.