Clinical Chemistry

An AI Model (LORIS) to Predict Immune Checkpoint Blockade Response in Cancer: A Clinical Data Science Perspective

Article:
Thomas E Tavolara, Wenchao Han, David S McClintock.
An AI Model (LORIS) to Predict Immune Checkpoint Blockade Response in Cancer: A Clinical Data Science Perspective
Clin Chem 2025; 71(3): 345–7. https://doi.org/10.1093/clinchem/hvae196

Guests: Drs. David McClintock and Thomas Tavolara from the Division of Computational Pathology and Informatics within the Department of Laboratory Medicine and Pathology at the Mayo Clinic in Rochester, MN.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett.

The development of immune checkpoint inhibitors has provided oncologists with a revolutionary new tool in cancer treatment, but the effectiveness of these drugs varies widely. Some patients respond well, while others show no response at all. Biomarkers like tumor mutational burden (TMB) and programmed death-ligand 1 (PD-L1) have been approved by the FDA for the prediction of immune checkpoint blockade response, but their performance is inconsistent. Clearly, immune checkpoint response is more complex than these two standalone markers, but is it possible to capture this complexity in a single diagnostic test?

A recent research article suggests that the answer is "yes." The study team developed a new machine learning model, LORIS, that incorporates several different data points and showed that it outperformed standalone markers that represent the current standard of care. A perspective article, appearing in the March 2025 issue of *Clinical Chemistry*, summarizes this new research study, showing LORIS's promise of improved performance relative to existing biomarkers and highlighting the need for laboratorian involvement while implementing clinical decision support tools based on new machine learning models.

In this podcast, we are joined by two of the article's authors. Dr. David McClintock is the Chair of the Division of Computational Pathology and Informatics within the Department of Laboratory Medicine and Pathology at the Mayo Clinic at Rochester, Minnesota. Dr. Thomas Tavolara is a Senior Associate Consultant in AI at the Division of Computational Pathology and AI Department of Laboratory Medicine and Pathology, and an Assistant Professor of Laboratory Medicine and Pathology at the Mayo Clinic.

So, Dr. Tavolara, let's start with you. The LORIS model integrates multiple clinical features, moving beyond single biomarkers like TMB and PD-L1. Given the inconsistencies of

these traditional biomarkers across cancer types, what advantages does LORIS offer in terms of predictive power and clinical utility? And how does it compare with other multi-biomarker approaches?

Thomas Tavolara: Yeah. So, you're right. LORIS was designed with the intent to overcome the inconsistency of traditional biomarkers, specifically TMB and PD-L1. It works by leveraging multiple commonly available clinical features, at least the ones that were common in their data sets. So, what this allows for is a high predictive power across a broad range of clinical tasks.

So, even though LORIS is trained to predict immunotherapy response--I'm probably going to start calling it "ICB [immune checkpoint blockade] response" throughout the rest of the podcast--that response probability number, it's actually a number that the model produces, can actually do a lot more that you couldn't normally do with other traditional biomarkers. So, for example, they have some results showing that it can stratify overall survival. It can also stratify progression-free survival. Again, that's something that single biomarkers can't do.

In fact, probably the strongest aspect of the predictive power of LORIS comes from, I think it's their Figure 3 and in our perspective it's the only figure we have, shows that LORIS can identify a subset of patients that would traditionally be marked for high likelihood for ICB response that by LORIS's estimation would actually not benefit. Vice versa, LORIS can identify a subset of patients that would traditionally be marked for a low likelihood for ICB response that by LORIS's estimation would actually benefit. So, here again traditional meeting the mutational burden, a single biomarker.

As far as how this compares to other multi-biomarker approaches, I think that was the second part of your question. I'm not really well-versed in the ICB response prediction literature, but the authors of LORIS showed that their method achieves a higher AUC, other metrics as well, not just AUC, than previously published multi-biomarker approaches across I think like five different external data sets. This is pretty par for the course for machine learning papers. They tend to always show that their method is better, so that's not too surprising. But it's better than the previous multi-biomarker method.

Bob Barrett: While LORIS has demonstrated strong performance retrospectively, what are the key hurdles for its adoption in clinical practice? And how do issues like data harmonization, standardization across different labs, and regulatory considerations impact the model's reliability and scalability?

Thomas Tavolara: So, definitely the biggest challenge would be prospective validation, which the authors actually mentioned as a limitation of their study in their discussion section. I think it was in the last paragraph. The reason this prospective validation is important is because retrospective studies tend to suffer from biases in data collection, patient selection, and treatment variability. That's one of the reasons why we'll say the gold standard for any novel test is a clinical trial, which follows patients prospectively rather than seeing how -- well, I guess if you're doing a drug trial you can't do retrospective, but in machine learning you technically could. But anyway, the gold standard is prospective.

Moreover, the true test for LORIS would actually be to see whether it's modifies physician's decisions or recommendations, like whether they actually use it, then to see whether those patients end up doing better. Without that, and LORIS just becomes like an interesting tool.

And let me tell you, I've been in this machine learning space in medicine for a while and there are a lot of interesting machine learning tools out there, but they never end up actually being used.

You also mentioned the data harmonization and standardization, how that impacts the model's reliability and scalability. I'm not so sure about scalability. I think that's somewhat independent of the harmonization and standardization, but those aspects would definitely affect reliability.

So, for example, one of the features that LORIS utilizes in its predictions is this tumor mutational burden and I know that different sequencing platforms measure that differently. So, some use whole exome sequencing and some use a panel-based measure, so, like a subset of exome sequencing. I'm not exactly sure whether the data LORIS was validated on has those variations in this example of tumor mutational burden. But I'm for sure the other stuff that their model uses could suffer from the same sorts of issues.

And I'm sure your listeners would definitely be a better judge of whether those features would be affected by harmonization and standardization because I'm not a laboratorian. I think Dave is going to take a crack at those regulatory considerations.

David McClintock: Yeah. I think from my perspective, there's actually some both practical and regulatory perspectives to take into mind when looking at a model like this.

The first thing is that as designed, LORIS really is best designated as a calculator. It's taking actually data from the

EHR that you can manually extract, and essentially being input by the clinician and run when needed. This really allows it to be considered a clinical assistance support tool.

So, per current FDA guidelines, this means it would be a non-medical device and it's much more easily deployed by clinical practice then, in this case, because you don't need to go through a lot of additional regulatory burden to get it into your practice. In this case for sure, it could literally just be a bookmark in a web browser because there's an online tool for this. The clinician goes in, inputs their labs and then gets their likelihood scores out.

However, this does limit the tool scalability as it requires a clinician or clinical service to actually go out and input that data, like I said, manually into the calculator for each patient. You might be at risk of copy paste errors. You might have the wrong data from the patient written on a piece of paper or looking at a different screen. You could assign the wrong score to a patient once you have that score back.

So, there's a lot of these different issues that come up that ideally you would be handling in a much more automated fashion. So, those are some of the issues that when you look at the actual scalability of this, it's really hard to scale up a really manual process like that.

There's also an issue of reliability here. If the clinical practice starts using this tool regularly in the practice and the site goes down, or the model changes and they're not aware of those changes, the group decides to take it back for whatever reason. This can actually impact their practice as well in the subsequent clinical care of the patients, because the clinicians have begun to rely on this, and it's not a tool that they actively manage. So, all those things kind of come into consideration beyond what Thomas talked about from the actual data harmonization perspective.

Bob Barrett: Doctor, the paper focuses on predicting immune checkpoint blockade response, but many of its findings relate to overall survival and progression-free survival. Would a direct prediction of clinical outcomes be a more powerful approach? And how might this shift improve the utility of machine learning models in oncology?

Thomas Tavolara: Yeah. So, a direct prediction of overall survival or progression-free survival could be more valuable than predicting ICB response. This is actually a point that we harp on in the perspective.

There are a lot of studies out there that train models to predict what I call "clinical intermediates" rather than the thing that's most important, which is usually outcome. So,

even though those intermediates are important for physicians and patients in making informed decisions, for machine learning models, it just is noise.

Those intermediates aren't necessarily 100% predictive of outcome. So, even if you get a machine learning model to perfectly predict an intermediate, it'll still fail to exceed that intermediate's predictive performance on outcome.

Now, there are good reasons why researchers use these intermediates. Primarily, the reason is outcome data is really hard to get. When you can get it, it's expensive, and sometimes outcome is just hard to define. Even those studies that try to predict outcome directly struggle to achieve a high accuracy.

So, there's actually an argument to be made, I'm arguing against myself now here, but there's actually an argument to be made about leveraging what is already known about a certain disease, meaning these clinical intermediates, over having machine learning model to implicitly figure out these intermediates directly from the data. Maybe these intermediates are kind of like -- I kind of think them as like a hand holding for machine learning models.

There's an argument to be made for both sides predicting outcome directly or predicting these clinical intermediates. And I think you also asked about how that might impact machine learning models in oncology. So, if we shifted machine learning models to predict outcome directly, I would hope it would make them more accurate, but I'm not exactly sure if they would.

Bob Barrett: Dr. McClintock, let's go to you here. LORIS is currently positioned as a clinical decision support tool available via web-based calculator. What steps would be necessary to integrate such a tool directly into clinical workflows? And what role should clinical laboratorians, pathologists, and data scientists play in ensuring AI models like LORIS are effectively implemented?

David McClintock: So, as I mentioned before, LORIS really exists primarily as an unregulated clinical decision support tool for clinicians with that web-based calculator existing online. Again, clinicians can go to this website, bookmark it, and put their patient's data into this tool. Again, de-identified data there.

To get an immunotherapy response score, that would be in the form of a percent likelihood that the patient would respond to immune checkpoint blockade therapy, and this would then help guide treatment for their patients. So, the clinician would get the score and potentially decide, "Do I want to use this or not" on their patients themselves.

This information would be provided strictly as an additional tool for the clinician, not intended for use as a prior decision-making tool. Thus, making it pretty much just this little bit of extra added information and not something that would qualify as a medical device under current FDA guidelines. Again, this tool's only useful when the clinician goes out and pulls that information back to them.

The clinical group wanted to automate this process. So, for example, they wanted to have every one of their patients get a LORIS score that qualified for it. You could potentially imagine a scenario where the original group who created the model could offer up an API, an application programming interface, that would allow you to send de-identified patient data to the calculator and then bring back the actual LORIS score to the practice that sent that data to it.

When that would happen, you actually would have some additional regulatory concerns coming up, specifically around how a clinician would use the score and if they would still retain full autonomy over its use. That's where the key point here, is that if a clinician doesn't really have the ability to fully understand how the score is being generated and how they can use it in their own practice, if it just shows up and they just blindly follow it, that's when the FDA gets really much more picky about how do we consider this as a medical device or not?

Following best practices around clinical system support tools, this tool would have to obviously be vigorously validated to ensure the scores were accurate and precise by the practice, especially as you automate it. In addition to making sure that any changes to their own laboratory practice that is generating these lab values that are going to be used for the score would not cause the model to drift, such as a change in instrumentation or reagents, or other lab changes that might change the lab results themselves.

So, overall, if a clinical group really wanted to go ahead and bring this tool into practice, they really should be engaging the labs early on, especially the lab directors, pathologists, data scientists that might exist in the lab to go ahead and understand how changes in the lab might affect these models in the future.

This also brings up the issue of how the laboratorians see these models. A lot of times these are seen as completely separate tools that they don't really have a role in. And this really requires our lab directors, pathologists, and laboratorians to be upskilled in how AI models work, how the different input data that we have can affect these types of models, and how awareness needs to be raised for getting

the actual laboratorians involved in model deployment at an institutional level early on in the process.

Bob Barrett: Well, finally, doctors, as machine learning models become more integrated into cancer diagnostics, what are the next steps for improving predictive algorithms like LORIS? Are there ethical or practical concerns in relying on AI for critical treatment decisions, particularly given issues of interpretability and data bias?

Thomas Tavolara: Yeah. So, I already mentioned the idea of predicting outcome directly. So, that's one way we could improve predictive algorithms like LORIS in the future.

Other than that, the authors mentioned transcriptomic and immune microenvironment data, which they say have been shown to play a significant role in ICB response, but were not included in their model due to the data availability constraints. I'm pretty sure that data was available for their training cohort, but none of their testing cohorts.

Other than that, I'm not sure of any other meaningful features you could add to improve the model. Again, not my area of expertise. I'm sure the laboratorians know way more about that. Another thing you could do is to make a cancer specific model and the authors actually do that. They show that when they train LORIS on just non-small cell lung cancer, it works better than the standard, which I think in the case of that lung cancer was PD-L1 and tumor mutational burden. Not only that, we'll say the lung cancer specific model was better than the LORIS general model for all cancers.

So, again, you could do the same thing for other organs. But I think the authors also mentioned that the reason they didn't do that is because their data was limited.

I think they might have shown some examples in their supplementary data for liver and for endometrium, but again, small cohort LORIS didn't perform as well.

Other things you could do to improve the predictive power of models like LORIS, you could leverage something like federated learning. A lot of papers will throw federated learning out there as a solution or to improve predictive power. And so, the idea behind that is it allows hospitals to benefit from one another's data when training machine learning models, without actually having to share any of that data directly.

It's a pretty cool technique. It's been around for a while and there's lots of studies that show that it can be just as good as having all the data in one place at the same time. So, yeah, another way we could improve these sorts of models. Maybe

Dave can provide some perspective on the ethical or practical concerns for relying on AI in treatment decisions.

David McClintock: I think overall we should all remember that when you treat a patient, this is typically something that you don't use a single piece of data to go ahead and make a critical treatment decision.

With these kind of AI tools, the reason the FDA has taken a harder stance on making sure that they don't become medical devices when used strictly as clinically decision support tools is that with that reliance on a piece of data like this or on a tool like this, it is too easy for someone just to quickly go through and make a decision without using all the available clinical data from, say, a multidisciplinary team and stuff like that.

So, fundamentally, the real practical concern here and the real ethical concern is making sure that this just becomes, again, one more piece of data that you use to help with the treatment decision. It may or may not be something that skews clinicians to go one way or the other when assigning treatment. But the whole point is that it's taken as one piece of a larger whole that allows you to go ahead and make that clinical decision for that patient.

Bob Barrett: That was Dr. David McClintock and Dr. Thomas Tavolara from the Mayo Clinic in Rochester, Minnesota. They authored a perspective article in the March 2025 issue of *Clinical Chemistry* describing a new machine learning model to predict immune checkpoint blockade response. I'm Bob Barrett. Thanks for listening.