Clinical Chemistry

Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group

Article:
Stephen R Master, Tony C Badrick, Andreas Bietenbeck, and Shannon Haymond. *Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group.*
Clin Chem 2023; 69(7): 690–8. https://doi.org/10.1093/clinchem/hvad055

Guests: Dr. Shannon Haymond, current President of AACC, from the Ann & Robert H. Lurie Children's Hospital of Chicago, in Illinois, and Dr. Stephen Master from the Children's Hospital of Philadelphia, in Pennsylvania.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the American Association for Clinical Chemistry. I'm Bob Barrett. For many conditions, disease diagnosis, treatment selection, or outcome prediction can be performed using only a handful of clinical laboratory tests. In other scenarios, this approach provides insufficient information and endpoints are best determined by using machine learning to distill large, complex datasets into a predictive algorithm.

Machine learning has been successfully applied to address certain diagnostic questions, but it is not a magic wand, and there is potential for misuse. Errors in experimental design, bias, and specimen selection or annotation, or general failure to follow best practices may result in an ineffective or unreliable tool that provides incorrect or misleading information. As clinical laboratorians become increasingly interested in machine learning, what are the strategies for success and what pitfalls should be avoided? A new special report appearing in the July 2023 issue of *Clinical Chemistry* addresses these questions by summarizing consensus recommendations by the IFCC machine learning working group.

In this podcast, we're excited to talk with two of the article's authors. Dr. Shannon Haymond is currently the President of AACC. She is also Vice Chair for Computational Pathology and Director for Clinical Mass Spectrometry at Ann & Robert H. Lurie Children's Hospital of Chicago. Dr. Stephen Master is Chief of the Division of Laboratory Medicine at the Children's Hospital of Philadelphia, where he also holds a secondary appointment in the Division of Pathology Informatics. So, Dr. Master, let's start with you. Why did you feel there was a need for this report?

Stephen Master: Well, I think the first thing to say is that machine learning continues to become more and more accessible for many people in laboratory medicine. We're seeing more and more publications and potential applications in the field. So, we and others have previously published some general guidance for labs in this area. But I guess in light of everything I said, the time seemed right to convene a larger and more

internationally representative group to produce an expanded consensus document under the auspices of IFCC. So our hope, I think, in this is that by having a more formal document, that will provide not only clear guidance for developers and authors, but also an increasing awareness of the importance of the best practices that we're advocating.

Bob Barrett: The first table of your group's report lists scenarios in which machine learning should and should not be used. Dr. Haymond, why do you think it was important to include this at the start?

Shannon Haymond: That's a great question. With all the excitement around machine learning, as Steve just alluded to, it seems like there's a rush to apply this to every problem, whether or not that really makes sense. And in our working group, we were all concerned by this trend that we've observed individually, as we've reviewed abstracts or been asked to do peer review for manuscripts in the literature. And so, we wanted to start by reminding the reader that machine learning does have a lot of potential, but that we need to carefully consider whether it's the right solution for a given scenario.

Bob Barrett: Did this potential for misuse drive your working group to publish these recommendations?

Stephen Master: Yeah, absolutely. Because even though the tools are more widely available, there are still some really important considerations in how those tools are used. I would actually say that it can be fairly easy for any of us to inadvertently do something in a very complex workflow that can give an answer that's too good to be true. So it might look like good performance initially, but it might lead to something that isn't generalizable, and then therefore, not generally useful for patients, which is, of course, the goal. So once again, we wanted to create a document that provides guidance on how to avoid those specific pitfalls.

Bob Barrett: So who's the intended audience for this report, and how might readers use all this information?

Shannon Haymond: Well, we're not the first people to write about this. But we did draft this report for laboratory medicine professionals and scientists, really those who want to better understand the best practices for machine learning development and validation. And that's whether they want to be involved in doing those activities themselves, or in the peer review, or evaluation of applications that use machine learning. And so, this is why the report includes not only a description of each step of the process, and then we have recommendations that are associated with that step, but we also included a summarization of two examples from the literature and tried to demonstrate how those reports line up against the best

practices. So readers can use it as a tutorial, or again, as an example of what you should do in each of these steps, and what are some of the potential pitfalls.

Bob Barrett: Well finally, taking a look at the report, are there any recommendations either of you would like to highlight?

Stephen Master: Well, yeah, one thing I'd like to highlight actually, is the section on ethical machine learning. This is definitely becoming a more prominent theme in the last couple of years as we've seen examples of how machine learning can go wrong and even exacerbate inequities. One simple example you could think about in this regard is that if you add a biological characteristic of a population, or an apparently biological characteristic of a population, it was being used to predict the outcome.

But if that characteristic was also correlated with, say, access to healthcare, then in your training dataset, it might look as if that group of patients has a worse outcome because of who they are rather than because of their existing access to healthcare. And if you use that as the basis of building your machine learning model, you might actually predict that those patients will not benefit from intervention because we know they're not going to do well, and you might increase an inequity that already exists. So I think that's an important thing that we've tried to emphasize.

Bob Barrett: Dr. Haymond, anything to add?

Shannon Haymond: Yeah, I would add -- I really liked the sections that we talk about data leakage, where we really try to point that out as a pitfall. In addition, we've included recommendations about utilizing metrics beyond accuracy through the area under the receiver-operator curve, which is really widely used—it's familiar to us into the machine learning field--but can be problematic in healthcare, particularly when we have imbalanced datasets. And so, I feel by raising awareness about using things beyond just AUC ROC, that this can help the field really assess the impact of automated decisions through machine learning applications on clinical scenarios, and help us better address the effects of imbalanced datasets on these performance metrics.

Bob Barrett: That was Dr. Shannon Haymond from Lurie Children's Hospital of Chicago and Dr. Stephen Master from the Children's Hospital of Philadelphia. They published a summary of the IFCC's recommendations for the use of machine learning and laboratory medicine in the July 2023 issue of *Clinical Chemistry*, and they've been our guests for this podcast on that topic. I'm Bob Barrett. Thanks for listening.