Clinical Chemistry

How Can We Ensure Reproducibility and Clinical Translation of Machine Learning Applications in Laboratory Medicine?

Opinion:
Shannon Haymond and Stephen R Master.
How Can We Ensure Reproducibility and Clinical Translation of Machine Learning Applications in Laboratory Medicine?
Clin Chem 2022;68:392–5 https://doi.org/10.1093/clinchem/hvab272

Guests: Dr. Shannon Haymond is the Director of Clinical Chemistry and Mass Spectrometry Laboratories at the Ann & Robert H. Lurie Children's Hospital of Chicago, and Assistant Professor of Pathology at Northwestern University Feinberg School of Medicine. Dr. Stephen Master is Chief of the Division of Laboratory Medicine at the Children's Hospital of Philadelphia and an Associate Professor of Pathology and Laboratory Medicine at the Perelman School of Medicine at the University of Pennsylvania.

Bob Barrett: This is a podcast from *Clinical Chemistry*, sponsored by the Department of Laboratory Medicine at Boston Children's Hospital. I'm Bob Barrett. Recent studies demonstrating problems with prediction models for COVID and sepsis have helped raise awareness about the need for better practices in developing and reporting machine learning or artificial intelligence methods in healthcare. This so-called reproducibility crisis has been recognized and extends beyond clinical applications. In fact, this issue was not specific to ML. Many scientific journals, including *Clinical Chemistry*, have adopted principles specified in the submission guidelines to facilitate reproducibility, rigor, and transparency in published findings.

Though there are common elements that support transparency and rigor in science, each technology has its own set of pitfalls that must be addressed. This is particularly true for the rapidly developing field of machine learning. An opinion piece on the importance of best practices for development and reporting the machine learning appears in the March 2022 issue of *Clinical Chemistry*. It outlines how investigators should develop, validate, and report machine learning-based methods in laboratory medicine, and also provides guidance for all journal editors and reviewers when evaluating submitted papers. We are pleased to have the two authors of that article with us in this podcast.

Dr. Shannon Haymond is the Director of Clinical Chemistry and Mass Spectrometry Laboratories at the Ann & Robert H. Lurie Children's Hospital of Chicago. She is also Assistant Professor of Pathology at Northwestern University Feinberg School of Medicine. Dr. Stephen Master is Chief of the Division of Laboratory Medicine at the Children's Hospital of Philadelphia and is an Associate Professor of Pathology and Laboratory Medicine at the Perelman School of Medicine at the University of Pennsylvania. And Dr. Master, let's start with you. What prompted you to write this opinion piece? Why is this topic important and why now?

Stephen Master: Well, there are really two converging trends that we think make this topic important and timely. The first is that software tools to perform machine learning have become much more widely available and accessible. Just by way of background, machine learning, which is part of the field of artificial intelligence, refers to the development of computer algorithms that can predict something based on complex input data. So, for example, we might want to predict whether a patient has a disease based on the results of all their laboratory tests taken together. And typically in an example like that, we would start with laboratory data from a large number of patients where we know whether they have the disease and then train the machine to recognize how to use data to make a prediction where it doesn't know the answer.

So, one trend is that the tools to do this kind of work have become accessible to more investigators. And then the second trend is that there's been an increasing recognition of the importance and potential of machine learning for laboratory medicine. Clinical laboratories always generated lots of data, obviously, and machine learning provides a new way to more directly use this information to diagnose diseases and aid in operational decisions in a way that we couldn't before. Now the downside to this convergence is that in this rush to develop and apply machine learning methods, there have been some missteps. So, you can look for example at recent reports describing significant problems with clinical predictive models for COVID-19 and sepsis for example. So, it's clear that we need to be discussing pitfalls and best practices for developing and validating predictive models in healthcare.

Bob Barrett: Dr. Haymond, your piece is about ensuring reproducibility and clinical translational in machine learning for folks like me. What exactly does that mean?

Shannon Haymond: Well, we mean that we want to make sure the results are reliable and can be reproduced by other groups as we do with any method we develop for use in clinical laboratories and medicine. Specifically, for the reliability piece, it's first necessary to guarantee that there weren't any errors or biases made during the machine learning development and validation process that would provide a misleading prediction for new or unknown cases. And this can be especially problematic when it leads to overestimating the performance of an algorithm, which is often the case for these common issues. That's why in the article, we discussed several known pitfalls that should be avoided or properly mitigated.

One of the ways we can check for this is to have other groups and investigators recheck the process to see if they get the

same answer as the original authors. This is particularly important when you think about the complex and sometimes subtle details that can have a major influence on the accuracy and generalizability of an algorithm. And so, just like with wet lab methods, in order to reproduce the machine learning results of another group, you really need to know exactly what they did and how they did it, and that requires a very detailed, specific description of their methods and the data. That detailed description, then, along with the data can give us clues as to whether the machine learning algorithm is likely to work well in a new population. By that, I mean, another group of patients and another medical center, and that's really central to whether this model will be what we say is clinically translatable.

Meaning, will it be able to be successfully used in medical diagnosis or decision-making? I would also add that clinically translatable refers to the feasibility and suitability of implementing an algorithm in clinical practice. You know, we realized that there's a large gap in machine learning applications that are published versus those that are actually implemented today. In the article, we discussed the importance of things like model formulation, selection, and explainability or interpretability. And where most people focus on the predictive performance, which is obviously an important result, we feel these other factors are also key when considering the clinical utility and feasibility of a proposed machine learning application.

Bob Barrett: So who is the intended audience for this piece and what are the key points that you hope to get across?

Shannon Haymond: That's a really good question, Bob. So, our intended audience in this case is first and foremost groups of medical researchers who are beginning to develop prediction algorithms for medical diagnosis, and particularly those who are using laboratory medicine data, which is quite a lot of folks. And we list a set of specific things that we and other people in the field think need to be carefully considered and this ranges from how the data are collected through how the machine learning model itself is chosen and implemented. We even discuss how best to interpret the factors that drive the prediction. We hope that this summary provides a useful checklist for those in the field of laboratory medicine, who are beginning to utilize machine learning approaches.

Bob Barrett: So, Dr. Master, you discussed the limitations of the traditional scientific publication process and truly allowing for critical evaluation of machine learning based publications, and ultimately you make the case for authors to submit code and data use for the analysis. Why does this support reproducibility and clinical translation of the published

methods? And why is this not something commonly done by authors today?

Stephen Master: Well, we certainly aren't the first ones to say this by any means, and in fact, a number of statisticians and researchers over probably the past 10 or 15 years now have noted that the complexity of the data sets and of the computational methods that are now being used in machine learning approaches have led to a real problem in the literature. So, as Dr. Haymond just mentioned, there are subtle problems that can sometimes creep in when you're building these applications and the processing details are very, very important. The fundamental problem is, if I try to summarize the way I created a model in a paragraph or two in the methods section of a paper, there's no way I can possibly mention all of the details, software parameters, and other things that would be necessary for someone to really check my work.

So, ultimately, we argue that the best way to do this is to give someone else your actual computer code and your data, so that they can not only run for themselves and check your final answer, but also then unambiguously note exactly what you did, and you know, if necessary, critique any decisions that might cause a problem. I should say that one of the traditional challenges to doing this has been that there are many different software packages and computer languages in which machine learning models can be created. And you might think it's probably more trouble than it's worth to try and manage all these possibilities. Unfortunately, over the last, I would say, decade or so, there's been an increase in consolidation though to a smaller number of languages and packages that are being used for most of this work. And also, there are more people in medicine who are learning computer programming. So, understanding someone else's code has gotten comparatively easier, and the challenge is just holding authors responsible for providing these details. And I'm very happy to see that more and more journals are doing this.

Bob Barrett: Finally, for both of you. What are the roles of editorial boards and manuscript reviewers in ensuring reproducibility in clinical translation of machine learning models for laboratory medicine? Dr. Haymond, let's start with you.

Shannon Haymond: Sure. Ultimately, it's the editors and peer reviewers who really are serving as the quality check for what makes its way into the medical literature. We think it's important that these groups especially are aware of best practices in machine learning. And in the case of editors, are able to draw on a pool of peer reviewers, who have the experience in rigorously evaluating machine learning methods in the ways that they're applied to medical diagnosis.

Stephen Master: Yeah, I completely agree. Beyond wanting to simply help the authors of papers, we really want to use this opinion piece to provide a convenient resource that editors and reviewers can use to identify high quality studies, and ultimately this serves all of us. It helps the editors because they're able to publish reliable information, helps the reviewers because they can point to specific recommendations, they can use to hold authors to a high standard, helps the authors because I think they can have more confidence in the results. And then ultimately, of course, it helps patients because it ensures that we as a laboratory community are using new advances in medical diagnostics in the best and most reliable way possible.

Bob Barrett: That was Dr. Stephen Master from the Children's Hospital of Philadelphia and the Perelman School of Medicine of the University of Pennsylvania. He was joined by Dr. Shannon Haymond from the Ann & Robert H. Lurie Children's Hospital and the Northwestern University Feinberg School of Medicine in Chicago. Their opinion piece on the importance of best practices for development and reporting of machine learning appears in the March 2022 issue of *Clinical Chemistry*. I am Bob Barrett.