Clinical Chemistry

A Machine Learning Model for the Routine Detection of "Wrong Blood in Complete Blood Count Tube" Errors

Article:
Christopher-John Farrell, Charles Makuni, Aaron Keenan, Ellena Maeder, Gareth Davies, and John Giannoutsos.
A Machine Learning Model for the Routine Detection of "Wrong Blood in Complete Blood Count Tube" Errors
Clin Chem 2023; 69(9): 1031–7. https://doi.org/10.1093/clinchem/hvad100

Guest: Dr Chris Farrell is the Clinical Director of Chemical Pathology at Liverpool Hospital in Sydney, Australia.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett. Clinical laboratories have many tools at their disposal to detect analytical errors or mistakes that occur during the process of sample testing. But there are limited tools to detect errors during the pre-analytical phase before testing begins. This is unfortunate because pre-analytical errors, such as placing a label with Patient A's, information on a tube containing blood from Patient B, are the most common causes of misleading test results. Existing tools can help identify some "wrong blood in tube" errors, but they only catch a small fraction of these mislabeled specimens. "Wrong blood in tube" errors that go undetected have the very real potential to cause patient harm, ranging from the selection of wrong drug dose to administration of incompatible blood products that lead to severe transfusion reactions. To address this problem, previous work has evaluated the potential of machine learning. But these initial models do not work in samples with missing parameters and generate a high number of false positive results.

A new article appearing in the September 2023 issue of *Clinical Chemistry* proposes a slightly different approach for the detection of wrong blood and tube errors, with improved results relative to previous work in this area. In this podcast, we're excited to talk with the article's lead author. Dr. Chris Farrell is the Clinical Ddirector of Chemical Pathology at Liverpool Hospital in Sydney, Australia. Well, Dr. Farrell, let's get right to it. Why is machine learning a good tool for detecting wrong blood and tube errors?

Chris Farrell: Can I just say thanks for having me on, Bob. I'm really thrilled to be speaking with you.

Bob Barrett: Well, thanks. Right back at you.

Chris Farrell: I think the first thing to say about "wrong blood in tube" really is that it is a significant issue. I hadn't appreciated before I started this work, just how much angst they cause in departments outside of chemistry. And I came across one audit, from the US actually, that looked at laboratory related

patient safety events and they found "wrong blood in tube" was responsible for a third of those. So it does seem to be something that we should be trying to do better on. Now for why machine learning might be good here, I think the first thing to say there is that it's the kind of problem that machine learning tends to do well on. Particularly the modern algorithms can take into account multiple input parameters and learn their relationship to some outcome that you want to predict. And they can do that even with highly complex relationships between input and outcome. And they could also take into account multiple interrelationships between those input parameters themselves. So that's exactly the scenario we have with "wrong blood in tube."

We're going to be asking a model to look through changes across multiple analytes within a panel and interpret those changes in the context of the patient's age and sex to determine if "wrong blood in tube" is likely. So I think it's the right kind of problem. But then really a big advantage of working on "wrong blood in tube" specifically is the ability that we have to simulate these errors on the computer. So what we do for that is we just take download of genuine patient results and we extract those, and then we can get sets of current and previous results for individual patients. Then when we want the computer to simulate a "wrong blood in tube" error, we can just get the computer to throw out the genuine current set of results for the patient and plug in results from a random other patient.

So in this way it's quite easy to generate data sets with tens of thousands of examples of "wrong blood in tube." And when you think about what lab staff are exposed to, if they come across a handful of "wrong blood in tube" errors in an entire year, they're doing well. So if we train up a model with tens of thousands of "wrong blood in tube" error examples, then it's in a sense equivalent to giving the model thousands of years-worth of human experience. So there is the potential there to generate very powerful models. And then the last thing that we can do in the "wrong blood in tube" topic is do studies on data sets that contain these simulated errors. And these have been really valuable in helping us to get an idea of what might work best in practice, really. So I guess firstly those studies showed that machine learning models could outperform delta checks, just the standard delta checks for detecting wrong blood in tube.

More recently, what's been exciting is that we've seen data come out showing the machine learning models outperforming lab staff both in hematology and biochemistry departments. And I guess what I've gotten fascinated in is that interplay between the model and staff and where the machine learning model might fit within a laboratory workflow. So I guess the initial idea there was that you could

send a model prediction to staff for them to take on board that prediction, but then review the results themselves and add to the model's prediction their own insight and knowledge of physiology, and then come to a better prediction than either they could alone or the model could by itself. This is something we could test. And when we did that, it actually turned out that despite being a good idea in theory, the machine learning models were actually so far ahead of human level performance that this addition of human interaction didn't add to the model's performance at all. It actually degraded it back down towards baseline human performance.

And so, these kind of studies have been really valuable in helping us to work out what algorithms we should be using, what panels we can apply them to, because they tend to do best on CBC results than general biochemistry results. And then it's probably best to allow these models to operate autonomously in flagging results for further investigation. So these simulation studies have really been a big head start for us when we started to think about detecting "wrong blood in tube" errors in practice.

Bob Barrett: Well, I'm interested in what were the challenges you encountered when moving from these simulation studies and trying to detect "wrong blood in tube" errors in the real world?

Chris Farrell: I guess the main thing there was error prevalence. So on the simulation studies, we were using an error prevalence of 50%. And that was great for model training because it was a balanced data set and you could show the model lots of examples of "wrong blood in tube." But when it comes to the real world, "wrong blood in tube" error prevalence is something more like 0.03%. So that really has a big effect on model performance in terms of positive predictive value. We calculated that for the best performing model from those simulation studies, which had sensitivity and specificity up at 99%, but if that was deployed in the real world, we would expect positive predictive value of only 3%. So 97% of times it was flagging a result as "wrong blood in tube" or likely "wrong blood in tube," that it'd be a false positive flag. So we needed to do something to improve positive predictive value to get a useful model for practice.

There are a range of things we could have done. We just took the simplest approach, which was essentially a threshold adjustment. We were working with an extreme gradient boosting model applied to CBC results. We just had the model output, its predicted probability of "wrong blood in tube" for each set of samples that it reviewed. And then what we could do is at the end of each week, just flag for further investigation those samples with the very highest probabilities assigned by the model. So we took the five samples with the highest probability each week and we sent

those for further investigation. So that was consisting of extended red cell phenotyping and blood group typing of the current and previous samples for that patient. But this threshold adjustment approach actually worked very well. We picked up "wrong blood in tube" errors at six times the rate of standard procedures and we achieved a positive predictive value of over 10%. The drawback of this kind of workflow, it was useful for the study, but clearly in practice, we don't want to be identifying all our "wrong blood in tube" errors at the end of the week after we've reported them. So we needed really to define a threshold for that probability, above which we should be flagging results as likely "wrong blood in tubes" just so we could have real time decision making.

So our paper defines two of those. One threshold is prioritizing sensitivity. So we're imagining that those in transfusion departments would be keen on that one. That's still able to achieve a positive predictive value of over 10%. We're also proposing a threshold that prioritizes specificity, which might be more attractive in high throughput hematology labs, that threshold is still able to pick up three quarters of the "wrong blood in tube" errors that are picked up by the other threshold. But it does it with a very low flagging rate. So it only flags four samples in every 10,000 reviewed and it has a positive predictive value of 60%. So it looks like we can get good performance characteristics out of the model if it's applied in practice.

Bob Barrett: This all sounds very promising, but what work remains to be done before this model can be rolled out to other labs?

Chris Farrell: Yes. I guess what we don't understand yet is how well this model generalizes to other labs. We trained the model using data from the same lab that we did the study in. So essentially what we don't know is how much work we'd need to do. If we were setting up the model in anew lab, it might work very well straight out of the box without any additional work. It may be that we need to do a bit of work with some retraining on local data, or we may need something a bit more involved where we have to retune as well as retrain the model. So that retuning process is just where we select the hyperparameters for the model, which are parameters that go into the model, but the model can't learn during that training process. So it needs to be selected by that analyst and we select it by a formalized trial and error process. And that's what's known as tuning. So we may have to go through that at other sites.

I guess the other thing that we don't have a lot of data behind at the moment is those thresholds that have been proposed, that's only based on a relatively small amount of data. So we'd love more data coming out to be able to verify those

thresholds and I guess to try and encourage others to get involved in some of this work. We've released the model publicly and there's a reference to the repository that it's in our paper.

So I encourage people that might be interested to go and have a look and see if the model is helpful for them in their labs and even conduct some formal investigations of it. I guess the thing I haven't mentioned yet is probably the big one, and that's if we're going to be using these models in the lab, we need to be able to integrate them with our existing IT systems. I'm not an IT person, so I don't have any answers here. But when I went with the question to our IT department, they said that to get the model working within existing systems in real time, that would be very time consuming and costly.

I do think we should be pushing our IT departments to be able to integrate models like this, but it looks like it would also be worthwhile for us to be pushing middleware and LIS vendors to create modules so that we can plug in machine learning models into their systems to be able to use them. Because I think this would really allow the research to be accelerated and then translate those research findings into practical benefits for us in the lab.

Bob Barrett: Well, finally, Dr. Farrell, what do you think people should be taking away from this work?

Chris Farrell: Well, I guess the first thing is just that machine learning can be helpful for us on this "wrong blood in tube" error issue. We should be still trying to stop these errors occurring in the first place, for sure. But given that they're likely to continue to appear in the lab, this is just another tool that we can use to help detect them once they arrive and hopefully keep patients safe from the consequences that might arise from reporting those errors. I guess the other thing more generally is that the work kind of provides an illustration of the kind of problem that machine learning can help us with and that machine learning can have practical benefits for us in the lab. And I think to a large extent, that's been the outcome of having a laboratory in leading the work.

And so I'd be encouraging others that are interested, but maybe haven't gotten involved in machine learning yet, really to get involved, because I think that it'd be really valuable for the profession to have lab people leading this kind of work. So I do think that that would be really valuable for us. So I guess historically, clinical chemists have led the way with adopting new technologies into our laboratories, and I think this is just another example where we can continue to contribute there. And I think certainly with the increasing

number of papers that are coming out, that the future for lab-based machine learning applications is very bright.

Bob Barrett: That was Dr. Chris Farrell from Liverpool Hospital in Sydney, Australia. He and his colleagues published a study describing the application of machine learning to detect mislabeled specimens in the September 2023 issue of *Clinical Chemistry*, and he's been our guest in this podcast on that topic. I'm Bob Barrett. Thanks for listening.