Clinical Chemistry

A Multianalyte Machine Learning Model to Detect Wrong Blood in Complete Blood Count Tube Errors in a Pediatric Setting

Article:
Brendan V Graham, Stephen R Master, Amrom E Obstfeld, Robert B Wilson.
A Multianalyte Machine Learning Model to Detect Wrong Blood in Complete Blood Count Tube Errors in a Pediatric Setting
Clin Chem 2025; 71(3): 418–27. https://doi.org/10.1093/clinchem/hvae210

Guests: Brendan Graham and Dr. Stephen Master from the Children's Hospital of Philadelphia, in Philadelphia, Pennsylvania.

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

Wrong blood in tube errors are events in which a collection tube contains blood from patient A, but is marked as having been collected from patient B. These undetected pre-analytical errors have the potential to cause serious adverse outcomes, including death. Fortunately, they occur infrequently but often go undetected because existing tools to identify them have inherent limitations. Traditionally, wrong blood in tube errors are detected using "delta checks," logic rules that notify lab staff if the difference between a patient's previous and most recent result exceeds a certain predetermined threshold. While better than nothing, delta checks are notoriously imperfect, often generating more false alarms than actionable notifications of wrong blood in tube errors.

A new research article, appearing in the March 2025 issue of *Clinical Chemistry*, describes a better way. A multi-analyte machine learning model to identify wrong blood in tube errors in a pediatric population, which brings unique challenges relative to the adult care setting.

In this podcast, we're joined by two of the article's authors. Brendan Graham is a data scientist in the Department of Pathology and Laboratory Medicine at the Children's Hospital of Philadelphia, where he leads machine learning projects within the Division of Pathology Informatics.

Dr. Stephen Master is Chief of the Division of Laboratory Medicine, Director of Metabolic and Advanced Diagnostics, and Director of the Center for Diagnostic Innovation at the Children's Hospital of Philadelphia.

And Dr. Master, let's start with you. Readers of *Clinical Chemistry* may have seen wrong blood in tube articles previously so what is unique about this study?

© 2025 Association for Diagnostics & Laboratory Medicine

Stephen Master: Well, first of all, we should probably review what the problem itself is. So, laboratorians have known for a long time that the most common sources of error occur in the so-called pre-analytics phase, before a measurement's actually made in the clinical laboratory and one important class of these errors is the so-called wrong blood in tube problem, which is often abbreviated as WBIT, or "wibit." So a WBIT occurs when the label on the tube doesn't accurately reflect the patient's sample that's in the tube. That can be due to a number of reasons. It could be due to drawing the wrong patient. It could be due to putting the incorrect label on the tube, so other similar errors.

So now it's clear that you can have an important clinical impact that comes from this kind of an error. So, over the last few years, a variety of groups have started to ask whether it's possible to automatically detect WBIT errors using machine learning to compare, for example, CBC results that were collected over time from a patient, and the hope of course is to improve patient safety by preventing an incorrect diagnosis.

So now, your question is what's new about our study. As a children's hospital, we were interested in determining whether previous machine learning WBIT detection methodologies would be appropriate for our population. There's a famous saying in pediatrics that "children are not just small adults," so we were interested, of course, to see how the methodology that was developed for an adult population would work for our population here at Children's Hospital of Philadelphia.

Now additionally, we also know that WBITs are rare events so we really wanted to dig into whether it was possible to achieve a level of predictive performance that would be useful to a clinician.

Bob Barrett: Well, those seem like important questions. Mr. Graham, what did your study reveal?

Brendan Graham: Well, with respect to that first question, whether prior results would generalize to our population, we found that we could adopt the same methodology used in adult WBIT models in a younger patient population and this wasn't entirely surprising. We had hoped that it would be true and so it was a good confirmatory signal that the methodology generalized not only across institutions but also across populations. Since there has been a lot of pilot work in machine learning for laboratory medicine that has come from single sites, this was an important initial finding.

In a previous *Clinical Chemistry* podcast interview with Dr. Chris Farrell about WBIT machine learning work at Liverpool

© 2025 Association for Diagnostics & Laboratory Medicine

Hospital in Sydney, Australia, he noted that this issue of replication is a really critical question for the field and we really hope that our current paper contributes to answering that question.

The second question has to do with whether we can apply this in a clinical context and as Dr. Master said, WBITs are rare. Prior estimates have suggested that they occur at a rate of about 1 out of 3,500. And so our ultimate goal is to apply this in the clinical lab but first and foremost, we really needed to know that it would work in the pediatric setting with acceptable predictive performance. And finally, because our end goal is implementing this in a clinical workflow, we realized during model development that since we're dealing with a rare outcome, that we needed a way to validate the model in a more realistic way.

Bob Barrett: You also mentioned that wrong blood in tube errors are rare. So why does this matter from a machine learning perspective?

Brendan Graham: That's a great question. Clinical laboratorians already understand the importance of positive predictive value [PPV] for clinical tests, particularly in settings where a condition is very rare. And as mentioned previously, WBITs are very rare, which means that our classifier has to have extremely high specificity in order to be useful to a laboratorian and clinician. So, in machine learning terms, that means we have two main problems.

First, we have to create a classifier. Second, we have to ensure that it will work in a low-prevalence setting. So to create the classifier, we used a similar strategy to prior groups by simulating how a WBIT might occur. We created artificial WBITs by finding pairs of patients that had CBCs around the same time in the same department of the hospital and then then we replaced the results of one of those patients with the results of the other patient's CBC. But in order to train a classifier effectively, we had to apply this matching concept to half of the data. Meaning, that we increased the prevalence of WBIT errors in the training set to around 50%, which was necessary to train the model effectively. However, to address the second problem of realistically assessing model performance, we then have to use a low-prevalence, unbalanced data set in order to ensure that our model performance will be fit for purpose in the clinical setting.

So we went back and used that same simulation approach to create what we call a low-prevalence validation data set that we could use to assess the model. Assessing the models this way means that we will likely see some drops in model performance, but it allows us to get a much better sense of how the models might behave in a real-world lab setting.

© 2025 Association for Diagnostics & Laboratory Medicine

And had we not done this step, we wouldn't have been able to truly assess how the model would perform if and when we start to alert clinicians in the real-world hospital setting.

Bob Barrett: Now, we hear a lot about alerts in the medical record. Mr. Graham, did you worry about alert fatigue?

Brendan Graham: Yes. And this is a very important topic that we considered throughout the project. In fact, this is exactly why it was important for us to demonstrate our model's performance in a low-prevalence setting. When thinking about implementation for a model like this, we know that the model won't be perfect and while we hope every sample we flag is a true positive, we know that the model will generate some false positives. And if clinicians start to receive too many of these false positives, they may start to ignore the message entirely and this is the alert fatigue you just mentioned. When we talked to clinical colleagues about how often an alert like this would have to be a true positive, the number we got in response is about 20%. Given the prevalence of WBITs, that seemed to be the sweet spot where it was high enough to act on, or to be aware of even if the sample wasn't redrawn, and really, this is how we knew that we have been successful in this project because we were able to reach that threshold for PPV while having very few, or even zero, false negatives.

In addition to the volume of alerts, another aspect of this under consideration at the moment is the wording of the alert. We're making somewhat nuanced statements about the probability that a sample is a WBIT and we want to make sure that however we deliver this alert, that it's interpreted within the wider clinical context of the patient.

Bob Barrett: One of the challenges with the large scale screen like a wrong blood in tube detection model is validating that you've truly identified a mislabeled specimen. So Dr. Master, how did you approach prospective validation to make sure that your model was working in the real world?

Stephen Master: Oh, we realized that we actually had a fairly interesting opportunity to use real-world data to test our model because several years ago, our institution implemented positive patient ID, or PPID. What that means is that we added a process where a nurse or a phlebotomist collecting a specimen scans the barcode on the tube and on the patient at the time of collection and therefore provides an independent check that we have the right sample in the right tube. And we already knew from the literature that PPID had been shown empirically to lower the WBIT rate. So we were then able to apply our model to historical data from our

© 2025 Association for Diagnostics & Laboratory Medicine

institution before and after this change to see if there was a change in the number of WBITs predicted by our model.

Now, we'd already used a lot of our recent patient data to build and test the model in the first place. So we initially didn't have as much post-PPID data as we would like in a perfect world, but even on a preliminary analysis described the paper, we saw that we detected WBITs before PPID had roughly the rate reported in the literature, and we saw no predictive WBITs after PPID. Now, of course, subsequent to submitting the paper, we have continued to collect new data, and this new analysis confirms a clinically and statistically significant drop in the rate of predicted WBITs after instituting PPID. The new predicted WBIT rate post-PPID isn't zero, but it is clearly lower, and our analysis shows a clear change after real-world intervention. So this provides really strong corroborative evidence that our predicted WBITs are in fact identifying real pre-analytical errors.

Now, you might say, "if your WBIT rate is very low after PPID, why do you still need to use a machine learning model?" and we think there are at least two answers to that question. And the first is that even if that WBIT rate is low, the risk to a patient of a WBIT is so important that we want to have many systems in place to ensure that we don't miss something. But second, we know that there are many health systems who don't yet have the resources to fully implement PPID and we think that our machine learning model provides a cost-effective way for them to take some safety steps until they're able to implement a more expensive operational change.

Bob Barrett: Well finally Dr. Master, you mentioned some possible next steps with regard to implementation. Where do you see the future of this research at CHOP and at other institutions?

Stephen Master: Well, as we said at the outset, we didn't want to develop the model just to see if it was possible. The goal was always to implement this in a clinical workflow and that takes us into the realm of something called machine learning operations, also known as MLOps. So for us, this is still a work in progress. We do have the advantage of working with several other teams or institutions that have been building the infrastructure to facilitate real-time data access and to deploy models to make this possible, and it's worth pointing out that this is a great cascade from machine learning in the clinical lab because it's really focused on operational safety alerts that may or may not have some of the same regulatory issues that you'd have if you were creating a patient diagnostic result. So we're very excited to see the latest plan over the next year and maybe we'll be fortunate enough to come back in the future and tell you about that experience.

© 2025 Association for Diagnostics & Laboratory Medicine

Bob Barrett: That was Dr. Stephen Master and Brendan Graham, both from the Children's Hospital of Philadelphia. They co-authored a new research article in the March 2025 issue of *Clinical Chemistry,* describing a multi-analyte machine learning model to detect wrong blood in tube errors, and they've been our guests in this podcast on that topic.

I'm Bob Barrett. Thanks for listening.

© 2025 Association for Diagnostics & Laboratory Medicine