Clinical Chemistry

Generalizability of a Machine Learning Model for Improving Utilization of Parathyroid Hormone-Related Peptide Testing across Multiple Clinical Centers

Article:
He S Yang, Weishen Pan, Yingheng Wang, Mark A Zaydman, Nicholas C Spies, Zhen Zhao, Theresa A Guise, Qing H Meng, and Fei Wang.
Generalizability of a Machine Learning Model for Improving Utilization of Parathyroid Hormone-Related Peptide Testing across Multiple Clinical Centers
Clin Chem 2023; 69(11): 1260–9. https://doi.org/10.1093/clinchem/hvad141

Guests: Dr. He Sarina Yang from New York Presbyterian Hospital/Weill Cornell Medicine and Dr. Fei Wang from Weill Cornell Medicine.

Bob Barrett: This is a podcast from *Clinical Chemistry*, a production of the Association for Diagnostics & Laboratory Medicine. I'm Bob Barrett. Laboratory stewardship, or the practice of ensuring physician-ordered clinical laboratory tests will help guide care decisions, is often performed manually on a case-by-case basis. This time intensive process is generally carried out by laboratory medicine trainees, taking them away from other responsibilities. As an alternative, machine learning models have the potential to streamline test review by analyzing the wealth of information in a patient's electronic medical record and assessing whether that patient is likely to have the clinical condition evaluated by the ordered laboratory test. This would reduce the time spent on manual review and would facilitate stewardship efforts at hospitals without training programs. While machine learning has shown tremendous promise in this area, one major limitation is the ability to transfer models developed at one hospital to another institution that may have different instrument platforms or overall ordering practices. A new article appearing in the November 2023 issue of *Clinical Chemistry* addresses this challenge by developing a machine learning model to guide use of a frequently overused test and then assessing strategies to successfully implement the model at other hospitals.

In this podcast, we are excited to talk with the article's lead and senior authors. Dr. Sarina Yang is an associate professor in the Department of Pathology and Laboratory Medicine and the Medical Director of Clinical Chemistry and Toxicology at New York Presbyterian Hospital, Weill Cornell Medicine. Dr. Fei Wang is an associate professor of Health Informatics in the Department of Population Health Sciences at Weill Cornell Medicine, where he is also the founding director of the Institute of AI for Digital Health. Let's start with you, Dr. Yang. Can you give us some background on this and explain why you chose to work on the parathyroid hormone-related protein project?

Sarina Yang: Sure. This work started from the AACC Machine Learning Data Challenge in 2022. The Data Analytics Steering

Committee and the Washington University School of Medicine in St. Louis organized the first machine learning competition last year, which challenged the participants to develop an algorithm to predict the PTHrP results based on patients' other laboratory test results available at the time of PTHrP ordering. The parathyroid hormone-related peptide, or PTHrP, is a tumor marker, which is able to mimic certain action of PTH. PTHrP can stimulate calcium resorption from bone and reabsorption in the kidney, thereby increasing calcium concentration in blood. PTHrP is the most common cause of malignancy-related hypercalcemia. Detectable concentrations of PTHrP are most often due to malignant solid organ tumors, but also reported to be detectable in patients with hematological malignancies.

So PTHrP testing can help diagnose hypercalcemia of malignancy when the source of elevated calcium is not evident. In theory, a logical, systematic approach to the workup of patients with hypercalcemia should include confirmation of elevated calcium, ideally through the determination of ionized calcium, followed by measurement of intact PTH. If PTH is high, patient is diagnosed as primary hyperparathyroidism, and if PTH is low, then order PTHrP. However, in real practice, elevation of total calcium often prompts a simultaneous order of both PTH and PTHrP tests, which is an inappropriate strategy that wastes laboratory resources. PTHrP testing is often ordered on patients with a low pretest probability. Excessive PTHrP testing can lead to expensive, unnecessary, and potentially harmful procedure. As a result, many institutes use a manual, rule-based approach.

For example, residents review PTH and calcium results in order to identify inappropriate PTHrP orders and try to convince the ordering physician to cancel the PTHRP. This approach is labor-intensive and time-consuming. Therefore, the purpose of this work is to develop a machine learning model which could help laboratorians to quickly and accurately identify potentially inappropriate PTHrP test orders and improve the PTHrP test utilization.

Bob Barrett: So how does your model work, and why is it better than the traditional manual approach?

Sarina Yang: Our model was developed based on the dataset provided by Washington University Department of Pathology and Lab Medicine, and they provided a real, de-identified clinical dataset consisting of 1,330 PTHrP orders from 2012 to 2022, along with patients' other laboratory test results available at the time of PTHrP ordering. The dataset was anonymously divided into 80% training and 20% test set. After data cleanup and normalization, we selected laboratory features that had a missing rate less than 50% and also showed a

significant difference between PTHrP normal and abnormal patient cohorts. We then tried four different popular models and determined that the gradient boosted decision tree model performed the best in the five-fold cross-validation in the training set.

So the model can analyze the pattern of laboratory test result profile and then output a score which predicts the likelihood of an abnormal PTHrP result. In the independent test set, our XGBoost model achieved an area under the ROC curve of 0.936. When sensitivity was set to 0.9, specificity was 0.842, and this performance was the highest among 24 participating teams in the data challenge. The organizer of this data challenge also said that in the CLN News that our model achieved a significant improvement compared to their manual approach for identifying patients at risk for PTHrP testing.

Bob Barrett: Okay, well Dr. Wang, let's turn to you. In the title of your paper, you focused on the generalizability of machine learning models. Can you tell us what generalizability means and why is it an important issue?

Fei Wang: Yeah, thanks for this great question. So once a machine learning model is built, it is critical to understand its performance on other data sets that are independently collected from patient populations with a different geographic or demographic or different laboratory settings. This is to avoid becoming overly optimistic about the model. Many factors can affect the model's generalizability in addition to population characteristics. In clinical laboratory setting, the differences among various laboratories, including instrument platforms, test methodology, sample handling, and the use of standout laboratories all pose technical challenges for model generalization. Therefore, it is highly recommended to evaluate model performance in external datasets before a machine learning model can be deployed. Here in this study, we evaluated our model's performance in datasets collected from Weill Cornell Medicine and MD Anderson. In fact, the differences among the three laboratories did have a great impact on model's performance.

Bob Barrett: Dr. Wang, what challenges did you encounter when evaluating the generalizability of your model?

Fei Wang: So when we directly apply the model to the data collected from Weill Cornell Medicine and MD Anderson, its performance moderately deteriorated in MD Anderson, but substantially dropped in Weill Cornell. As I mentioned above, this could be caused by the different chemistry analyzers in the lab and different standout labs for PTHrP, which represents a major challenge when applying machine learning model to different sites. We cannot assume that a predictive model built on one site can automatically be transported to

other sites. To further understand why our model's performance drops at other sites, we investigated the discrepancy of data distribution across different sites. We leveraged something called maximum mean discrepancy, or MMD, to quantify the degree of distribution shift between pairwise datasets. A higher MMD indicates greater distribution shifts, which may lead to poorer model transportability. In our case, the MMD between the data from WashU and Weill Cornell is larger than the MMD between WashU and MD Anderson. So the WashU model performance was worse in Weill Cornell than in MD Anderson.

The next question is, if a machine learning model cannot be directly transported to external sites, so what can we do? A straightforward thought is to develop a new model using site-specific data. We have demonstrated in the paper that if we have enough site-specific data, the new model can perform much better than the transported model. However, in the real world, there could be sites with limited resources, such as small suburb hospitals or hospitals in low socioeconomic status areas, where we do not have large enough sample sites to train a good site-specific model.

In this case, we propose a model fine-tuning strategy, which treats the parameters of the transported model as a warm start and fine-tunes them with limited site-specific data. We demonstrated that in the case of transporting the model trained from WashU data to Weill Cornell, our fine-tuning strategy worked the best when the number of available samples at Weill Cornell is less than 200, compared to direct model transporting or learning the site-specific model for Weill Cornell.

Bob Barrett: Well, this certainly has been a thorough study. Finally, what do you hope people should take away from this?

Fei Wang: I think there are a few important points here. So, first, our proposed PTHrP prediction model is a feasible and promising model that has the potential to improve PTHrP test utilization. Machine learning can potentially improve laboratory workflow and efficiency. Second, we cannot expect a ready-made model developed in one institution to automatically work in other institutions. Model generalizability should be rigorously evaluated in external datasets. Third, when the model cannot be directly transported to other hospitals, site-specific customization strategies can be employed to improve the model performance.

Bob Barrett: That was Dr. Fei Wang and Dr. Sarina Yang from Weill Cornell Medicine in New York City. They published a study describing the generalizability of a machine learning algorithm to guide test utilization in the November 2023 issue of *Clinical*

Chemistry, and they've been our guests in this podcast on that topic. I'm Bob Barrett. Thanks for listening.