Clinical Chemistry 69:11
1260–1269 (2023)

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

He S. Yang,[a,†] Weishen Pan,[b,†] Yingheng Wang,[c] Mark A. Zaydman,[d] Nicholas C. Spies,[d] Zhen Zhao,[a] Theresa A. Guise,[e] Qing H. Meng,[f,*,‡] and Fei Wang[b,*,‡]

BACKGROUND: Measuring parathyroid hormone-related peptide (PTHrP) helps diagnose the humoral hypercalcemia of malignancy, but is often ordered for patients with low pretest probability, resulting in poor test utilization. Manual review of results to identify inappropriate PTHrP orders is a cumbersome process.

METHODS: Using a dataset of 1330 patients from a single institute, we developed a machine learning (ML) model to predict abnormal PTHrP results. We then evaluated the performance of the model on two external datasets. Different strategies (model transporting, retraining, rebuilding, and fine-tuning) were investigated to improve model generalizability. Maximum mean discrepancy (MMD) was adopted to quantify the shift of data distributions across different datasets.

RESULTS: The model achieved an area under the receiver operating characteristic curve (AUROC) of 0.936, and a specificity of 0.842 at 0.900 sensitivity in the development cohort. Directly transporting this model to two external datasets resulted in a deterioration of AUROC to 0.838 and 0.737, with the latter having a larger MMD corresponding to a greater data shift compared to the original dataset. Model rebuilding using site-specific data improved AUROC to 0.891 and 0.837 on the two sites, respectively. When external data is insufficient for retraining, a fine-tuning strategy also improved model utility.

CONCLUSIONS: ML offers promise to improve PTHrP test utilization while relieving the burden of manual review. Transporting a ready-made model to external datasets may lead to performance deterioration due to data distribution shift. Model retraining or rebuilding could improve generalizability when there are enough data, and model fine-tuning may be favorable when site-specific data is limited.

Introduction

About 90% of total hypercalcemia cases are diagnosed as primary hyperparathyroidism and malignancy-related hypercalcemia (1, 2). The latter is primarily mediated by parathyroid hormone-related peptide (PTHrP), which stimulates calcium resorption from bone and reabsorption in the kidneys (3). Hypercalcemia mediated by PTHrP is most frequently caused by malignant solid organ tumors, and is indicative of a poor prognosis (4). Clinically, measuring PTHrP levels can aid in diagnosing the humoral hypercalcemia of malignancy when the source of elevated calcium levels is not immediately evident (5). However, PTHrP testing is often ordered on patients with a low pretest probability of this condition. As a result, many institutes employ a manual, rule-based approach in which the laboratory medicine residents review PTH and calcium results and attempt to identify inappropriate orders in instances where the likelihood of an abnormal PTHrP result is low (e.g., high calcium levels and high PTH levels). This approach is labor-intensive and time-consuming. This inadequate laboratory utilization practice results in increased healthcare costs, drains laboratory resources, and can trigger unnecessary patient anxiety.

[a]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, United States; [b]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States; [c]Department of Computer Science, Cornell University, Ithaca, NY, United States; [d]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, United States; [e]Department of Endocrine Neoplasia and Hormonal Disorders, Division of Internal Medicine, The University of Texas, MD Anderson, Houston, TX, United States; [f]Department of Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, United States.

*Address correspondence to: F.W. at Division of Health Informatics, Department of Healthcare Policy and Research, Weill Cornell Medicine, 425 E. 61st Str., New York City, NY 10065, United States. E-mail few2001@med.cornell.edu. Q.H.M. at Department of Laboratory Medicine, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, R4.1441G, Unit 37, Houston, TX 77030-4009, United States. E-mail qhmeng@mdanderson.org.
[†]Contributed equally.
[‡]Joint corresponding authors.
Received March 27, 2023; accepted August 23, 2023
https://doi.org/10.1093/clinchem/hvad141

To improve the utilization of PTHrP tests, the AACC Data Analytics Steering Committee organized the first data competition, which challenged participants to develop an algorithm for predicting the normalcy of PTHrP results based on other laboratory results available for a patient at the time when the PTHrP test was ordered (5). Machine learning (ML) holds tremendous potential for uncovering intricate relationships among complex laboratory parameters and identifying variables that are not included in traditional diagnostic algorithms (6, 7). A successful predictive model would help laboratorians to identify inappropriate PTHrP orders when the laboratory data available at the time of order already suggests a normal PTHrP result and it would also provide timely clinical information to ordering physicians (8).

In addition to model development, evaluating a model's performance on external datasets that are independently collected from different geographic or demographic populations is crucial to understand its real-world utility (9, 10). However, the differences among various laboratories, including instrument platforms, testing methodologies, sample handling, or use of send-out laboratories, pose technical challenges for model generalization. Based on a recent review of ML papers in the field of laboratory medicine (6), only a small proportion of studies have conducted external validation to demonstrate cross-center generalizability. Therefore, there is a pressing need for the rigorous evaluation of model generalizability.

We developed a ML model that achieved the best predictive performance among the 24 participating teams in the AACC ML data challenge (8). To further evaluate our model's generalizability, we evaluated it on unseen datasets obtained from two independent clinical centers. In this paper, we present the workflow of data collection, data preprocessing, model development, and evaluation, as well as a comprehensive analysis of feature distributions among the three sites and different strategies to improve model generalizability when deploying to external sites.

Methods

DATASETS

A real, de-identified, clinical dataset consisting of 1330 PTHrP orders from 2012 to 2022 along with patients' other laboratory results available at the time of PTHrP order was provided by Washington University School of Medicine in St. Louis (WUSM) in the contest. For patients who had multiple PTHrP orders, only the first order and its associated data were retained. The day and time when each laboratory test was ordered and performed, as well as its corresponding reference interval, were also provided. PTHrP testing offered by WUSM was performed by Mayo Clinic Laboratories (method in the Supplemental Materials). This dataset was anonymously divided by the organizer of the contest into 2 parts, including 1064 patient data (80%) used for training and 266 patient data (20%) for testing model performance. Data were collected in the same format from 2 independent external institutes, Weill Cornell Medicine (WCM, New York, NY) and University of Texas MD Anderson Cancer Center (MDA, Houston, TX). A total of 1101 PTHrP orders from 2017 to 2022, performed by Quest Diagnostics (method in the Supplemental Materials), were collected from WCM and 1090 PTHrP orders from 2021 to 2022, performed by Mayo Clinic Laboratories, were collected from MDA. The proportion of positive samples, i.e., PTHrP values greater than the reference interval, in WUSM, WCM, and MDA were 17.5% (232/1330), 15.9% (175/1101), and 23.9% (260/1090), respectively. Instrumentation and methodologies of routine laboratory tests offered by each site are listed in the Supplemental Materials. This study was approved by the Institutional Review Board of each site (WUSM:202202087 and 202204007; WCM: 21-03023422; MDA: 2022-0760).

DATA PREPROCESSING

The input feature vectors of the prediction model were constructed with the laboratory tests collected within a 1-year observation window prior to a specific PTHrP test. Only laboratory tests that had available measurements during the observation window from at least 50% of the patients were selected. The missing rate of each laboratory test is shown in Supplemental Table 2. Since methodologies of some tests and the reference intervals have changed in the past years, we normalized each laboratory result value (V) by its corresponding reference interval (RR) using the following formula: $V_{\text{norm}} = (\text{V—lower limit of RR})/(\text{higher limit of RR—lower limit of RR})$. After normalization, the statistics of each laboratory test within the observation window were calculated, including minimum value (min), maximum value (max), mean, latest value, and rate of change (slope of the fitted linear regression model). If there were insufficient measurements to calculate the statistics for a given patient, the corresponding statistics were treated as missing values and were imputed with the median value of the statistics across all patients. Comparisons between different imputation methods are shown in Supplemental Table 5. Next, statistics of the laboratory tests were selected if they showed a significant difference (P value after false discovery rate correction (11) less than 0.05) between the PTHrP normal and abnormal patients. Finally, to ensure consistency among the selected features, z-score normalization was employed, given the lack of reference intervals for certain laboratory tests.

MODEL DEVELOPMENT AND EVALUATION

We evaluated 4 popular classifiers including the random forest, support vector machine, extreme gradient boosting

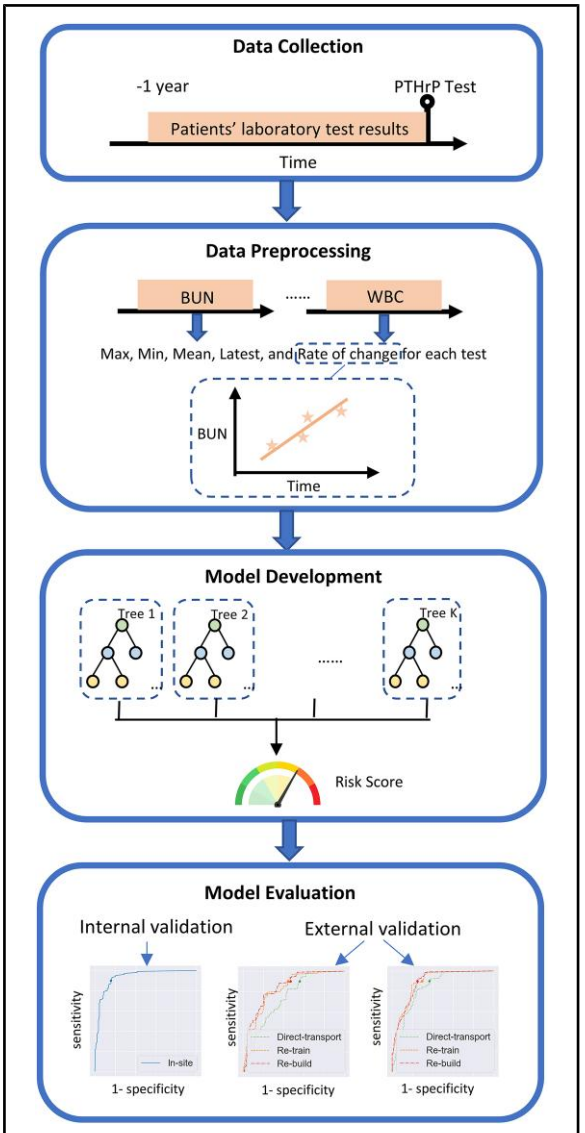

Fig. 1. Illustration of the modeling workflow. Patients' laboratory test results available within 1 year prior to the order of PTHrP testing were used to construct the feature vectors, on which an extreme gradient boosting (XGBoost) classifier was developed to predict the normalcy of PTHrP results. For each laboratory test, 5 statistics, including maximum, minimum, mean, latest values, and the rate of change, were calculated with the measurements during the collection window. These laboratory features were concatenated to construct the feature vectors. The model outputs a probability score ranging from 0 to 1, indicating the likelihood of an abnormal PTHrP result. The PTHrP model was evaluated on the internal test set and 2 external datasets. Color figure available online at clinchem.org.

(XGBoost), and multilayer perceptron models on the WUSM dataset, using the scikit-learn package v.1.1.3 with the sklearn.model_selection.StratifiedKFold function. For each model, the appropriate hyperparameters were determined through a standard 5-fold cross-validation, where the training data were randomly split into 5 equal folds with the same positives/negatives ratio as the ratio of overall cases. Parameters of each model and comparison of their performance in cross-validation are shown in the Supplemental Fig. 1 and Supplemental Table 7. Once the hyperparameters were determined, the entire training set (80%) was used to train a model, which was then applied to the test set (20%) to evaluate the performance measured by area under the receiver operating characteristic curve (AUROC). For the best performing model, we also measured its specificity and precision (or positive predictive value) at an operating point that was set to sensitivity (or recall) at 0.900, given that this model is primarily intended for a screening purpose. The partial AUROC, which is calculated as the area above the sensitivity line of 0.9 on the receiver operating characteristic (ROC) curve, quantifies the predictive performance with sensitivity exceeding 0.900. This metric ranges between 0 and 0.1. The Shapley Additive Explanations technique (12) (http://github.com/slundberg/shap, v.0.41.0) was employed to interpret the selected model and explain the impact of each feature on the model predictions. The features were ranked based on their global Shapley values, which were the average of the magnitudes of their Shapley values with respect to each sample. A force plot of top impactful features illustrates how features act as "force" to push the model to make a prediction of normal or abnormal PTHrP result. Overall, a pipeline of the XGBoost modeling framework is illustrated in Fig. 1.

INVESTIGATION OF MODEL GENERALIZABILITY

Both WCM and MDA data were randomly split into a training set (80%) and a test set (20%) with the same ratio of PTHrP positives/negatives in their respective overall sample populations. The datasets were preprocessed using the same process as in the original WUSM dataset. Initially, the model developed on WUSM data was directly applied to the test sets of the WCM and MDA data. AUROC and specificity/precision calculated at sensitivity level of 0.900 were reported. ΔAUROC was calculated as the difference between AUROC of the model evaluated in the training site and AUROC obtained from directly transporting the model to the testing site. Moreover, we have implemented 2 additional strategies: (*a*) retraining the model using site-specific data with the same model architecture, feature sets (intersection of the feature sets present in both the training and test datasets), and hyperparameters; and (*b*) rebuilding the model using site-specific

data including feature selection, hyperparameter tuning, and model parameter learning (Details of the retraining and rebuilding strategies are in Supplemental Materials. Features and their missing rate for WCM and MDA datasets are shown in Supplemental Tables 3 and 4, respectively). Both model retraining and rebuilding were conducted on the same training sets (80%) and model performances were evaluated on the same test sets (20%).

We also investigated a low-resource scenario where a target institution had limited PTHrP test results that were not enough to retrain a good model, in which case we implemented a fine-tuning strategy for the model developed from the source institution to be adapted in the target institution. In our investigation, sample subsets from WCM were used to mimic a target institution with limited PTHrP orders, and WUSM was used as the source institution for model development. New decision trees of the XGBoost model were added to the original model during the fine-tuning process. The learning rate for fine-tuning was set to one-tenth of that used to train the original model, while other hyperparameters remained unchanged. Cross-validation was utilized on the available samples to determine the optimal number of decision trees for fine-tuning (details in Supplemental Materials).

To further understand the different performances of the model across different sites, we quantified the distribution discrepancies between the data acquired from 3 institutes (13). Specifically, we first picked the top 20 important features of each dataset, which were selected based on their global Shapley values, and then calculated the maximum mean discrepancy (MMD) (14) on the distribution of the intersection of the top 20 features for each pair of institutes. A gaussian kernel was selected, which was used to measure the similarity between pairs of data points. MMD was calculated as the difference between the mean of the kernel values for pairs of data points drawn from the 2 distributions being compared. Given samples from 2 distributions: $X = \{x_1, \ldots, x_N\}$, $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_M\}$, the calculation of MMD was as:

$$MMD = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{i \neq j}^{N} \kappa(x_j, x_i) + \frac{1}{M(M-1)} \sum_{j=1}^{M} \sum_{i \neq j}^{M} \kappa(\tilde{x}_j, \tilde{x}_i) - \frac{2}{NM} \sum_{j=1}^{N} \sum_{i=1}^{M} \kappa(x_j, \tilde{x}_i)$$

where $\kappa(x_j, x_i)$ was the kernel function that was used to determine the latent space where the data points were projected to.

Results

DEVELOPMENT AND EVALUATION OF THE PTHRP PREDICTION MODEL ON THE WUSM DATA

In the WUSM dataset, a total of 48 laboratory tests (listed in Supplemental Materials) were selected based on their missing rate during the 1-year observational window. For each of these tests, 5 statistics, including minimum, maximum, mean, latest, and rate of change, were calculated, resulting in a total of 240 features. After feature selection as described in the Method section, 159 features, which exhibited statistical significance for discrimination between PTHrP normal and abnormal patients, were used to build the ML model. To select the model with the best performance, AUROCs of the random forest, support vector machine, XGBoost, and multilayer perceptron models were compared using 5-fold cross-validation. The XGBoost model outperformed the other 3 models in cross-validation (Supplemental Fig. 1). In the test set, the XGBoost model achieved an AUROC of 0.936. At the operating point with a sensitivity (recall) of 0.900, the model achieved a specificity of 0.842 and a precision (or positive predictive value) of 0.539 (Fig. 2A).

The force plot illustrates the impact of top features on the predictive performance of the PTHrP model according to their Shapley values shown in Fig. 3. For instance, the latest result of albumin level, the maximum total calcium level, the mean phosphorus level, the latest count of white blood cells (WBC), and the mean sodium level within a year prior to the PTHrP order were the 5 most important predictors in the model. Moreover, lower levels of albumin, intact PTH, and phosphate drove the model to make a prediction of abnormal PTHrP results, while higher levels of total calcium and WBC counts led to the same prediction.

EXTERNAL EVALUATION OF THE PTHRP PREDICTION MODEL IN 2 INDEPENDENT INSTITUTES

To further investigate the generalizability of the model developed on the WUSM data, the XGBoost model's performance was assessed in 2 unseen external datasets obtained from WCM and MDA. The rate of positive PTHrP results were 15.9% in WCM and 23.9% in MDA, compared to 17.5% in WUSM. First, when the ready-made model was directly applied "as-is" to the 2 independent datasets, its performance moderately deteriorated in MDA (AUROC = 0.838) but substantially in WCM (AUROC = 0.737). Next, with the fixed model architecture and hyperparameters, as well as the selected input features, retraining the model parameters using local data led to an improved performance at both sites (MDA AUROC = 0.889, WCM AUROC = 0.819). Further improvements on AUROC at both sites were achieved by rebuilding the model with site-specific

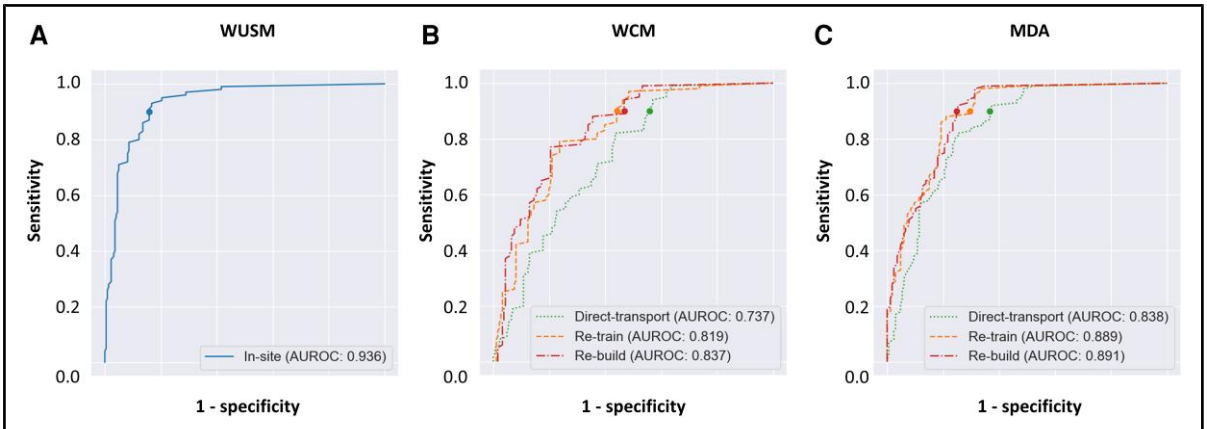

Fig. 2. Comparison of the extreme gradient boosting (XGBoost) model performance on different datasets and training strategies using the ROC curves. The ROC curves depict the performance of the XGBoost model on various datasets and training strategies. The *y*-axis of the ROC curve is sensitivity, and *x*-axis is 1-specificity. AUROC indicates the ability of a classifier to distinguish between 2 classes. (A), The model's ROC curve when trained on 80% of the WUSM dataset and evaluated on the remaining 20% test data; (B, C), The model's performance when tested on the 20% of the WCM (B) and MDA (C) datasets. The green dotted line indicates that the model was trained on the entire WUSM dataset and directly transported to the WCM (B) and MDA (C) test data. The orange dashed line indicates that model was retrained on the 80% local data and accessed on the 20% test data, whereas the red dotted line indicates that the model was rebuilt on the 80% local data and tested on the 20% test data. The large colored dot on each line indicates the operating point with a sensitivity of 0.9. Color figure available online at clinchem.org.

data, which optimized both the selected feature sets and model hyperparameters (MDA AUROC = 0.891, WCM AUROC = 0.837). The ROC curves of each scenario are shown in Fig. 2B (WCM) and 2C (MDA). The force plots illustrating the impact of the top features on the rebuilt WCM and MDA models were shown in Supplemental Fig. 2A and B. Overall, the model that was retrained or rebuilt using data from the test site achieved better performance (Table 1).

ANALYSIS OF THE DIFFERENCES IN FEATURE DISTRIBUTIONS ACROSS THREE SITES

On transporting the ready-made model developed on WUSM data to new patient data collected from WCM and MDA, there was a deterioration in the model's performance, measured by AUROC, specificity, and precision. To better understand the reasons causing such degradation of predictive performance, we calculated the MMD between each pair of datasets (13), which quantified the degree of distribution shift between them, with a higher MMD value indicating a larger shift. As shown in Table 2, the MMD between WUSM and MDA data was smaller than the MMD between WUSM and WCM data, which was consistent with the observed deterioration in cross-site performance, as measured by a drop in AUROC (ΔAUROC).

ANALYSIS OF THE PERFORMANCE OF MODEL FINE-TUNING IN LOW-RESOURCE SCENARIOS

We also considered a scenario where smaller hospitals do not have sufficient local PTHrP and other laboratory data to retrain or rebuild the model. To explore how the ready-made model can be applied to hospitals with limited training data, we assessed the effectiveness of a model fine-tuning strategy. We compared the performance of the following 3 strategies on WCM data since it showed a larger MMD with WUSM data: (*a*) directly using of the WUSM model; (*b*) retraining the WUSM model with varying amounts of data from WCM; and (*c*) fine-tuning the WUSM model with varying amounts of available data from WCM. The results are shown in Fig. 4, which demonstrates that the fine-tuning strategy performed best when the amounts of available data samples were relatively small (<200). However, when the number of available samples exceeded 200, model retraining appeared to be a better option.

Discussion

In this study, we built an ML model to predict the normalcy of PTHrP level using routine laboratory test results available at the time when the patient's PTHrP test was ordered. When evaluated in the WUSM internal test data, the model exhibited a high AUROC, as well as

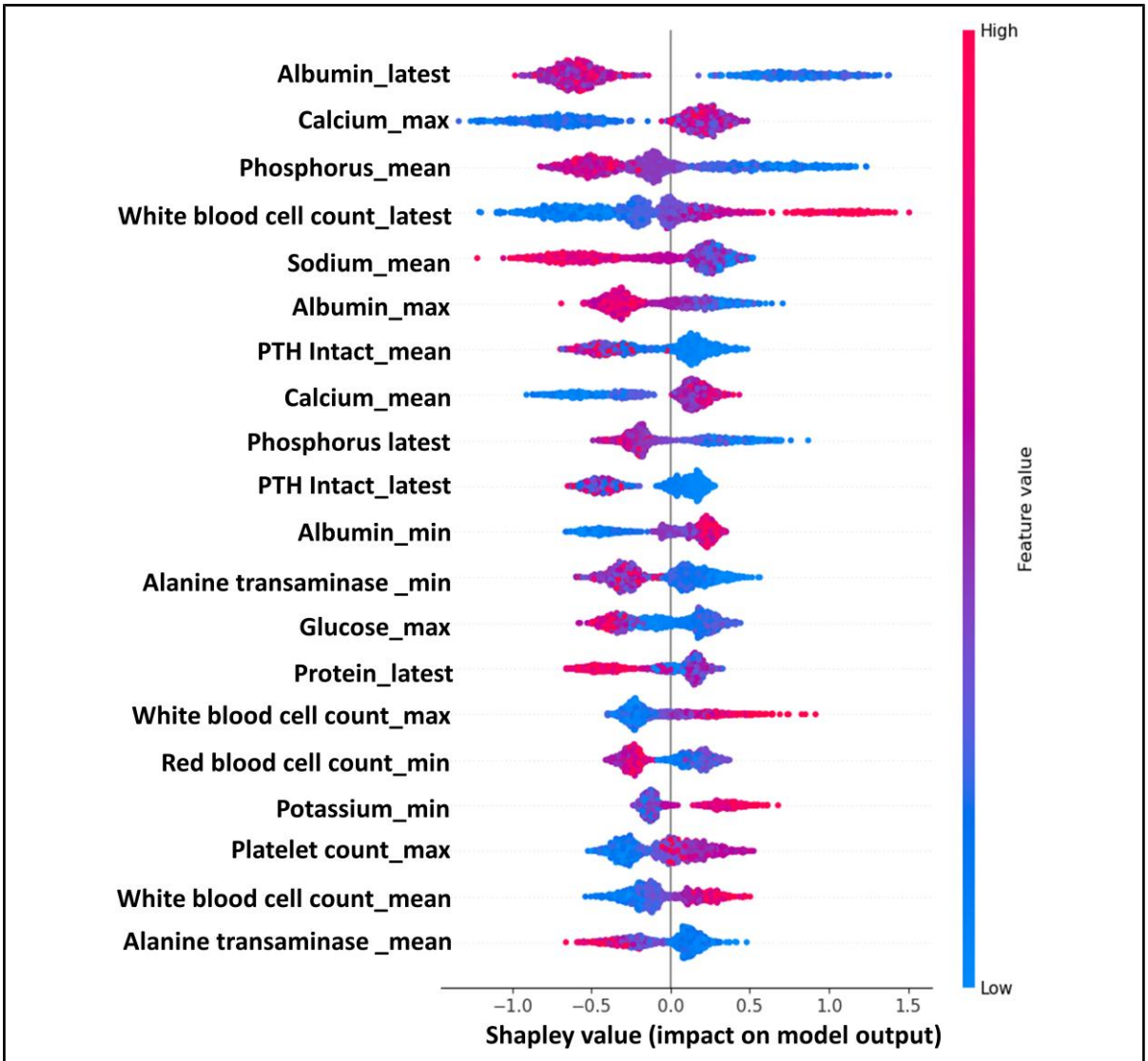

Fig. 3. Impact of each laboratory test statistics on the predictive performance of the PTHrP model, using the Shapley Addictive Explanations (SHAP) technique. The force plot of top features illustrates how features act as "force" to push the model to make a prediction of normal or abnormal PTHrP results. Individual values of each laboratory test statistics for each patient are colored according to their relative values, with the blue color representing lower values of laboratory results, and the red color representing higher values. The laboratory test statistics were ranked based on their global Shapley values shown on the *x*-axis. Positive Shapley values to the right-hand side indicate predictions of abnormal PTHrP results, and negative SHAP values to the left-hand side indicate predictions of normal PTHrP results. The thickness of the line represents the number of value points. Color figure available online at clinchem.org.

reasonable clinical interpretability. We further assessed the generalizability of this model on patient data collected from two independent external sites. Not surprisingly, transporting the ready-made model "as-is" led to a decreased model performance on both datasets. Nevertheless, after retraining and rebuilding the model using site-specific data, we observed significant improvements on model performance at both sites. In the low-resource scenario where a site does not have enough data to retrain and rebuild a customized model, we demonstrated that a fine-tuning strategy could be a favorable choice. Overall, our study developed an ML model that shows promise in improving PTHrP test utilization and demonstrated a comparison of strategies

Table 1. A summary of the extreme gradient boosting (XGBoost) model performance using different training and test datasets.

Method	AUROC	Partial AUROC given sensitivity ≥ 0.900	Specificity given sensitivity = 0.900	Precision (or positive predictive value) given sensitivity = 0.900
Testing: WUSM				
In-site test	0.936	0.068	0.842	0.539
Testing: WCM				
Off-the-shelf model	0.737	0.037	0.441	0.235
Retrain the model	0.819	0.044	0.559	0.281
Rebuild the model	0.837	0.046	0.532	0.269
Testing: MDA				
Off-the-shelf model	0.838	0.050	0.633	0.435
Retrain the model	0.889	0.061	0.705	0.490
Rebuild the model	0.891	0.064	0.753	0.534

Partial AUROC is calculated as the area above the sensitivity line of 0.9 on the ROC curve, which quantifies the predictive performance with sensitivity exceeding 0.9.

Table 2. A summary of the MMD and XGBoost model performance using different training and test datasets.

Training site	Testing site	Maximum Mean Discrepancy (MMD)	AUROC	ΔAUROC[a]
WUSM	WCM	0.084	0.737	0.199
	MDA	0.073	0.838	0.098
WCM	WUSM	0.076	0.707	0.130
	MDA	0.050	0.743	0.094
MDA	WUSM	0.011	0.858	0.033
	WCM	0.038	0.633	0.258

[a]ΔAUROC is calculated as the difference between AUROC of the model evaluated in the training site and AUROC obtained from directly transporting the model to the testing site.

for improving the generalizability of ML in external health systems.

In contemporary clinical practice, measuring PTHrP level aids in determining the cause of unexplained hypercalcemia, characterized by elevated calcium levels without a concurrent increase in PTH (15). However, elevation of total calcium levels often prompts simultaneous requests for both PTH and PTHrP tests, which is an inappropriate utilization of the PTHrP test (5). Excessive PTHrP testing can result in unnecessary and expensive procedures, such as invasive laboratory tests to identify a cancerous tumor that may not even be present (16). However, manually reviewing PTHrP orders by comparing them with a patient's PTH and calcium results can be a cumbersome and time-intensive process. The organizer of this ML competition found that our XGBoost model achieved a significant improvement compared to their manual approach for identifying patients at risk for PTHrP (8). In addition, in the WUSM dataset, if we build a XGBoost model using only the total calcium and PTH intact results available at the time of the PTHrP order to predict the PTHrP normalcy, the AUROC of the model would be 0.762, and specificity would be 0.471 when sensitivity is set to 0.900. The predictive performance would be significantly worse compared to the XGBoost model incorporating other laboratory tests. Thus, if implemented, the proposed ML model that predicts normal and abnormal PTHrP results has the potential to complement the current workup algorithm by detecting inappropriate PTHrP orders, thus facilitating automation of the decision-making process and test utilization. Furthermore, the ML-based data-driven approach detects variables that are presently not included in the existing workup algorithm consisting of intact PTH and total calcium. For instance, patients who have hypercalcemia of malignancy may exhibit lower levels of albumin partly due to liver dysfunction, nephrotic syndrome, or malnutrition. In addition, hypercalcemia of malignancy may be associated with systemic inflammatory response leading to higher levels of WBC and lower levels of albumin. The clinical interpretability of the ML model is crucial as laboratorians and clinicians prefer to use models that can be comprehended and aligned with their knowledge and experience (6).

Before an ML model can be deployed in clinical practice, its generalizability and transportability, i.e.,

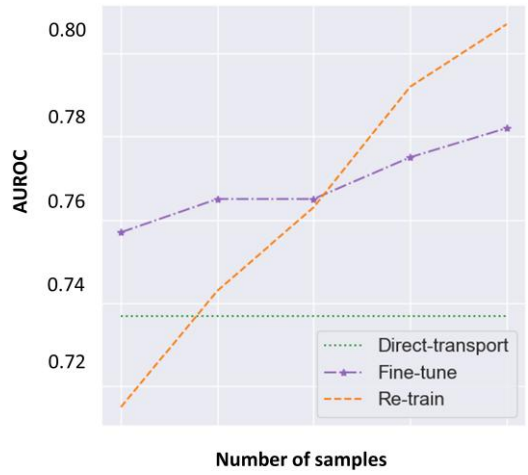

Fig. 4. Comparison of the extreme gradient boosting (XGBoost) model's performance when deploying the model from WUSM to WCM using different amount of WCM samples. Three strategies, including direct transporting (green dotted line), retraining (orange dashed line), and fine-tuning (purple line with star), were evaluated for the model developed from the source institution (WUSM) to be adopted in the target institution (WCM). The performance of each strategy was compared using AUROC. When number of samples was <200, fine-tuning was a more favorable strategy; whereas retraining generated a better result when number of samples was >200. Color figure available online at clinchem.org.

the ability of a model to perform well on independent datasets collected from different geographic or demographic populations or different hospital settings, need to be assessed (17). In the setting of clinical laboratory medicine, various factors such as instrument platforms, test protocols, sample handling, and send-out laboratories, can affect a model's generalizability. Here, we observed that, when transported directly, although the model built on the WUSM data showed deterioration on both WCM and MDA data, its performance was better on MDA data than on WCM data. This difference could be partially attributed to the fact that both WUSM and MDA laboratories use the same analyzers to conduct routine chemistry tests and send their PTHrP samples to the same reference laboratory. By contrast, the WCM laboratory employs a different vendor's chemistry analyzers and sends their PTHrP to another reference laboratory. In fact, we observed that some laboratory test features exhibited distinct distributions between WUSM and WCM, which could be due to variations in clinical or laboratory processes, or different patient populations. Based on our analysis, if a ready-made model cannot be directly transported to new data due to the shift of data distribution, some local customization strategies can be utilized to improve model performance, such as retraining or rebuilding the model using site-specific data.

We have further investigated the quantitative relationship between the performance drop when transporting the model from one data set to another and the discrepancy between their data distributions measured by MMD. Supplemental Fig. 6 shows that MMD and ΔAUROC correlate, suggesting a larger discrepancy could lead to a greater performance degradation. This suggests that the MMD value, which has been a widely adopted statistical metric for measuring distribution shift in transfer learning, could be used to predict performance deterioration of ML models when transported to external sites. A previous study on an acute kidney injury prediction model across 6 institutes also drew a similar conclusion (13). While there is not a specific threshold of MMD value that ensures successful generalization, calculating MMD can facilitate the adoption process of ML models in external hospitals.

Our investigation on low-resource scenarios where the testing site has a limited amount of PTHrP requests and other available laboratory test results showed that fine-tuning model parameters works better than retraining a customized model. In practice, we can store models trained from different sites into a "model bank." When a model is needed for a new site, we can first assess whether there are enough samples on this site to retrain a customized model; if not, we can pick a model (from the model bank) trained from the site that is the most similar to the new site (according to hospital operation patterns, patient populations, etc. or MMD between data distributions if possible) and fine-tune it.

Unfortunately, patient demographic and other clinical information was not provided in the original WUSM dataset, and thus was not incorporated into the model's input. It would be interesting to explore whether model performance can be further enhanced with additional clinical features.

In conclusion, our proposed PTHrP prediction model is a feasible and promising technique that has the potential to improve utilization of PTHrP testing. While directly transporting the ready-made model to external datasets led to a deterioration of model performance due to a shift of data distribution, site-specific customization strategies were employed to improve the predictive performance in a new context. Calculating MMD prior to model fitting on a new dataset could provide valuable insight into the degree of model generalizability, thereby facilitating the model adoption process.

However, the path toward fully operational implementation is still long and laden with obstacles that span technical challenges, regulatory requirements, and the need for clinical education. The successful deployment of this model into clinical practice will require a collaborative effort among IT specialists, clinical teams, and laboratory professionals.

Supplemental Material

Supplemental material is available at *Clinical Chemistry* online.

Nonstandard Abbreviations: PTHrP, parathyroid hormone-related peptide; ML, machine learning; MMD, maximum mean discrepancy; AUROC, area under receiver operating characteristic curve; WUSM, Washington University School of Medicine in St. Louis; XGBoost, extreme gradient boosting; WBC, white blood cells.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved. Nobody who qualifies for authorship has been omitted from the list.*

He Yang (conceptualization—lead, data curation—supporting, formal analysis—supporting, investigation—lead, methodology—supporting, project administration—lead, supervision-lead, validation—supporting, writing—original draft—lead, writing—review and editing—lead), Weishen Pan (conceptualization—supporting, formal analysis—lead, investigation—supporting, methodology—lead, software—lead, visualization-lead, writing—original draft—supporting, writing—review and editing—supporting), Yingheng Wang (formal analysis—supporting, investigation—supporting, methodology—supporting, writing—review and editing—supporting), Mark Zaydman (data curation—supporting, writing—review and editing—supporting), Nicholas Spies (data curation—supporting, writing—review and editing—supporting), Zhen Zhao (writing—review and editing—supporting), Theresa Guise (data curation—supporting, writing—review and editing—supporting), Qing Meng (conceptualization—supporting, data curation—supporting, writing—review and editing—supporting), and Fei Wang (conceptualization—equal, formal analysis—lead, investigation—lead, methodology—lead, resources—lead, supervision—lead, writing—original draft—equal, writing—review and editing—equal).

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form.*

Research Funding: F. Wang, support from NSF awards (1750326, 2212175) and NIH awards RF1AG072449, R01MH124740, and R01AG080991. T.A. Guise, Scholar of CPRIT (Cancer Research Prevention Institute of Texas), RR190108 Established Investigator Award.

Disclosures: Y. Wang, Presidential Life Science Fellowship from Cornell University. M.A. Zaydman, recipient of a grant from BioMerieux and an Intramural grant (Washington University); honoraria for invited speakership (Sebia) and invited speakership (Siemens); recipient of travel support from AACC; provisional patent US 20210311046 A1.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

Acknowledgments: We want to thank Richard Fedeli, Kelvin Espinal, and Ron A. Phipps for collecting and organizing the datasets of laboratory testing results.

Code Availability: The source code of the manuscript is publicly available online via a public repository on Github: https://github.com/weishenpan15/pthrp-prediction. The DOI of the repository is 10.5281/zenodo.7767390.

References

1. Koster M, Brandle M. [Hypercalcemia—diagnosis and management]. Praxis (Bern 1994) 2022;111:675–81.
2. Meng QH, Wagar EA. Laboratory approaches for the diagnosis and assessment of hypercalcemia. Crit Rev Clin Lab Sci 2015;52:107–19.
3. Ashrafzadeh-Kian S, Bornhorst J, Algeciras-Schimnich A. Development of a pthrp chemiluminescent immunoassay to assess humoral hypercalcemia of malignancy. Clin Biochem 2022;105–106: 75–80.
4. Donovan PJ, Achong N, Griffin K, Galligan J, Pretorius CJ, McLeod DS. PTHrP-mediated hypercalcemia: causes and survival in 138 patients. J Clin Endocrinol Metab 2015;100:2024–9.
5. Fritchie K, Zedek D, Grenache DG. The clinical utility of parathyroid hormone-related peptide in the assessment of hypercalcemia. Clin Chim Acta 2009;402:146–9.
6. Yang HS, Rhoads DD, Sepulveda J, Zang C, Chadburn A, Wang F. Building the model: challenges and considerations of developing and implementing machine learning tools for clinical laboratory medicine practice. Arch Pathol Lab Med 2023;147: 826–36.
7. Haymond S, McCudden C. Rise of the machines: artificial intelligence and the clinical laboratory. J Appl Lab Med 2021;6: 1640–54.
8. Michaud S. Learning by doing in data analytics. Clinical Laboratory News 2022. https://www.aacc.org/cln/articles/2022/october/learning-by-doing-in-data-analytics (Accessed March 2023).
9. Haymond S, Master SR. How can we ensure reproducibility and clinical translation of machine learning applications in laboratory medicine? Clin Chem 2022;68:392–5.
10. Yang J, Soltan AAS, Clifton DA. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit Med 2022;5:69.
11. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57:289–300.
12. Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: von Luxburg U, Bengio S, Fergus R, Garnett R, Guyon I, Wallach H, Vishwanathan SVN, editors. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017). Long Beach (CA): Neural Information Processing Systems; 2017. p. 4768–77.
13. Song X, Yu ASL, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, et al. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun 2020;11:5668.
14. Smola AJ, Gretton A, Borgwardt KM. Maximum mean discrepancy. In: King I,

Wang J, Chan L-W, Wang D, editors. Neural information processing. Proceedings of the 13th International Conference, ICONIP 2006; Hong Kong (China): Springer-Verlag Berlin; 2006. p. 3–6.

15. Asonitis N, Angelousi A, Zafeiris C, Lambrou GI, Dontas I, Kassi E. Diagnosis, pathophysiology and management of hypercalcemia in malignancy: A review of the literature. Horm Metab Res 2019;51: 770–8.
16. Kushnir MM, Straseski JA. What can we learn from sensitive and specific measurements of parathyroid hormone-related peptide (pthrp)? AACC Scientific Shorts 2021. https://www.aacc.org/science-and-research/scientific-shorts/2021/what-can-we-learn-from-sensitive-and-specific-measurements-of-parathyroid-hormonerelated-peptide (Accessed March 2023).
17. Master SR, Badrick TC, Bietenbeck A, Haymond S. Machine learning in laboratory medicine: recommendations of the IFCC working group. Clin Chem 2023;69:690–8.