VI CONGRESO LATINOAMERICANO DE BIOQUIMICA CLÍNICA
II CONGRESO INTERNACIONAL DEL COLEGIO NACIONAL DE BACTERIOLOGÍA
¡El riesgo es que te quieras quedar!
Cartagena, Colombia 3 al 6 OCTUBRE 2024

# New Tools for Reducing Errors in the Total Testing Process

COLABIOCLI
Confederación Latinoamericana de Bioquímica Clínica
CNB
Colegio Nacional de Bacteriología
www.congresocolabiocli.com

Weill Cornell Medicine

# Pre-pre-analytical: Developing a Machine Learning Model to Improve Test Utilization and Laboratory Stewardship

He Sarina Yang, PhD, MB, DABCC, FADLM
Associate Professor
Department of Pathology and Laboratory Medicine
Weill Cornell Medicine
Hey9012@med.cornell.edu

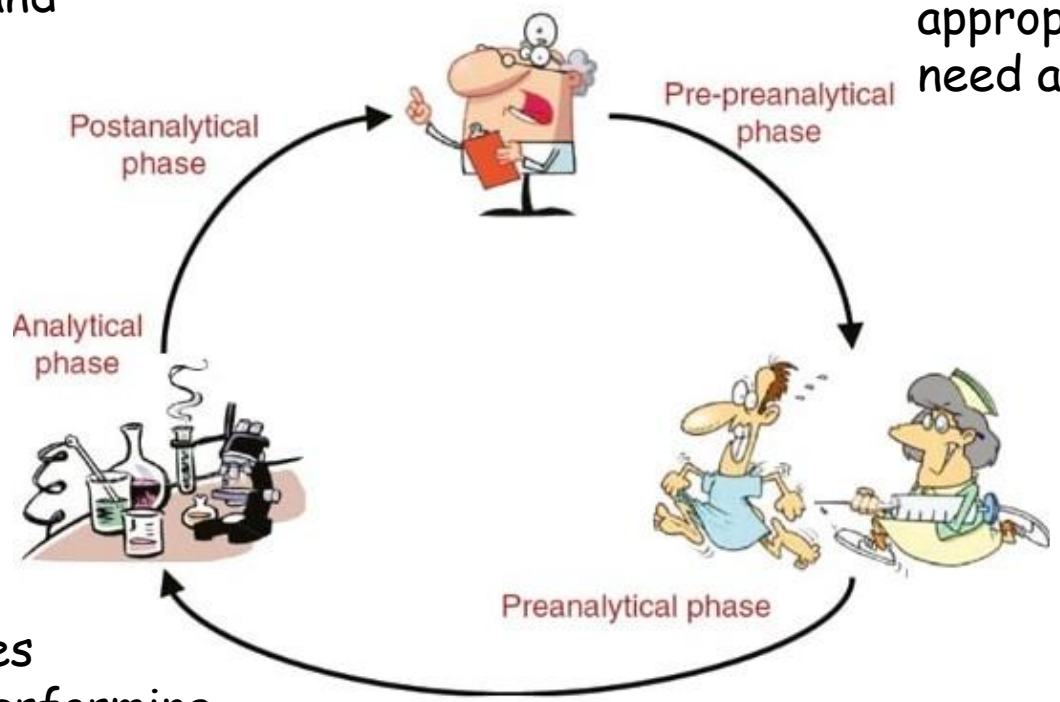
# Relevant Disclosure

- I don’t have any conflict of interest relevant to this presentation to disclose.

# Phases in the Total Testing Process

The **post-analytical phase** includes reporting results to clinicians, interpretation, and addressing any follow-up requirements.

The **pre-pre-analytical phase** involves activities before sample collection, including ordering tests, ensuring appropriate selection of tests on clinical need at the right time.




The **pre-analytical phase** covers specimen collection, handling, and transportation.

The **analytical phase** involves processing the specimens, performing tests, and generating results.

This figure is cited from Batt Lab. https://battlab.com/pre-analytical-errors-what-are-they-and-how-to-avoid-them/

# The Emerging Importance of the Pre-pre-analytical Phase in Reducing Testing Errors

Unlicensed Published by De Gruyter October 12, 2016

**Towards a new paradigm in laboratory medicine: the five rights**

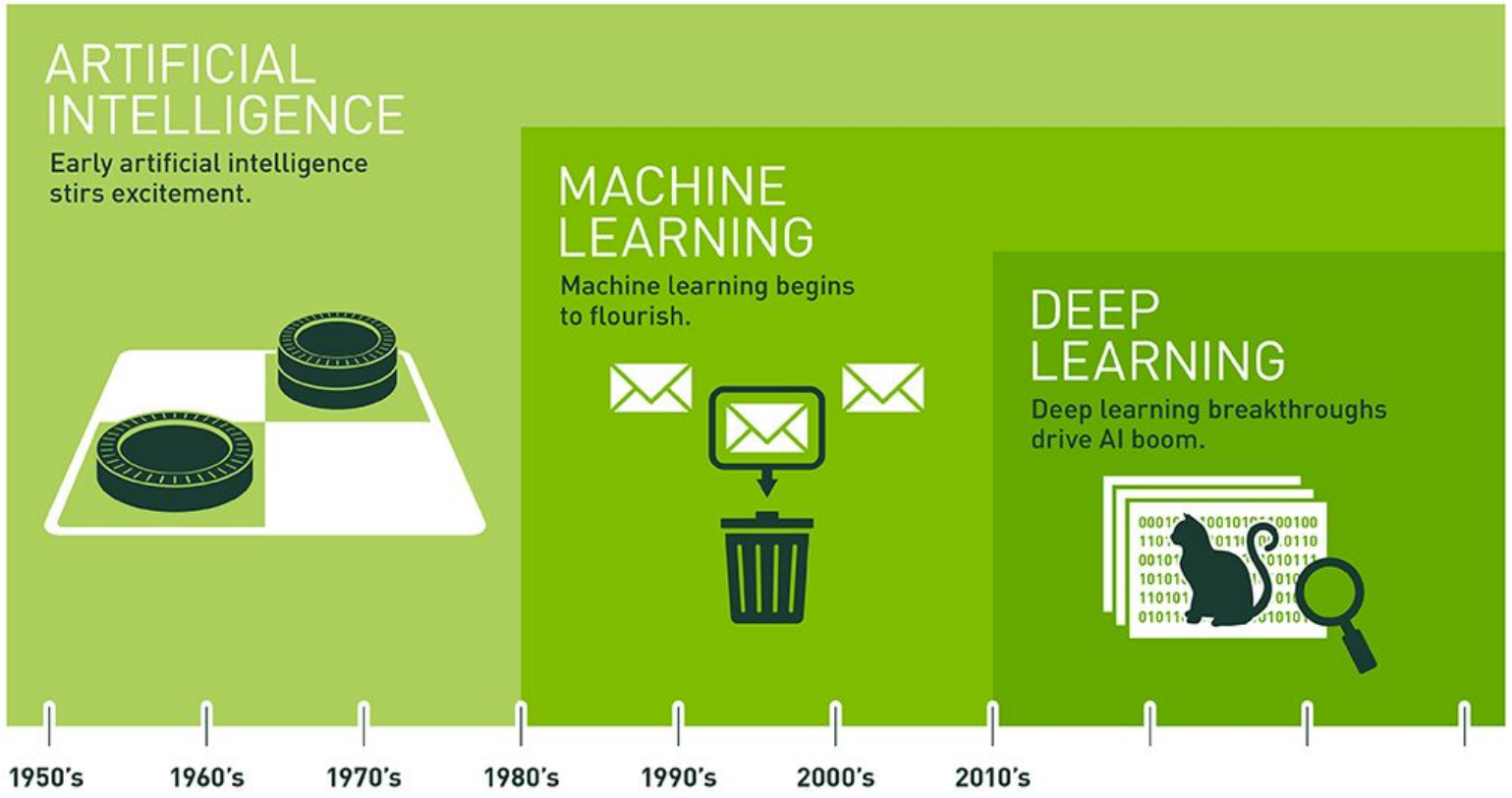
Mario Plebani

From the journal Clinical Chemistry and Laboratory Medicine (CCLM)
https://doi.org/10.1515/cclm-2016-0848

<u>The “Five Rights” in the pre-pre-analytical phase:</u>

- The **right patient** i.e. the use of the **correct sample**, the use of the **right test** at the **right time**, and the **right method** of sample collection and transportation.

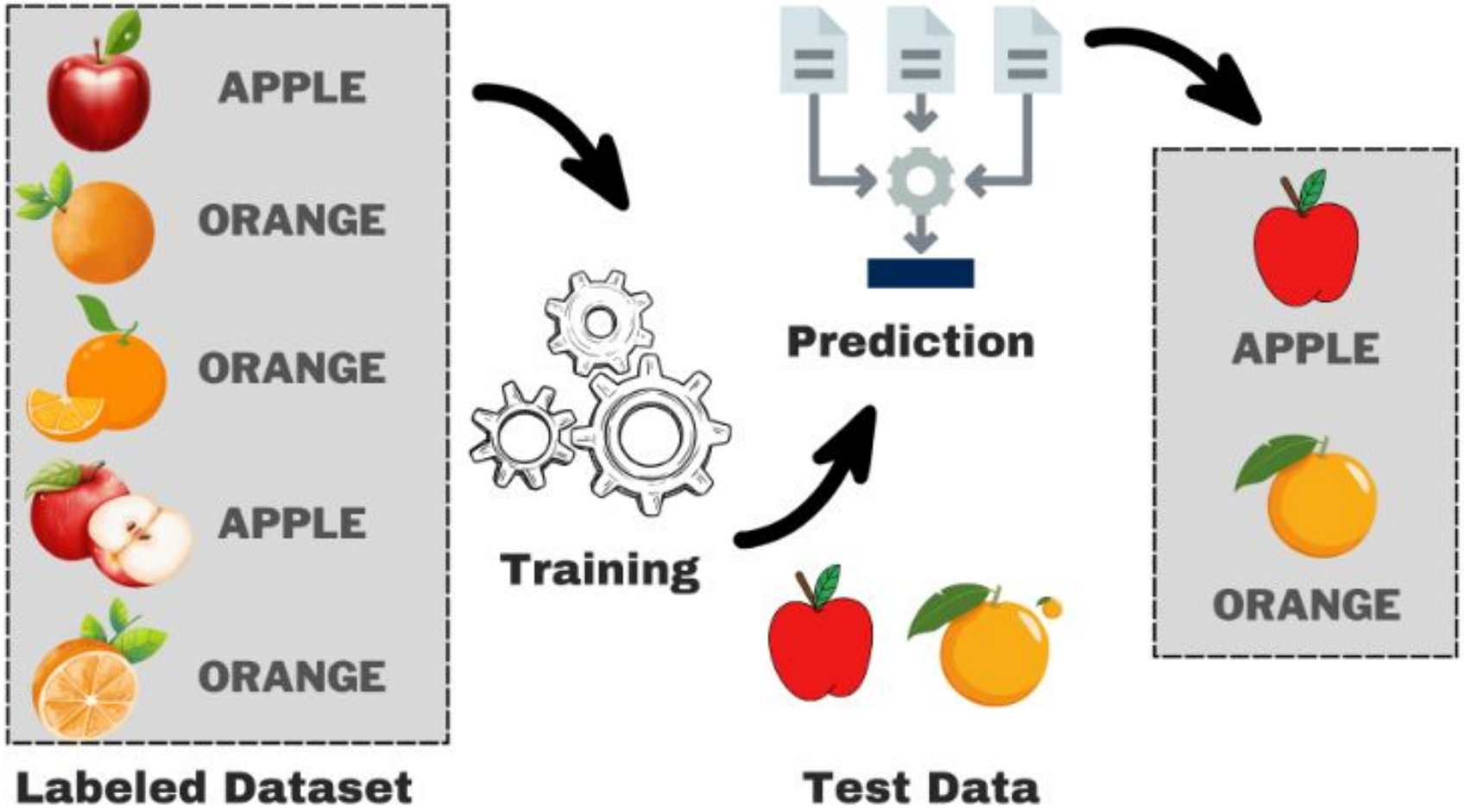
# What is Machine Learning?




Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

- Machine learning is a subset of AI that enables systems to learn from data, identify pattern, and make decision with minimal human intervention.
- It involves using algorithms to analyze large datasets, extract insights from the data, and improve performance over time as more data becomes available.

# Supervised Machine Learning



# ML Empowers Laboratory Data Analysis

- Up to 70% of data in the EHR are derived from the clinical laboratories.
- Most test results are reported as individual numerical or categorical values in structured formats.
- High dimensional and longitudinal data.
- Manual analysis and interpretation is challenging!
- Computational data analytics provides valuable information.

# Opportunities for Machine Learning in Laboratory Medicine

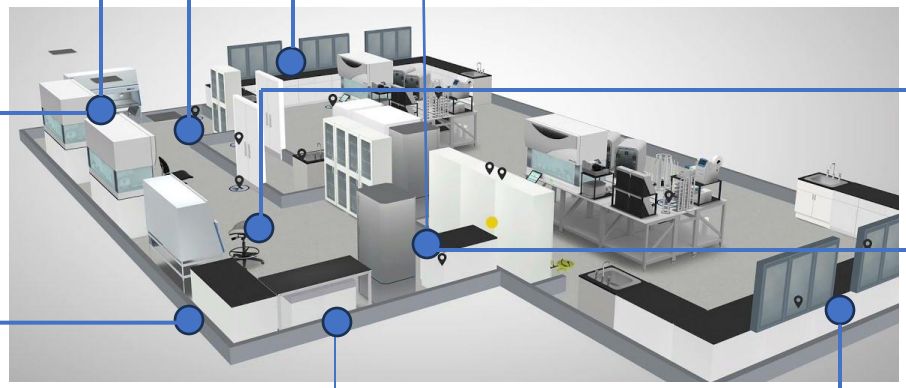
Improve laboratory workflow

Predict accuracy of measured laboratory results

Interpret complex laboratory results

Clinical decision support
Predict onset, progression, subtypes, outcome of diseases e.g. sepsis, AKI, COVID-19, cancer

Facilitate early cancer detection



Identify pre-analytical errors, e.g. wrong blood in tube, IV fluid contamination, hemolysis

Improve test selection

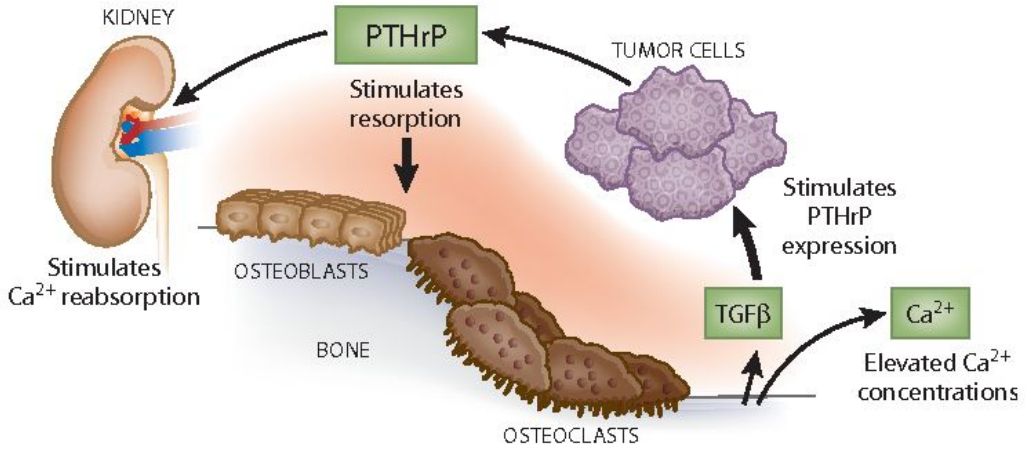
Enhance quality control
e.g. early warning

Establish population-based reference range

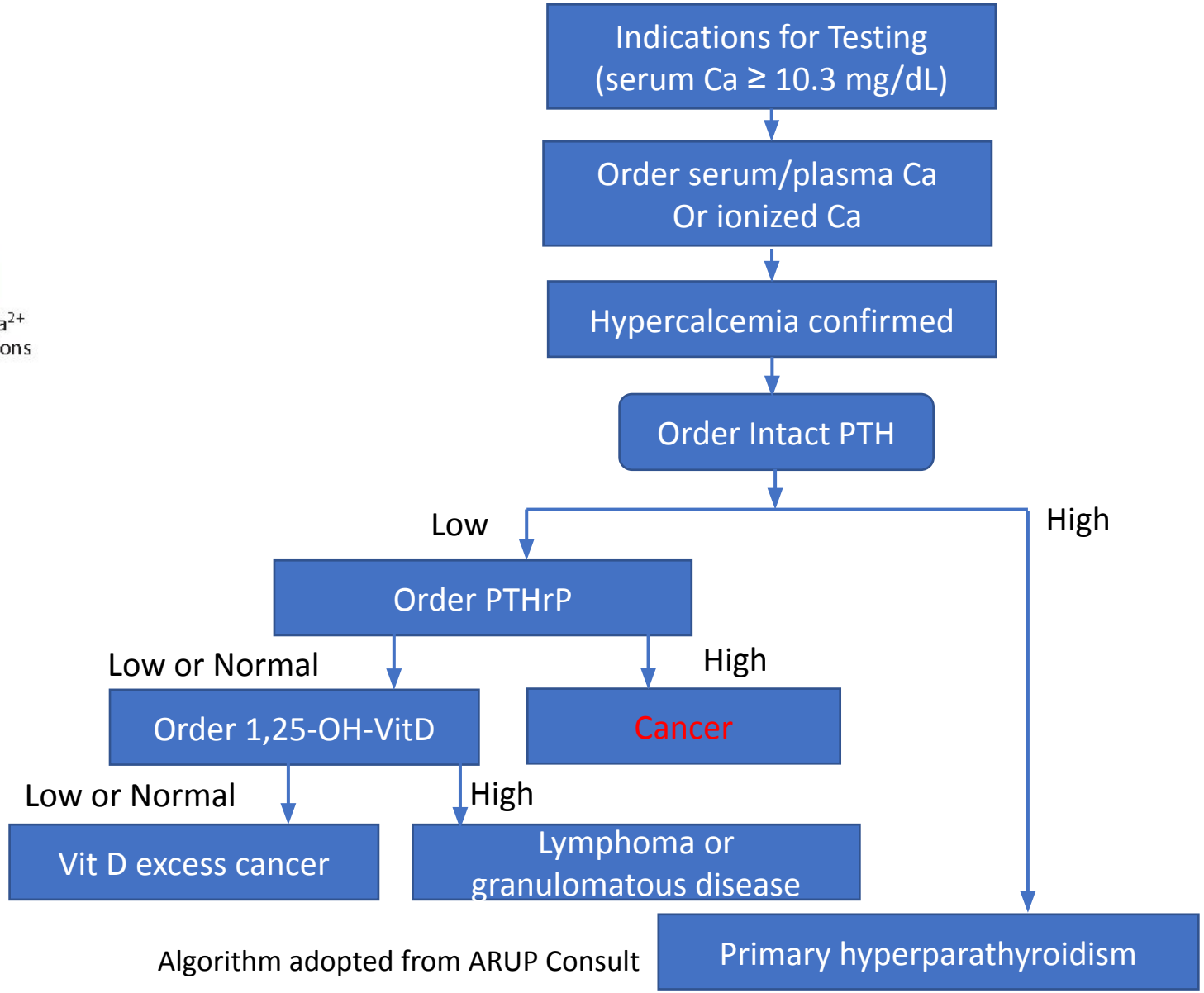
Analyze genetic and genomic data

Microorganism detection and identification
Disease classification

# Parathyroid Hormone-Related Peptide (PTHrP)




- 90% of total hypercalcemia cases are diagnosed as primary hyperparathyroidism and malignancy-related hypercalcemia.
- PTHrP is the most common cause of humoral malignancy-related hypercalcemia.
- Hypercalcemia mediated by PTHrP is most frequently caused by malignant solid organ tumors, and it is indicative of a poor prognosis.
- PTHrP testing can aid in diagnosing hypercalcemia of malignancy when the source of elevated calcium is not evident.




Mundy GR and Edwards JR, J Am Soc Nephrol 2008; 19 (4): 672-5

Algorithm adopted from ARUP Consult

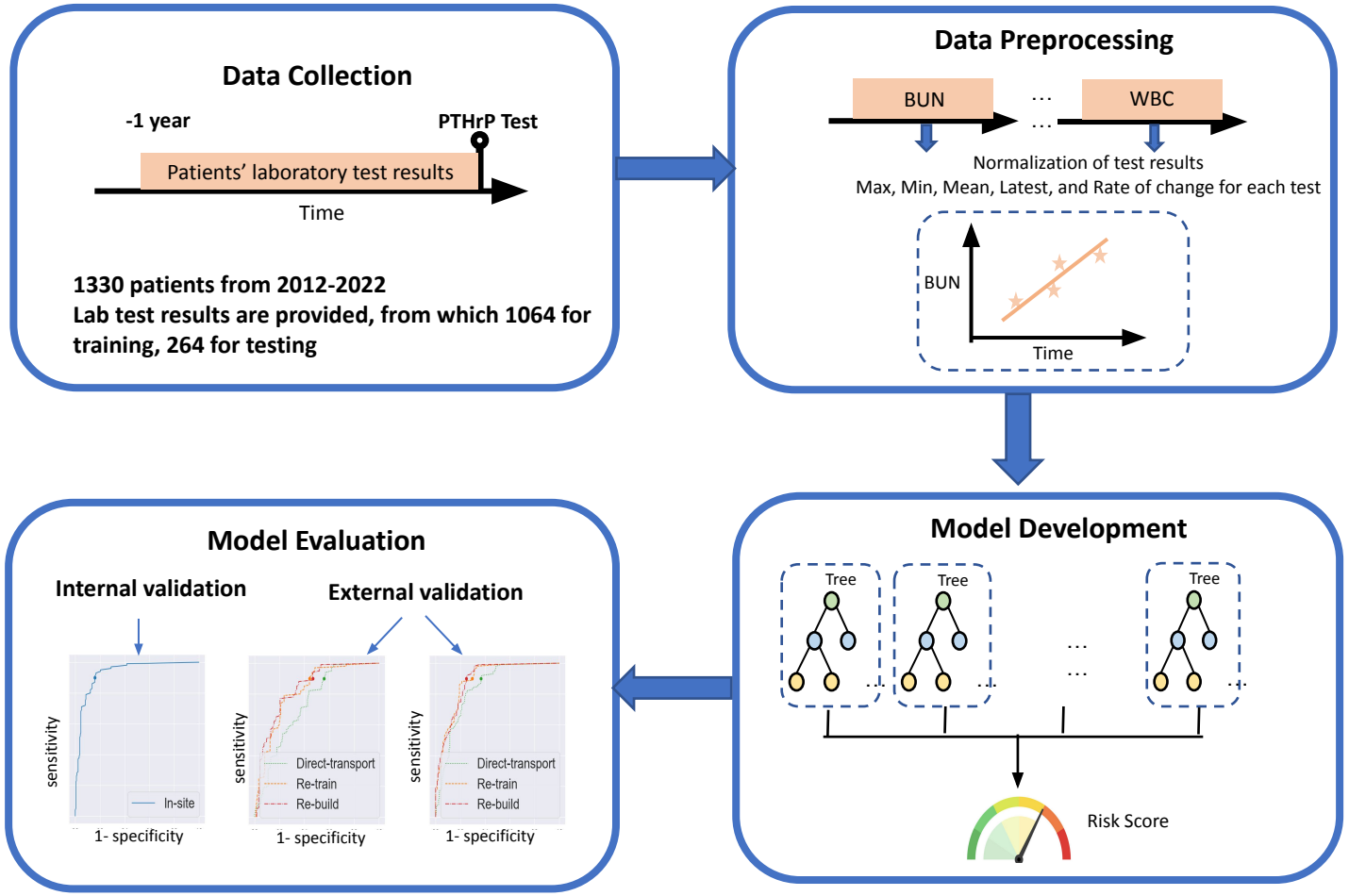
# Poor Utilization of the PTHrP testing

- PTHrP testing is often ordered for patients with a low likelihood of having hypercalcemia of malignancy, indicating a low pre-test probability.
- PTHrP is typically a send-out test to reference laboratories with TAT ranging from one to two weeks.
- For emergency department patients and inpatients, PTHrP testing may not be reimbursed if the results come back too late.
- This results in increased healthcare costs, wastes laboratory resources, and can trigger unnecessary patient anxiety.
- 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. This approach is labor-intensive and time-consuming.

# Objectives

- Develop a machine learning model to predict PTHrP results based on patients' other laboratory results available at the time of PTHrP ordering.
- Investigate whether the ML model can potentially complement the current hypercalcemia workup algorithm by identifying inappropriate PTHrP orders, thereby improving test utilization and laboratory stewardship.

# Overall Workflow of PTHrP Model Development and Evaluation

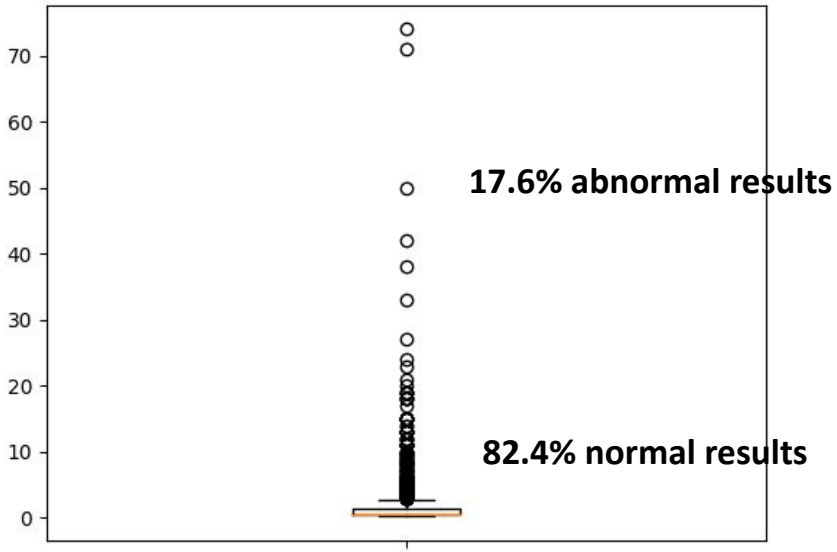


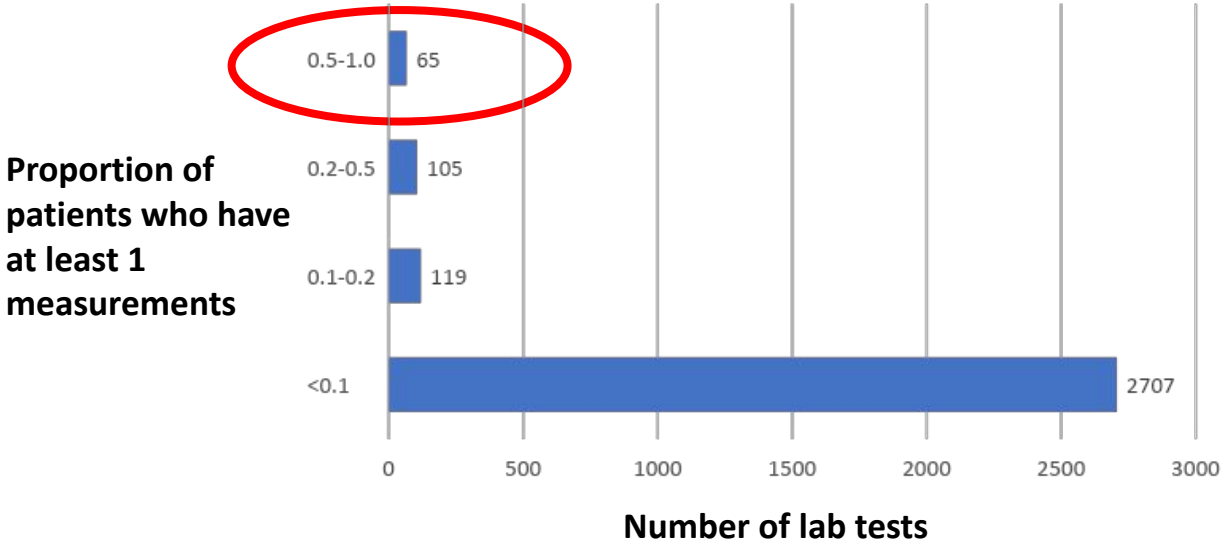

Yang et al. *Clin Chem* 2023; 69(11):1260-1269

# Overview of the Original Dataset

- A real, de-identified dataset consisting of 1330 PTHrP orders from 2012-2022 along with patients' other lab results available at the time of PTHrP ordering was provided by WUSM.
- PTHrP testing offered by WUSM was performed by Mayo Clinic Laboratories.
- This dataset was anonymously divided into training set (1064 patients) and test set (266 patients).

**The distribution of PTHrP results:**




- A total of 2,996 lab tests were ordered for all patients
- The majority of the tests were not ordered for most patients

# Data Normalization

- Test methodology and reference range of the PTHrP assay and some lab tests changed during the past 10 years
- Normalization of the lab test values based on their respective reference range

$$val_{norm} = \frac{val - normal\_low}{normal\_high - normal\_low}$$

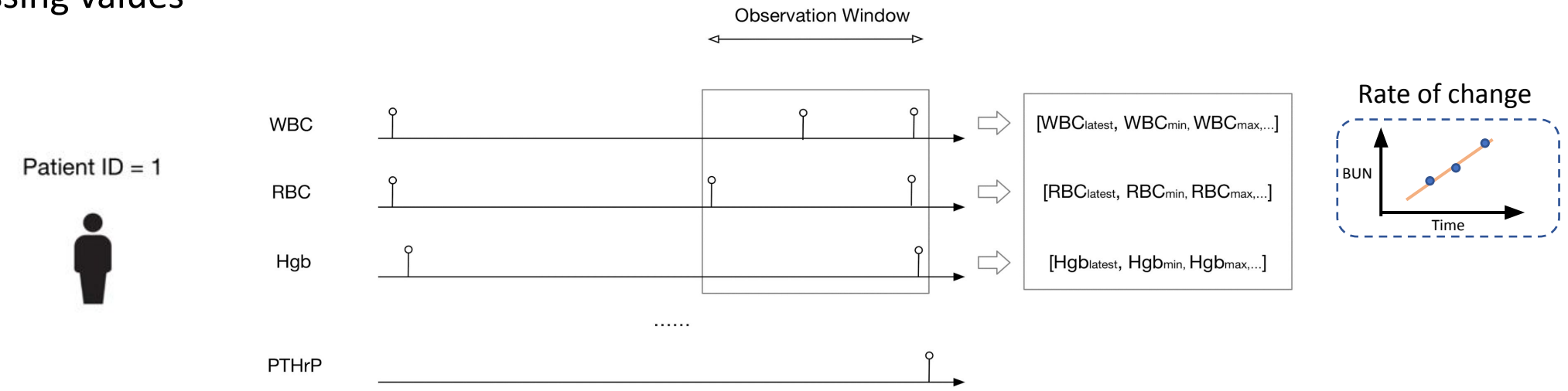
- Example

| PATIENT_ID | ORDERABLE | TASK_ASSAY | RESULT_VAL | RESULT_UNITS | NORMAL_LOW | NORMAL_HIGH | NORMALCY | RESULT_STATUS | ORDER_STATUS | ORDER_PROVIDER | ORDER_DT_TM | PERFORMED_DT_TM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | CBC Hosp_I | WBC | 17.7 | K/cumm | 4.8 | 10.8 | HI | Auth | Completed | 38 | 2012/7/24 23:06 | 2012/7/24 23:43 |

$$val_{norm} = \frac{17.7 - 4.8}{10.8 - 4.8} = 2.15$$

# Data Preprocessing

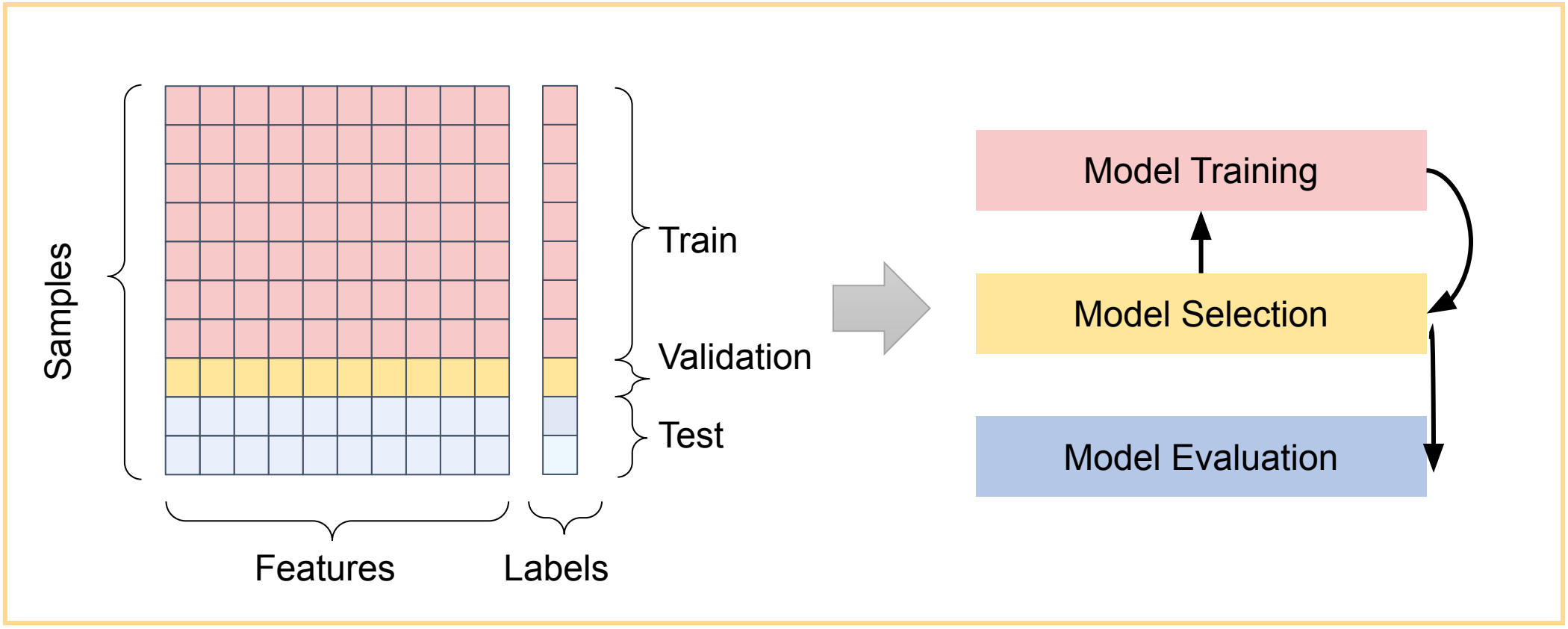
- Compare the distribution of each laboratory test between PTHrP normal and abnormal patient cohorts.
- Determined the observation window prior to the PTHrP order (1 year)
- Calculated statistics of each lab test in the observation window
  - Min, max, mean, latest results
  - Rate of the change
  - Number of measurements in the observation window
  - Missing values



# Feature Selection and Missing Value Imputation

- 159 lab statistics were selected based on their missing rate (< 50%) during the observation window and statistical testing between the PTHrP normal and abnormal patients (p value after false discovery rate correction < 0.05)
    - Mann-Whitney test for the continuous features
    - Chi-square test for the binary features
- Missing values were imputed with the median value of the statistics across all patients

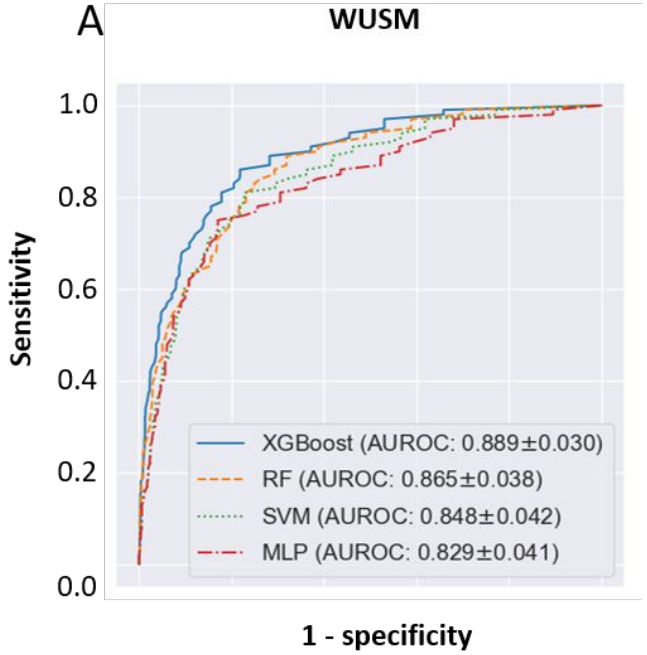
# Overall Process of Model Selection, Training, and Evaluation




Yang et al., Archives of Pathology and Lab Medicine, 2022

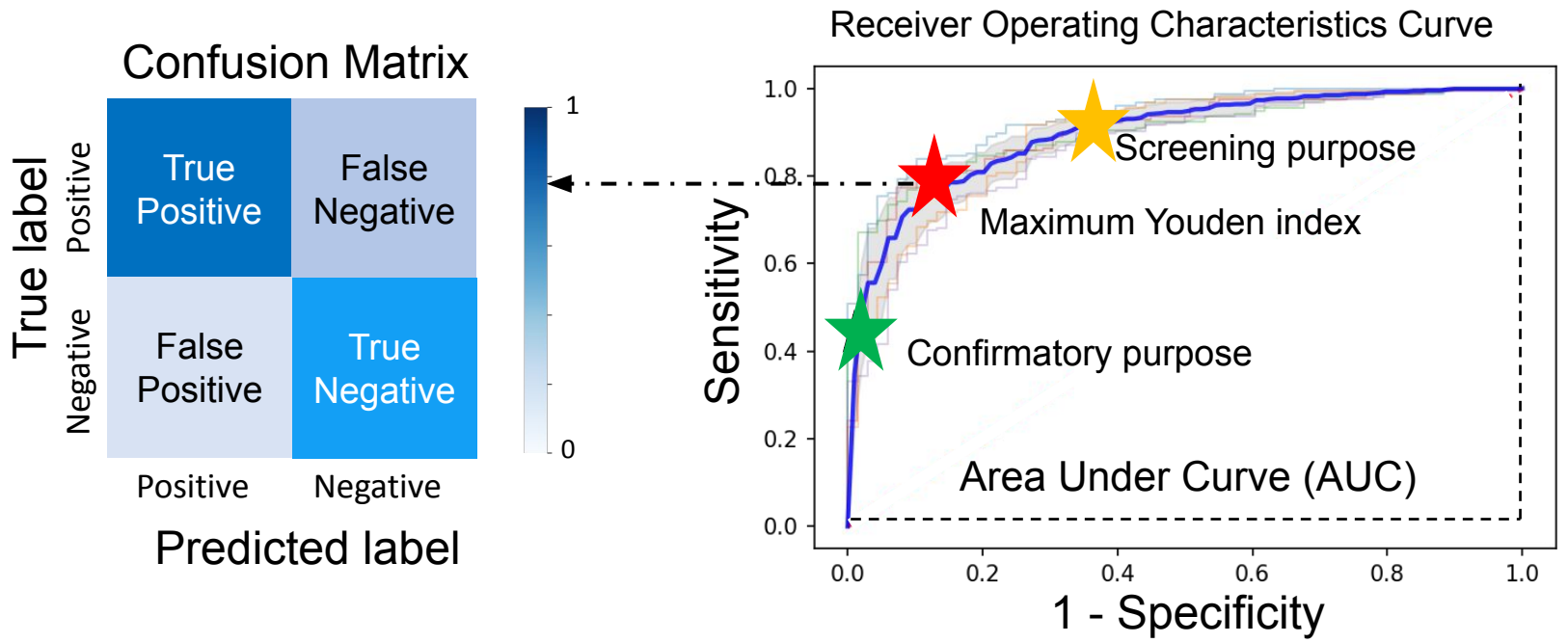
# Model Selection and Training

- Model: XGBoost
  - The XGBoost model outperformed random forest (RF), support vector machine (SVM), multi-layer perceptron (MLP) models in the cross-validation of training data.
- Critical hyperparameters for the XGBoost model
  - Number of selected features: 159 laboratory test statistics
  - Learning rate
  - Prediction threshold



# Model Evaluation

- It is recommended to use multiple criteria to comprehensively evaluate model performance

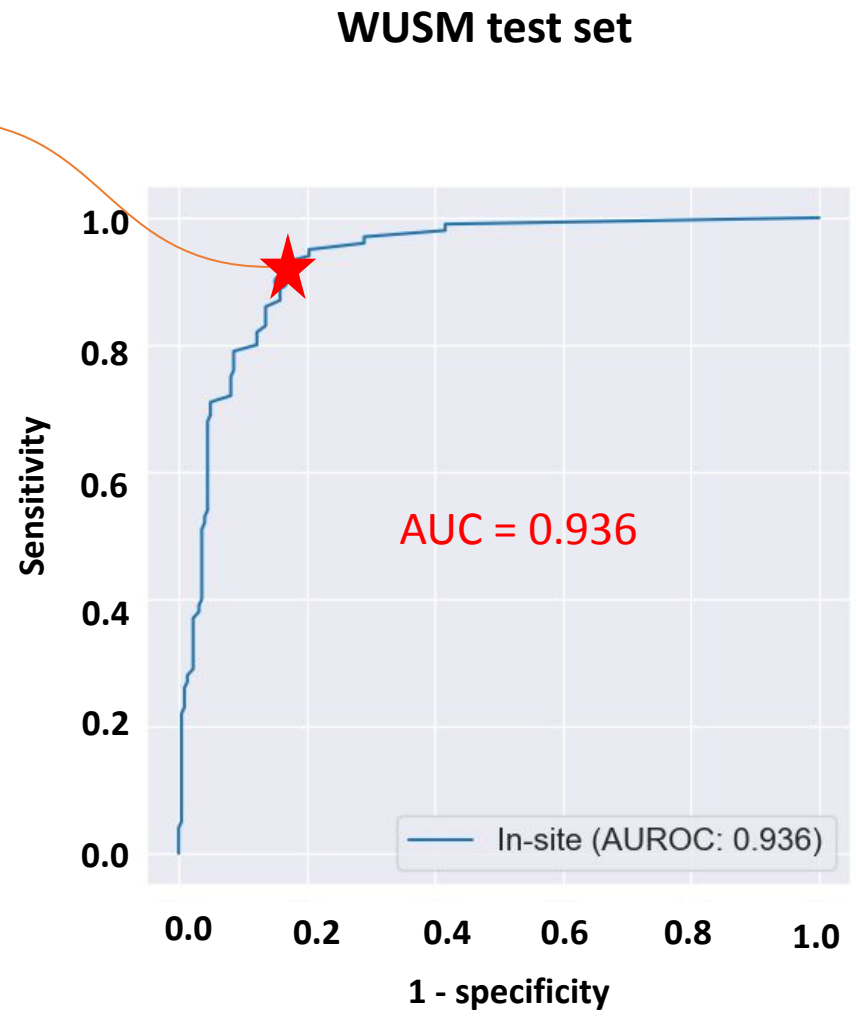



$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Sensitivity} = \frac{TP}{TP+FN} \qquad \text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Specificity} = \frac{TN}{TN+FP} \qquad \text{F1 score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

# Performance of the XGBoost Model in WUSM Test Set

Sensitivity = 0.900
Specificity = 0.842



# Does the ML Model Work better than the Calcium+PTH Algorithm?

- Yes!
- Using WUSM as an example, if we only use total calcium and intact PTH results available at the PTHrP order to predict the PTHrP normalcy, AUROC of an XGBoost model would be 0.762, and specificity would be 0.471 when sensitivity is sent to 0.900. The predictive performance is remarkably worse compared to our XGBoost mode incorporating other laboratory tests.

# Interpretability of the PTHrP Model

PERSPECTIVE

https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

**Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**

The way forward is to design models that are inherently interpretable.

Annals of Internal Medicine®

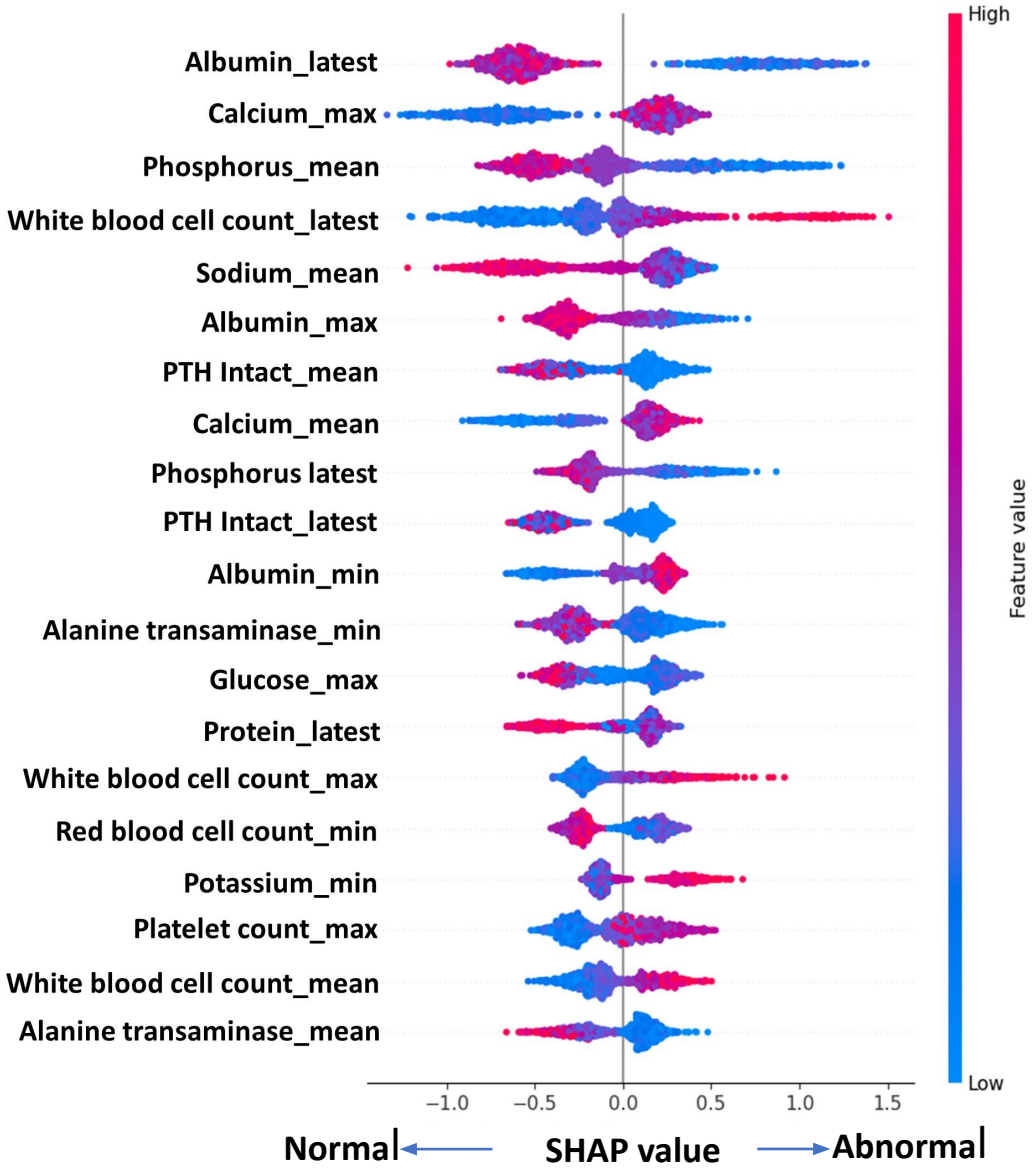




IDEAS AND OPINIONS | 7 JANUARY 2020

**Should Health Care Demand Interpretable Artificial Intelligence or Accept "Black Box" Medicine?**

Fei Wang, PhD; Rainu Kaushal, MD, MPH; Dhruv Khullar, MD, MPP



# Evaluation of ML Model Generalizability


Volume 68, Issue 3
March 2022

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

Shannon Haymond ✉, Stephen R Master

*Clinical Chemistry*, Volume 68, Issue 3, March 2022, Pages 392–395,
https://doi.org/10.1093/clinchem/hvab272

**Published:** 10 January 2022 **Article history** ▾

Generalizability is the ability of a ML model to perform well on independent datasets collected from different geographic or demographic populations or different hospital settings.


Volume 69, Issue 6
June 2023

JOURNAL ARTICLE

**Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group**

Stephen R Master ✉, Tony C Badrick, Andreas Bietenbeck, Shannon Haymond ✉

*Clinical Chemistry*, hvad055, https://doi.org/10.1093/clinchem/hvad055

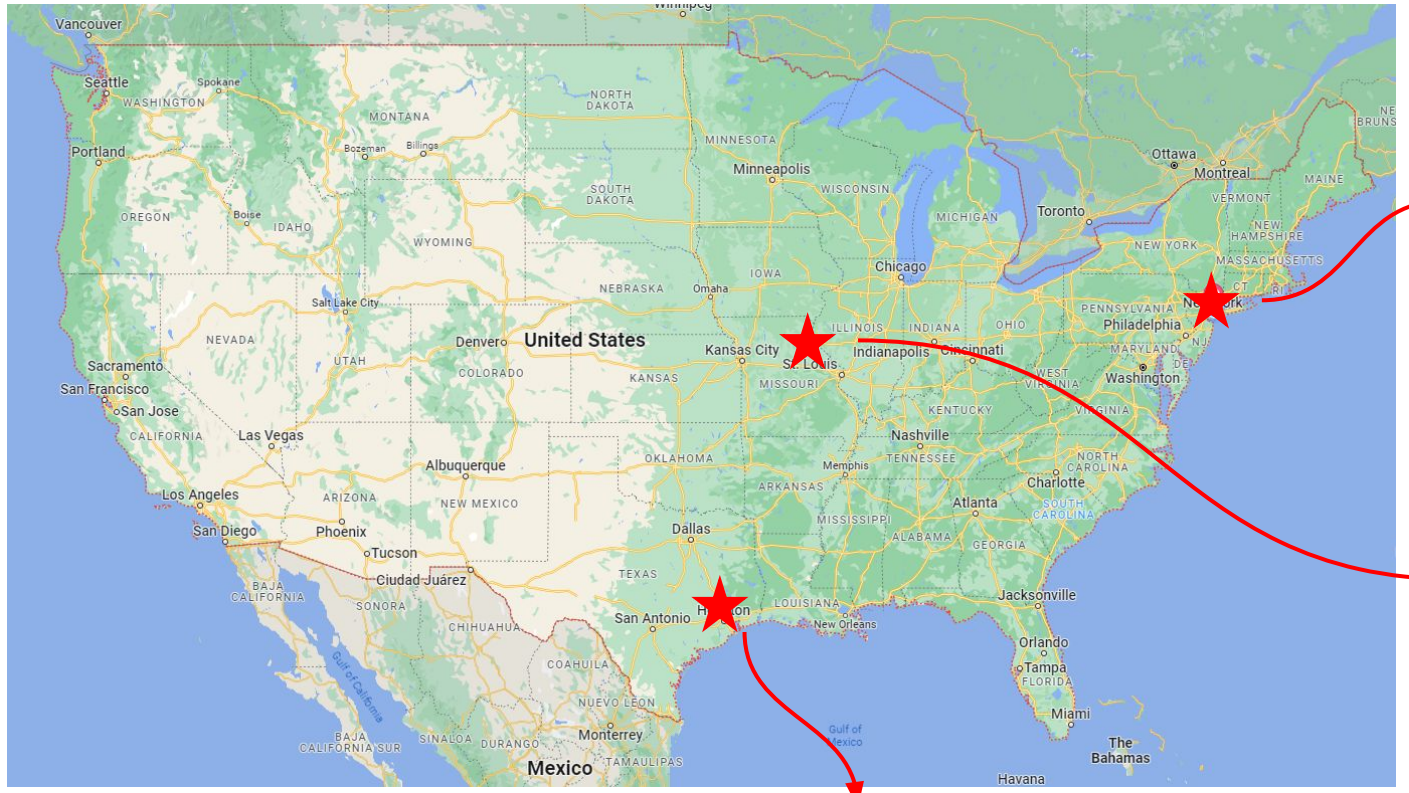
**Published:** 30 May 2023 **Article history** ▾

*Recommendation #15: Verify generalizability, particularly when applying a model outside its original training context.*

# Factors that Affect Model Generalizability

- Patient demographic characteristics
- Geographic features
- Instrument platforms
- Sample handling protocols and other pre-analytical factors
- Testing methodologies
- Send-out laboratories

# Model Evaluation – External Datasets




<u>Weill Cornell Medicine (WCM):</u>
1101 PTHrP orders from 2017 to 2022
PTHrP positive rate 16.9%
Send-out lab: Quest Diagnostics

<u>Washington University School of Medicine in St. Louis (WUSM)</u>
1330 PTHrP orders from 2012 to 2022
Positive rate 17.5%
Send-out lab: Mayo Clinic Laboratories

<u>University of Texas M.D. Anderson Cancer Center (MDA)</u>
1090 PTHrP orders from 2021-2022
PTHrP positive rate 23.9%
Send-out lab: Mayo Clinic Laboratories

# Directly Applying the PTHrP Model to Independent External Datasets

- When the ready-made model was directly applied “as-is” to the two independent datasets, its performance moderately deteriorated in MDA but substantially deteriorated in WCM.

| Sites | AUROC | Specificity given sensitivity = 0.900 | Precision (or PPV) given sensitivity = 0.99 |
|---|---|---|---|
| WUSM | 0.936 | 0.842 | 0.488 |
| MDA | 0.838 | 0.633 | 0.542 |
| WCM | 0.737 | 0.441 | 0.269 |

- The performance drop was due to the shift of data distribution from the original dataset to the new dataset.

# Maximum Mean Discrepancy

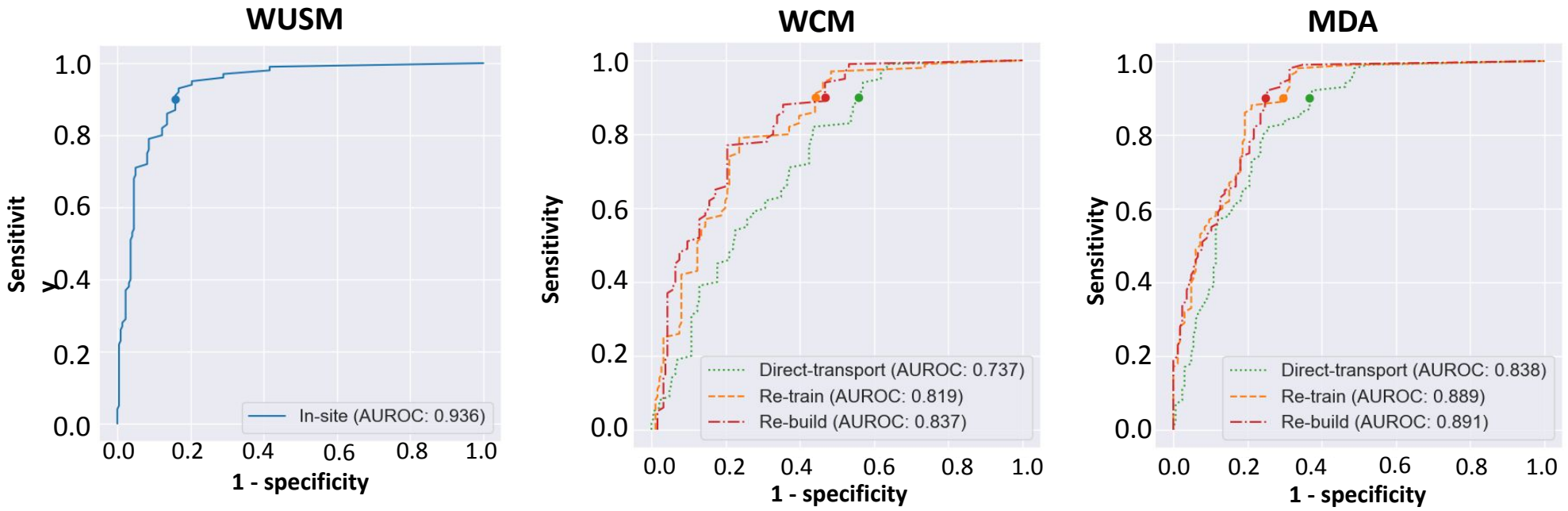
- MMD quantifies the degree of distribution shift between two datasets.
- A higher MMD between each pair of datasets indicates a greater distribution shift, which leads to a lower AUROC.

| Training site | Testing site | Maximum mean discrepancy | AUROC | ΔAUROC |
|---|---|---|---|---|
| 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 |

- MMD could be used to predict performance deterioration of ML models when transported to external sites. While there is not a specific MMD threshold to ensure successful generalization, calculating MMD can facilitate the adoption process of ML models.

# Strategies to Improve Model Performance

- **Strategy 1:** Re-training the XGBoost model using site-specific data with the same model architecture, feature sets, and hyperparameters
- **Strategy 2:** Re-building the model using site-specific data including feature selection, hyperparameter tuning and model parameter learning



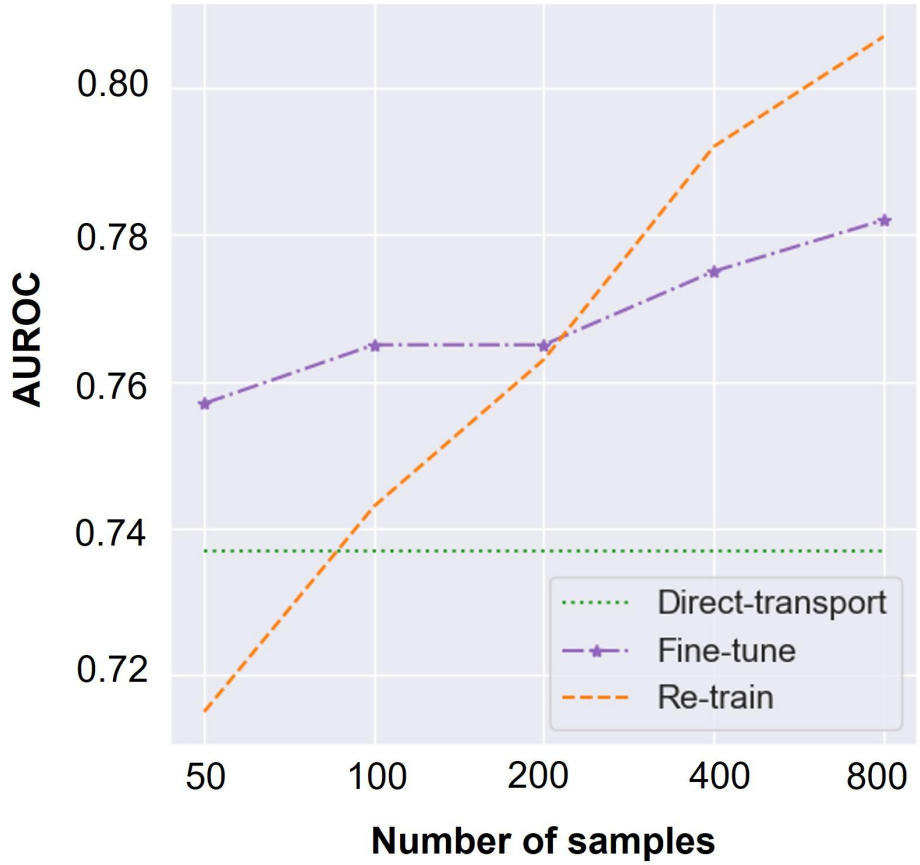
# Model Adaptation in External Datasets

| Method | AUROC | Specificity given sensitivity = 0.900 | Precision | Accuracy |
|---|---|---|---|---|
| **Testing: WUSM** | | | | |
| In-site test | 0.936 | 0.842 | 0.488 | 0.823 |
| **Testing: WCM** | | | | |
| Direct-transport the model | 0.737 | 0.441 | 0.269 | 0.787 |
| Re-train the model | 0.819 | 0.559 | 0.373 | 0.756 |
| Re-build the model | 0.837 | 0.532 | 0.406 | 0.787 |
| **Testing: MDA** | | | | |
| Direct-transport the model | 0.838 | 0.633 | 0.542 | 0.784 |
| Re-train the model | 0.889 | 0.705 | 0.505 | 0.766 |
| Re-build the model | 0.891 | 0.753 | 0.536 | 0.789 |

**Take home message**: When a ready-made model cannot be directly transported to external datasets due to the shift of data distribution, some local customization strategies can be utilized to improve model performance, such as re-training or re-building the model using site-specific data.

# What If a Hospital Has Limited Laboratory Data to Re-Train the Model?

- We explored a model fine-tuning strategy in which the ready-made model is applied to hospitals with limited training data (low-resource scenarios).
- The fine-tuning strategy performed best when the amounts of available samples were relatively small (< 200). However, when the number of available samples exceeded 200, model re-training appeared to be a better option.



# Model Bank


- A trusted authority, e.g. an NIH archive or repository, may serve as the “bank” storing models trained in different sites.
- All necessary metadata associated with each model should also be appropriately recorded following specified reporting guidelines.
- Institutes interested in deploying the model may select one or several models, deploying them either directly or with suitable fine-tuning.
- Institutes can also build their own model and deposit to the model bank following principles and instructions, which can continuously improve the model and make it more robust and generalizable in practice.

# Path to Model Implementation

- We are working on a pilot study to prepare for model deployment.

Clinicians attempt to order a PTHrP test;
Or clinicians order a PTH + calcium panel

Trigger the ML model to calculate a score

High Score → Indicating a high likelihood of abnormal PTHrP result. It suggests ordering PTHrP test to direct a focused cancer search

Low Score → Indicating a high likelihood of normal PTHrP result. It reminds clinicians to check primary hyperparathyroidism first.

- Determine the threshold for each scenario.
- Evaluate model fairness.
- Analyze model's performance in a prospective patient cohort.
- Collaborate with clinical teams to evaluate the model's real-world clinical utility.

# Summary

- Machine learning holds a great potential to improve laboratory efficiency and reduce pre-pre-analytical errors.
- The PTHrP model can potentially complement the current workup algorithm by identifying inappropriate PTHrP orders, and thus facilitating automation of the decision-making process, improving test utilization and laboratory stewardship.
- It will also help identify patients who need PTHrP testing and facilitate early cancer detection.
- Implementing a ML model in LIS/EHR system is still challenging and requires consensus among experts in our field.

# Acknowledgement


Weill Cornell Medicine

- Dr. Fei Wang
- Dr. Yu Hou
- Dr. Hao Zhang
- Dr. Chengxi Zang
- Dr. Weishen Pan
- Yingheng Wang
- Dr. Amy Chadburn
- Dr. Zhen Zhao
- Dr. Melissa Cushing

Cleveland Clinic

- Dr. Daniel D. Rhoads

GW School of Medicine & Health Sciences

- Dr. Jorge Sepulveda

Washington University School of Medicine in St. Louis

- Dr. Mark A. Zaydman

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

- Dr. Qing H. Meng