Original Article

Machine learning pipelines developed for the prediction of cancelation of inappropriate parathyroid hormone-related peptide orders demonstrate poor performance in predicting provider behavior

Nicholas C. Spies^, Christopher W. Farnsworth, Ronald Jackups, Mark A. Zaydman

Department of Pathology, St. Louis School of Medicine, Washington University, St. Louis, MO, USA

Contributions: (I) Conception and design: All authors; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: MA Zaydman; (V) Data analysis and interpretation: NC Spies; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Nicholas C. Spies, MD. Department of Pathology, St. Louis School of Medicine, Washington University, 660 S Euclid, St. Louis, MO 61137, USA. Email: nspies@wustl.edu.

Background: Quantification of circulating parathyroid hormone-related peptide (PTHrP) aids in the diagnosis of humoral hypercalcemia of malignancy. However, utilization of this test in the setting of low pre-test probability leads to false positive results, unnecessary follow-up testing, and patient anxiety. As part of an initiative to improve laboratory utilization, all PTHrP orders at our institution are reviewed by a laboratory medicine resident (LMR), who contacts the ordering physician when an order is deemed to have low utility. This review process is time- and labor-intensive, and may sow discontent with providers who feel they are being "second guessed". We sought to apply machine learning to automate this review process and minimize futile LMR interventions.

Methods: Retrospective, first-time PTHrP orders from 2019 to 2022 (n=1,144) were extracted from the laboratory information system of a single healthcare system. The dataset was partitioned into an 80:20 split between training and testing sets. XGBoost models were trained to predict order cancelation and order result, using laboratory data available at the time of the PTHrP order as features. After training and cross-validation, the models were applied to the held out test set and performance was evaluated using area under the receiver operating characteristic curve (AUC_{ROC}).

Results: Six hundred and forty-two (56%) PTHrP orders were ordered on patients without a recently suppressed PTH (<25 pg/mL), while 467 (41%) were placed on patients without recent hypercalcemia (>11 mg/dL). Of these, 450 were not canceled and only 9 (2%) were positive. The model trained to predict whether a PTHrP order was completed or canceled demonstrated little discriminatory power, with an AUC_{ROC} of 0.64 [95% confidence interval (CI): 0.59–0.68]. However, a model trained using the same pipeline to instead predict the PTHrP result demonstrated an AUC_{ROC} of 0.85 (95% CI: 0.82–0.87).

Conclusions: The performance difference between the models trained on the two different targets suggests that the physician's willingness to cancel in response to the LMR-driven intervention may be unrelated to the diagnostic value of the test in the context of other laboratory data.

Keywords: Laboratory utilization; parathyroid hormone-related peptide (PTHrP); machine learning (ML); hypercalcemia; medical decision making

Received: 04 February 2023; Accepted: 26 October 2023; Published online: 30 October 2023.
doi: 10.21037/jlpm-23-9
View this article at: https://dx.doi.org/10.21037/jlpm-23-9

^ ORCID: 0000-0002-5873-351X.

Introduction

Background

Quantification of circulating parathyroid hormone-related peptide (PTHrP) aids in the diagnosis of humoral hypercalcemia of malignancy (HHM), a common cause of malignancy-related hypercalcemia (1). PTHrP secreted by tumor cells drives the release of calcium from bone into serum, causing elevated calcium and secondary suppression of intact parathyroid hormone (PTH). Elevations in PTHrP may aid in the diagnosis of HHM, but PTHrP testing is often ordered in settings with low pre-test probability, such as in patients without a known malignancy (2).

At our institution, PTHrP requests are fulfilled by sending whole blood promptly to a reference laboratory for measurement, with a typical turn-around time of 4–5 days. Over-utilization of laboratory testing leads to increases in healthcare costs and absolute increases in the number of false positive results (3). These false-positives, in the setting of PTHrP testing, can lead to expensive, invasive, and anxiety-provoking "tumor hunts" (4).

Rationale and knowledge gap

Our institution has highlighted PTHrP testing as an opportunity to improve laboratory stewardship. To accomplish this, all orders for PTHrP are reviewed by a laboratory medicine resident (LMR), to check for evidence of recent elevations in serum calcium, a suppressed PTH, and a known or likely malignancy. If these conditions are not met, the LMR reaches out to the ordering provider and explains the initiative, with patient-specific justification for why PTHrP testing may not be indicated. However, this manual intervention is time-consuming and labor-intensive. The intervention also comes at an inopportune time (after the physician has finished seeing the patient), and through a suboptimal channel (a "cold" phone call from an often-unfamiliar phone number). These inefficiencies offer an opportunity for improvement in the form of clinical decision support (CDS), but a blanket CDS alert that triggers on all PTHrP orders contributes to alarm fatigue.

Machine learning (ML) techniques offer the potential to leverage complex relationships between high dimensional data to make inferences. Supervised classification refers to the paradigm of using a fully labelled set of training data to predict a binary outcome. In this work, we train supervised models on the complex input space of a patient's prior laboratory results to make inferences on the classification task of whether a PTHrP order is completed and if the measured result is abnormal.

Prior work has demonstrated the utility of multivariate modeling in the identification of laboratory testing that may provide low clinical utility (1), the effectiveness of a patient-specific approach to reducing over-utilization through provider education (2), and the use of recent laboratory results to predict the results of PTHrP testing (3,4). However, the combination of these ideas is relatively unexplored, especially as it pertains to the automation of these patient-specific interventions.

Objective

The objective of this study was to develop an ML pipeline for the prediction of PTHrP test cancelation using only laboratory data available at the time of order, with the ultimate goal being the improvement of improving the efficiency of our laboratory utilization initiative by highlighting orders that are likely to be canceled. We present this article in accordance with the TRIPOD reporting checklist (available at https://jlpm.amegroups.com/article/view/10.21037/jlpm-23-9/rc).

Highlight box

Key findings

- Physician compliance with a laboratory utilization management intervention is less predictable than the laboratory result itself, suggesting a lack of effectiveness of the current intervention in practice.

What is known and what is new?

- Parathyroid hormone-related peptide testing is often over-utilized or mis-ordered in the work-up of hypercalcemia. Clinical workflows aimed at improving this utilization are often time- and labor-intensive. We sought to automate this intervention through the application of a machine learning pipeline, but found laboratory data alone was not sufficient in predicting physician behavior, even if it is sufficient to predict testing results.

What is the implication, and what should change now?

- These findings highlight an interesting discordance between the biological information available to a clinician and their decision to reconsider a test order, while also reaffirming the importance of developing effective clinical decision support.

Methods

The study was conducted in accordance with the

Declaration of Helsinki (as revised in 2013). The study was performed after approval by Washington University Institutional Review Board (#202202087). The individual consent for this retrospective analysis was waived.

Aggregating clinical data

Laboratory data was extracted from the Cerner Millennium (Cerner Corporation, Kansas, USA) laboratory information system (LIS) of a single healthcare system for all inpatients and outpatients with an order for PTHrP testing between 2019 and 2022. This date range was chosen to coincide with a transition in the method for measuring PTHrP described in greater detail below. Features included: patient identifiers; order and result times; result values, units, and reference intervals; and additional clinical metadata such as clinical service and location for the encounter.

Inclusion/exclusion criteria

For each patient, only the first PTHrP order or result was considered, and all other laboratory results for these patients prior to the ordering time for this first result were included. No exclusion criteria were applied on the basis of patient demographics, as these data are not captured within the LIS.

Analyzing the laboratory data

After extraction from the LIS, data was loaded into R 4.1.2 (R Core Team, R Foundation for Statistical Computing) where all processing, analysis, and figure generation was performed. Using the {*targets*} (5) pipeline structure and the {*tidymodels*} (6) package set for modeling, ML pipelines were developed, validated, and assessed for performance.

Preparing the input data

Feature selection and engineering was performed using a knowledge-driven approach, rather than an unbiased one, such as recursive feature elimination, to reduce the risk of overfitting to spurious signal. We considered laboratory tests related to calcium homeostasis and high-volume routine tests (complete blood counts, metabolic panels) to be the most likely to contain valuable information. These features are summarized in Table S1.

All numerical laboratory features were centered and scaled to a mean of 0 and standard deviation of 1, without binning into categorical features. Features included the most recent, the minimum, and the maximum result for each analyte within 30 days of PTHrP order, as well as the medical service of the ordering provider, and the type of patient encounter. Any analyte for which over 25% of patients had no results within 30 days were removed, and the remaining missing values were imputed using a bagged decision tree classifier implemented in the tidymodels {*recipes*} (7) package. Identical procedures were followed for training the cancelation and result predictors. The input data was partitioned into an 80:20 training/testing split, stratified by the target outcome for both canceled tests and final results.

Building the ML pipeline

After initial screening of logistic regression, support vector machines, random forest, XGBoost (5), naïve Bayes, and multi-layer perceptron architectures, the XGBoost model was chosen for the final architecture based on the area under the receiver operating characteristic curve (AUC_{ROC}) (Table S2), with tunable hyperparameters of "*mtry*", "*learning rate*", and "*tree depth*". All other hyperparameters were left at their default values. Bayesian optimization was performed on a 10-fold cross-validation set and repeated ten times to find optimal tuning for these hyperparameters that maximized area under the precision-recall curve (AUC_{PR}). Final models were then fit onto the full training set using the best performing hyperparameters, then applied to the held-out test set. These predictions were used for the remainder of the performance assessment.

Clinical definitions and assay descriptions

PTHrP orders at our institution are fulfilled by sending ice-chilled, cold-centrifuged, K_2EDTA-preserved aliquots of patient plasma directly to Mayo Clinical Laboratories (Rochester, MN, USA), who performs a plate-based chemiluminescent assay targeting 1-86 PTHrP (6). A key benefit of this assay worth noting is that it no longer requires the phenylalanyl-prolyl-arginyl chloromethyl ketone (PPACK) tube that its predecessor was collected in. This procedural change motivated the exclusion of all results prior to the assay transition to minimize the effect of collection protocol on the decision to cancel the test.

Relevant laboratory thresholds were defined through a combination of literature review and clinical stakeholder

engagement. PTH suppression was defined as less than 25 pg/mL to reflect the findings of Szymanski *et al.* (7), while calcium elevation was defined as greater than 11 mg/dL after discussion with our endocrinology colleagues. PTHrP results were classified as normal or abnormal using the Mayo-reported clinical decision limit of 4.2 pg/mL for the newer assay, which began in 2019, motivating the exclusion of results prior to 2019 from the study set. This newer method, and definition of the clinical decision limit, are described in greater detail by Ashrafzadeh-Kian *et al.* (6).

Statistical analysis

All statistical tests were performed using R 4.2.1 (R Core Team, R Foundation for Statistical Computing). A significance threshold of 0.05 was used for all statistical tests. Chi-squared test was performed for the cancelation table, while Fisher's exact test was performed for the results table due to low numbers within the positive group. Summary statistics for the performance of the models were calculated using the default definitions of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), AUC_{ROC} and AUC_{PR} from the tidymodels {*yardstick*} package, where the positive event was defined as the less frequent class.

Reproducibility and replicability

The dataset used in this study was provided in anonymized form for the first annual AACC Kaggle Competition (8). The input data can be downloaded from the competition's Data page at https://www.kaggle.com/competitions/aacc-2022-predicting-pthrp-results/data, and the code used to develop and test the models can be found in notebook form on the Code page at https://www.kaggle.com/code/nickspies/aacc-kaggle-baseline-xgboost.

Results

PTHrP testing is often ordered in patients without suppressed PTH

A total of 1,144 patients' PTHrP orders were included in the study set, summarized in *Figure 1*. Five hundred and eighty-four patients (51%) were outpatient at the time of the collection, 552 (48%) were inpatient, and eight were unknown.

Figure 1A demonstrates a schematic overview of the LMR intervention workflow that was in place for the entirety of the study period. Six hundred and forty-two (56%) of the PTHrP orders were placed on patients without a recently suppressed PTH, while 467 (41%) were placed on patients without a recently elevated calcium result. *Figure 1B* summarizes these results. Patients with recently suppressed PTH or recently elevated calcium results were more likely to have their orders completed than those without (P=0.01 and 0.04, respectively).

Of the 847 orders that were not canceled, nine could not be performed due to issues with specimen integrity. Of the 838 completed orders, 88 (11%) were positive. These counts are summarized in *Figure 1C*. Patients with recently suppressed PTH or elevated calcium were significantly more likely to have an abnormal PTHrP results (P<0.001 for each).

ML models predict PTHrP results with greater success than order cancelation

Figure 2 demonstrates the summary of the performance of ML models for predicting PTHrP positivity or order cancellation in response to LMR intervention. The model trained to predict PTHrP positivity had a Matthews correlation coefficient (MCC) of 0.52 (95% CI: 0.39–0.62) and an AUC_{ROC} of 0.85 (95% CI: 0.82–0.87), while the model trained to predict whether the LMR-driven intervention would lead to a cancelation of the PTHrP order showed an MCC of 0.21 (95% CI: 0.02–0.39) and an AUC_{ROC} 0.64 (95% CI: 0.59–0.68).

Discussion

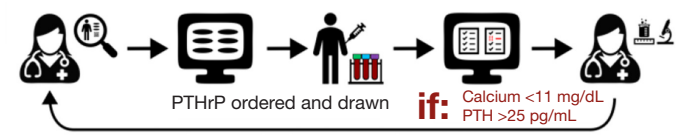
Key findings

In this work we sought to develop an ML pipeline that could predict cancelable orders by LMR intervention. However, these models exhibited poor predictive ability despite models trained on the same data successfully predicting which results would be abnormal. Altogether, this highlights a discrepancy between the biological information present in the laboratory data and the decision to resist efforts to cancel an inappropriate PTHrP order.

Strengths and limitations

To our knowledge, this study is the first of its kind to compare the potential of ML for improving the utilization

A



B

Characteristic	Completed, n=847[1]	Canceled, n=297[1]	P value[2]
PTH			0.005
Suppressed (<25 pg/mL)	397 (47%)	105 (35%)	
Normal/high	352 (42%)	141 (48%)	
Not measured	98 (11%)	51 (17%)	
Calcium			0.036
Elevated (>11 mg/dL)	519 (61%)	158 (53%)	
Normal/low	293 (35%)	120 (40%)	
Not measured	35 (4%)	19 (6%)	

[1] n (%)
[2] Pearson's Chi-squared test

C

Characteristic	Normal, n=750[1]	Abnormal, n=88[1]	P value[2]
PTH			<0.001
Suppressed (<25 pg/mL)	318 (42%)	79 (90%)	
Normal/high	341 (46%)	6 (7%)	
Not measured	91 (12%)	3 (3%)	
Calcium			<0.001
Elevated (>11 mg/dL)	433 (58%)	85 (97%)	
Normal/low	285 (38%)	2 (2%)	
Not measured	32 (4%)	1 (1%)	

[1] n (%)
[2] Fisher's exact test

Figure 1 Overview of intervention and result summary. (A) The workflow for LMR intervention in the setting of low pre-test probability. (B) Summary of PTHrP order resolutions. (C) Summary of PTHrP results. PTHrP, parathyroid hormone-related peptide; PTH, parathyroid hormone; LMR, laboratory medicine resident.

of PTHrP with that of the performance in predicting the PTHrP result. In doing so, we highlight that the ability to predict a result does not necessarily correlate with the more operationally meaningful outcome of successful utilization interventions. We also provide an anonymized version of the data set and code to encourage the replicability and reproduction of this approach.

There are several limitations of this study to consider. First, the data used to train these models is limited to information that can be extracted from the LIS. This excludes potentially informative features such as diagnosis codes to explain alternative causes of hypercalcemia, codes associated with malignancies known to secrete PTHrP, and more. Second, the performance of the model for predicting test cancelation falls below what would be considered clinically useful, suggesting the need to explore alternative options for improving the efficiency of our clinical workflow. Additionally, the prior protocol for

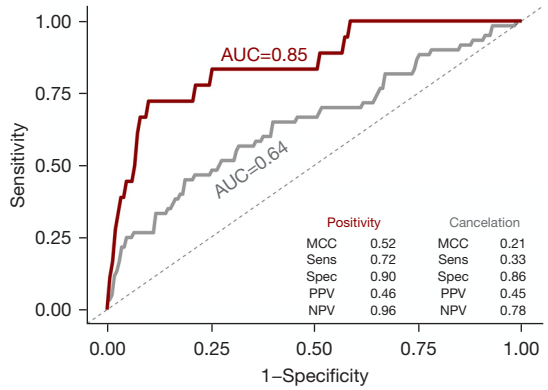


Figure 2 Receiver operating characteristic curves for the PTHrP results (red) and order resolutions (gray). Summary statistics when the models were applied to the held-out test set are shown in the bottom right. AUC, area under the curve; MCC, Matthews correlation coefficient; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; PTHrP, parathyroid hormone-related peptide.

specimen collection when a PTHrP is ordered required prior planning and the coordination of collection within a dedicated PPACK tube. While our new protocols no longer require this planning, and the dataset included in these analyses come from after this transition, it may be the case that physicians are hesitant to cancel their test orders due to an outdated assessment of the effort needed to collect the sample. Finally, the incorporation of laboratory information may be able to predict the ultimate PTHrP result to some extent, but the degree of performance that a clinical provider would deem sufficient to convincingly forgo testing is an opportunity for future research. It is likely that no single performance threshold will be universally applicable to all patients and providers, and how the clinical providers incorporate the model predictions into their final decision-making process remains to be seen.

Comparison to similar researches

Yang *et al.* (4) describe a similar ML approach for predicting the result of PTHrP orders using laboratory data available at the time of order which finds similar performance to our model. However, this work is the first to our knowledge that applies ML models to predict the success of a resident-driven intervention to improve PTHrP test utilization.

Explanation of findings

We believe that the most likely explanations for this observation are likely a combination of convenience and anchoring bias (9). Providers assessing a patient's hypercalcemia may elect to test for all possible causes upfront, including PTHrP, because a sequential testing scheme may inconvenience the patient with multiple blood draws. Furthermore, given that there is a time lag between the decision to order PTHrP and the phone call from the LMR describing the utilization improvement initiative, the default decision may be to proceed with the original plan rather than revisit prior work.

The 'five rights' of CDS is a paradigm that has been proposed for designing effective interventions (9,10). These include: the right time, right person, right information, right channel, and right format. While the LMR-based process described in this study often delivers the right information to the right person, it may be delivered at an inopportune time (well after the provider has moved on to other clinical duties), through an inconvenient channel and improper format (a phone call from the LMR). The current work may help to improve this approach by moving the intervention proximal to the time of order by integrating our models into the computerized provider order entry system. Doing so may simultaneously improve PTHrP utilization and lessen the burden of the current CDS effort on the LMR and ordering provider.

Implications and actions needed

In conclusion, what began as an attempt to automate a clinical workflow through the use of ML brought valuable insight into the nature of ordering practices surrounding PTHrP testing but did not result in an ML solution that performed up to the standard for clinical application. Future efforts will attempt to capitalize on these insights to improve the utilization surrounding this testing.

Conclusions

While ML models are capable of predicting the result of PTHrP testing, they are incapable of predicting the success of a utilization-focused initiative to cancel orders with low pretest probability, highlighting a discordance between the information content present in laboratory values at the

time of order and the information being used to drive these clinical decisions in practice.

Acknowledgments

We would like to thank Dr. Ann Gronowski, Professor at Washington University School of Medicine, for her invaluable expertise regarding PTHrP physiology, testing patterns, and intervention opportunities.
Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the Guest Editor (Lee Schroeder) for the series "Data-Driven Laboratory Stewardship" published in *Journal of Laboratory and Precision Medicine*. The article has undergone external peer review.

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://jlpm.amegroups.com/article/view/10.21037/jlpm-23-9/rc

Data Sharing Statement: Available at https://jlpm.amegroups.com/article/view/10.21037/jlpm-23-9/dss

Peer Review File: Available at https://jlpm.amegroups.com/article/view/10.21037/jlpm-23-9/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jlpm.amegroups.com/article/view/10.21037/jlpm-23-9/coif). The series "Data-Driven Laboratory Stewardship" was commissioned by the editorial office without any funding or sponsorship. CWF has received grants/contracts from Abbott, Roche, Siemens, Beckman Coulter, Biomerieux, Cepheid, Sebia, and Binding Site. MAZ has received contracts from Biomerieux. CWF and MAZ have received honoraria from ADLM (both) and API (MAZ). The authors have no other conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was performed after approval by Washington University Institutional Review Board (#202202087). The individual consent for this retrospective analysis was waived.

References

1. Xu S, Hom J, Balasubramanian S, et al. Prevalence and Predictability of Low-Yield Inpatient Laboratory Diagnostic Tests. JAMA Netw Open 2019;2:e1910967. Erratum in: JAMA Netw Open 2019;2:e1914190.
2. Elnenaei MO, Campbell SG, Thoni AJ, et al. An effective utilization management strategy by dual approach of influencing physician ordering and gate keeping. Clin Biochem 2016;49:208-12.
3. 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.
4. Yang HS, Pan W, Wang Y, et al. Generalizability of a Machine Learning Model for Improving Utilization of Parathyroid Hormone-Related Peptide Testing across Multiple Clinical Centers. Clin Chem 2023;69:1260-9.
5. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM; 2016:785-94. Available online: https://dl.acm.org/doi/10.1145/2939672.2939785
6. 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.
7. Szymanski JJ, Otrock ZK, Patel KK, et al. Incidence of humoral hypercalcemia of malignancy among hypercalcemic patients with cancer. Clin Chim Acta 2016;453:190-3.
8. Predicting PTHrP Result - AACC 2022 Annual Meeting. Kaggle. 2022. Available online: https://kaggle.com/competitions/aacc-2022-predicting-pthrp-results

9. Osheroff JA, Teich J, Levick D, et al. editors. Improving outcomes with clinical decision support: an implementer's guide. 2nd edition. Chicago, IL, USA: HIMSS Publishing; 2012:323.
10. Alós-Ferrer C, Hügelschäfer S, Li J. Inertia and Decision Making. Front Psychol 2016;7:169.

doi: 10.21037/jlpm-23-9
Cite this article as: Spies NC, Farnsworth CW, Jackups R, Zaydman MA. Machine learning pipelines developed for the prediction of cancelation of inappropriate parathyroid hormone-related peptide orders demonstrate poor performance in predicting provider behavior. J Lab Precis Med 2023;8:29.

Supplementary

Table S1 Laboratory features included in the models

Laboratory values
Sodium
CO2
Calcium
WBC
Albumin
Alkaline phosphatase
Ionized calcium
Hemoglobin
25 Vitamin D
Chloride
Creatinine
Glucose
AST
Total protein
Phosporus
Intact PTH
Hematocrit
1,25 Vitamin D
Potassium
BUN
Anion gap
ALT
Total/direct bilirubin
Magnesium
TSH
Platelets
Urine calcium

Table S2 Performance of all screened model architectures

Model	Result		Cancelation	
	Median AUC_{ROC}	95% CI	Median AUC_{ROC}	95% CI
XGBoost	0.85	0.82–0.87	0.64	0.59–0.68
Random forest	0.85	0.82–0.86	0.61	0.56–0.70
Multi-layer perceptron	0.84	0.81–0.86	0.62	0.58–0.67
Naïve Bayes	0.81	0.79–0.82	0.61	0.61–0.71
Support vector machine	0.79	0.78–0.80	0.58	0.52–0.66
Logistic regression	0.79	0.75–0.82	0.61	0.57–0.66

AUC_{ROC}, area under the receiver operating characteristic curve; CI, confidence interval.