Utility of disease probability scores to guide decision-making during screening for phaeochromocytoma and paraganglioma: a machine learning modelling cross sectional study

Christina Pamporaki,[a,m,**] Georg Pommer,[b,m] Ioannis D. Apostolopoulos,[c] Angelos Filippatos,[d] Mirko Peitzsch,[e] Hanna Remde,[f] Georgiana Constantinescu,[a] Annika M. A. Berends,[g] Matthew A. Nazari,[h] Felix Beuschlein,[i,j,k] Martin Fassnacht,[f] Aleksander Prejbisz,[l] Karel Pacak,[h] and Graeme Eisenhofer[a,*]

[a]Department of Internal Medicine III, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden 01307, Germany
[b]Institute of Clinical Genetics, University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden 01307, Germany
[c]Department of Energy Systems, University of Thessaly, Gaiopolis Campus, Larisa 41500, Greece
[d]Machine Design Laboratory, Department of Mechanical Engineering & Aeronautics, University of Patra, Patras, Greece
[e]Institute for Clinical Chemistry and Laboratory Medicine, University Hospital and Medical Faculty Carl Gustav Carus, Technische Universität Dresden, Dresden 01307, Germany
[f]Division of Endocrinology and Diabetes, Department of Internal Medicine I, University Hospital, University of Würzburg, Würzburg 97082, Germany
[g]Department of Endocrinology, University Medical Center, Groningen, the Netherlands
[h]Section on Medical Neuroendocrinology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, United States
[i]Medical Clinic IV, University Hospital, Ludwig Maximilians-Universität, University Hospital of Munich, Germany
[j]Medical Clinic for Endocrinology, Diabetology, and Metabolism, UniversitätsSpital and University of Zurich, Zurich, Switzerland
[k]The LOOP Zurich-Medical Research Center, Zurich, Switzerland
[l]Department of Epidemiology, Cardiovascular Prevention and Health Promotion, National Institute of Cardiology, Warsaw, Poland

eClinicalMedicine 2025;82: 103181
Published Online xxx
https://doi.org/10.1016/j.eclinm.2025.103181

Summary

Background Interpretation of plasma metanephrines and methoxytyramine to assess likelihood of phaeochromocytoma/paraganglioma (PPGL) during screening can be challenging. This study (study period: 2021–2023) introduces new methods to select machine-learning (ML) models and evaluate derived probability-scores to better interpret laboratory results.

Methods ML models were trained and internally tested using data from 2046 patients with and without PPGL and according to several features: age, pre-test risk of PPGL, plasma metanephrines and methoxytyramine. External validation involved a second cohort of 1641 patients with and without PPGL. The study employed several processes to select and evaluate the best model: concordance of models with human intelligence; intra- and inter-laboratory variability in derived probability-scores; and comparison of scores of the selected model to predictions of ten clinical care specialists before and after provision of those scores.

Findings External validation established equally excellent diagnostic performance for all five best ML models according to areas under ROC curves (0.988–0.994) and balanced accuracies (0.958–0.981). Probability-scores of models, however, varied widely and were poorly correlated. The additional selection processes indicated an artificial-network model as a superior and more robust model than others. Predictions of disease likelihood by specialists, according to six categories from disease highly unlikely to disease clear, varied widely for individual patients. Within each of the six predictive categories, median probability-scores of the artificial-network model were 70-, 175-, 59-, 15-, 3.5- and 1.7-fold higher ($P < 0.0001$) in patients with than without PPGL. This superiority of probability scores over variable predictions by specialists remained evident after specialists were tasked to modify their predictions according to those scores.

*Corresponding author. Department of Medicine III, University Hospital Carl Gustav Carus at the TU Dresden, Fetscherstraße 74, Dresden D-01307, Germany.
**Corresponding author. Department of Medicine III, University Hospital Carl Gustav Carus at the TU Dresden, Fetscherstraße 74, Dresden D-01307, Germany.
E-mail addresses: Graeme.eisenhofer@ukdd.de (G. Eisenhofer), Christina.pamporaki@ukdd.de (C. Pamporaki).
[m]These authors contributed equally to this work.

Interpretation This study employed several novel processes to establish an ML model with probability-scores superior to predictions of disease likelihood by specialists. However, the negligible improvement in interpretations by specialists after provision of probability-scores indicates this alone is insufficient to improve decision-making.

Funding Deutsche Forschungsgemeinschaft.

Keywords: Phaeochromocytoma; Paraganglioma; Metanephrines; Machine learning; Clinical decision support system

Research in context

Evidence before this study

We searched PubMed from 2022 onwards using the terms (diagnosis OR diagnostic) AND (phaeochromocytoma OR paraganglioma) AND (machine learning OR artificial intelligence). Most publications involved surgical robotics or radiomics studies. Nevertheless, we identified five studies that combined omics and machine learning models to distinguish patients with different causes of hypertension, including phaeochromocytoma and paraganglioma. We also used combinations of terms (clinical chemistry OR laboratory medicine AND validation AND machine learning AND laboratory) with and without title restrictions [TITLE] for a wider search of practices for validating machine learning models in laboratory medicine. We identified two studies that validated machine learning models using prospective comparisons with interpretations by clinicians, two that integrated models in clinical decision support tools before validation, and a study where validation was secured by prospective evaluation in real-time clinical settings. Another study proposed a novel methodology for appropriate external validation of machine-learning models.

Added value of this study

Our study introduces novel approaches for validation of machine learning models, which using commonly available laboratory tests and other data we applied to the diagnosis and management of patients with phaeochromocytoma and paraganglioma. Apart from external validation of machine learning models in a large population that differed from the training cohort in terms of setting, baseline characteristics and outcomes, generalisability was further ensured by assessment of robustness according to measurements in large numbers of identical patient samples run by twelve different laboratories. Comparisons of model-derived disease probability scores with interpretations of data by ten clinical care specialists were also employed to develop and test machine learning models against human intelligence. As a further defining feature of the study, disease probability scores of the finally selected model were provided to the specialists to determine improvement in interpretations of disease likelihood. Although machine learning model-derived probability scores were superior for assessment of disease likelihood than interpretations by specialists, they failed to improve interpretations.

Implications of all the available evidence

The established machine-learning model has potential to assist decision-making according to steps in the diagnostic and management process for patients with suspected phaeochromocytoma and paraganglioma. Nevertheless, there is need for further interpretative support to strengthen ability of physicians to appropriately utilise data from machine learning models. Integration of machine learning probability scores within a clinical decision support system may offer a solution. The approaches we employed for testing and validation of machine learning models, which move beyond traditional metrics, have wider relevance for bringing other machine learning applications into daily routine clinical practice.

Introduction

Phaeochromocytomas and paragangliomas (PPGL) are catecholamine-producing tumours that can result in life-threatening cardiovascular consequences. As their clinical presentation is quite heterogeneous, diagnosis is crucially dependent on the biochemical work-up. Measurements of plasma free or urinary metanephrines, the O-methylated metabolites of catecholamines, are the recommended screening tests.[1] The plasma test, with inclusion of free methoxytyramine in the panel, provides several advantages over the urinary panel including superior diagnostic accuracy.[2]

Interpretation of biochemical test results, even with an accurate test, remains challenging. Low pre-test prevalence of the PPGL results in considerably more false-positive than true-positive results. Pre-test prevalence though varies. Pre-test probability of PPGL is lower in patients tested due to signs and symptoms (0.2–0.6%) than those tested due to adrenal or extra-adrenal incidentalomas or hereditary risk (4–7%).[3] Difficulties for clinicians to employ pre-test probability to better interpret diagnostic results are well established.[4–6] Need to consider results for three catecholamine metabolites and differences of each from cut-offs of

reference intervals that define positive versus negative results adds to difficulties to interpret laboratory results.[2] Such problems are particularly relevant for unclear cases with metabolite concentrations close to those cut-offs, which may reflect a host of preanalytical causes of false-positive versus true-positive results.[7–9] Occasional false-negative results for tumours that produce limited amounts of any single metabolite due to small size or a poorly differentiated phenotype are also a problem.[10]

Advances in computational power have led to the introduction of digital technologies to support decision-making in health care.[11,12] Machine learning (ML) is one digital approach, that compared to conventional statistics prioritizes prediction accuracy, making it more suitable for tasks where future outcomes are of primary concern. ML has been recently applied to distinguish patients with PPGL from those with other causes of hypertension[13] and stratify patients with PPGL according to metastatic risk.[14] Despite the clinical potential of ML applications, few have been deployed.[15,16] In part this may reflect reliance to select models on conventional metrics (e.g., accuracy during training and testing), without assessment of generalisability, clinical applicability, or robustness.[17,18]

It was not the objective of this study to improve upon test results for binary classification of whether patients do or do not have PPGL, but rather to establish ML models to assist with interpretation of results for plasma free metanephrines and methoxytyramine. For any application of ML-based models to facilitate interpretation of laboratory results, it is important that generated probability scores are not only meaningful for clinicians but can also be reproducibly generated within and among clinical centres. With the above considerations in mind, the present study had three objectives: 1) establish ML models to provide information about disease likelihood in patients tested for PPGL; 2) introduce optimal methods for selection of ML models that provide clinically meaningful and robust disease prediction; and 3) establish whether provision of ML-based probability scores to clinical specialists improves their interpretation of laboratory results.

Methods

Study overview and patients

The study period was from year 2021–2023 and involved several steps to generate and test ML models, then select the best model according to concordance with human intelligence and robustness of probability scores after repeated measurements of catecholamine metabolites in identical plasma aliquots within and among laboratories (Fig. 1). Predictions of the finally selected model were then compared to predictions by ten specialists over three interpretative phases.

To train and internally test ML models we employed data from 2047 patients with and without PPGL from six tertiary centers recruited into the Prospective Monoamine-Producing Tumor (PMT) study (https://

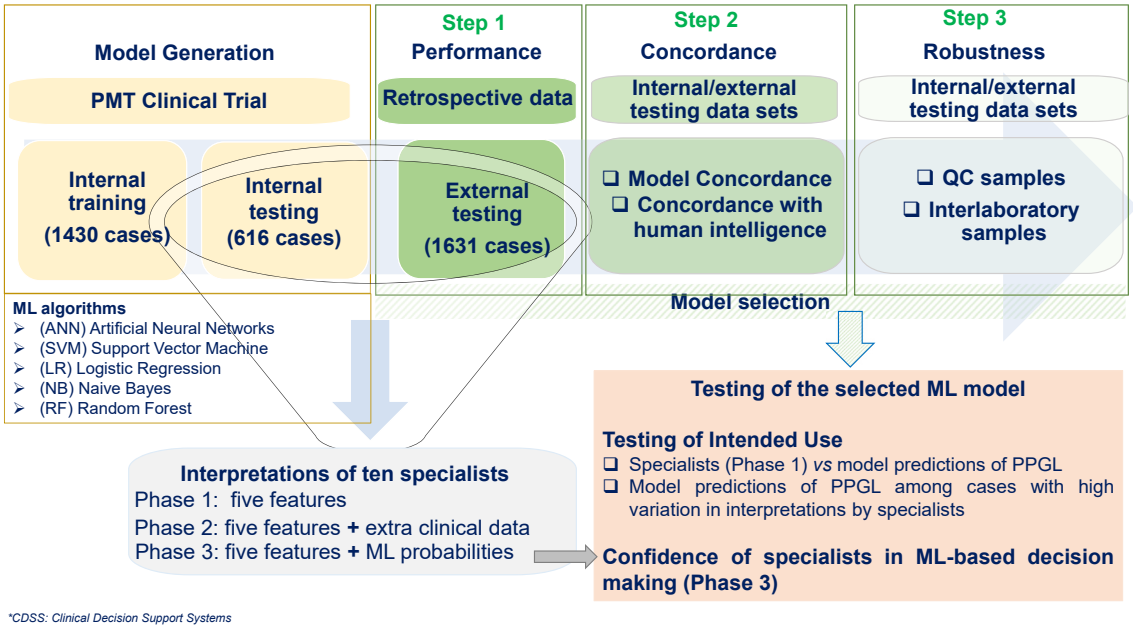

Fig. 1: Study work flow and "road map" for selection and evaluation of ML models. The steps include model development, statistical examination of model performance, concordance of artificial and human intelligence, examination of robustness according to within and between laboratory variability, and testing for intended use and interpretability.

pmt-study.pressor.org).[2] Criteria for disease confirmation and exclusion are detailed in the Appendix (p3). Models were then externally validated in a cohort of 1631 patients with and without PPGL enrolled under the PRESCRIPT trial at the University Medical Center of Groningen, the Netherlands, and an observational clinical study (Diagnosis, Pathophysiology and Molecular Biology of Phaeochromocytoma, 00-CH-0093) at the National Institutes of Health (NIH) in Bethesda, Maryland, USA. Protocols were approved by the local ethics committees and all patients provided signed informed consent for their participation, including parental consent for children (Appendix p3).

Clinical information included age, sex, a single screening determination of plasma concentrations of normetanephrine, metanephrine and methoxytyramine, as well as reasons for the screen. These included signs and symptoms of catecholamine excess, hereditary risk due to a pathogenic variant in a susceptibility gene, previous history of PPGL or presence of an incidental adrenal or extra-adrenal mass. Plasma normetanephrine, metanephrine and methoxytyramine were measured using liquid chromatography with electrochemical detection[19] or mass spectrometry.[20]

Model generation

The predefined features for supervised training of models are well defined predictors of PPGL and include risk group (low/high), plasma concentrations of normetanephrine, metanephrine and methoxytyramine. Patients were assigned to the low-risk group if tested due to signs and symptoms and to the high-risk group if tested due to hereditary risk, previous history of PPGL, or presence of an incidentaloma. As detailed in the Appendix (pp3-4), ML models were developed with Python programming language version 3.9.7 and trained using five algorithms: support vector machine (SVM), random forest (RF), naïve Bayes (NB), artificial neural networks (ANN) and logistic regression (LR). After data preparation and normalisation, the patient cohort was randomly split at a 7:3 ratio into training (1430 patients) and internal testing (616 patients) groups using the corresponding function of Scikit Learn's module in Python. ML models were then externally validated (1641 patients) to assess the predictive performance by several metrics: area under the receiver operating characteristic (ROC) curve, F1 score, sensitivity, specificity, accuracy, and balanced accuracy.

Interpretations by specialists

As outlined in Fig. 1 and detailed in the Appendix (p4-6), we invited ten clinical care specialists with expertise in PPGL to provide predictions of the likelihood of PPGL in internal and external test cohorts using six categories of classification: highly unlikely, unlikely, possible, likely, highly likely, and clear. Before review of the datasets, specialists were provided with definitions for each category (Appendix Box S1).

Specialists returned their predictions in three phases. Predictions by specialists in phase 1 were provided according to provision of information about age, sex, and plasma normetanephrine, metanephrine and methoxytyramine with corresponding upper cut-offs. After an interval of at least four weeks, all specialists received the same dataset, though scrambled with different identifiers and with additional clinical data relevant to pre-test risk (Appendix p6). They were again requested to return their predictions of disease likelihood. Finally, for phase 3 the specialists received the dataset with the five features together with their initial interpretations, supplemented this time by the probability scores of the selected best ML model. Specialists were again provided instructions on how to interpret data, including probability scores (Appendix Box S2, pp7-8), and were instructed to modify their initial interpretations if they considered it necessary.

Concordance with human intelligence

To determine how well probability scores for different ML models aligned with interpretations by specialists, we quantified agreement of probability scores with predictions of clinical care specialists selected according four criteria (Appendix p9). Agreement of models with the interpretations by those specialists was evaluated by relationships and concordance of digitised interpretations by specialists with ML probability scores (Fig. 1).

Model robustness

As summarised in Fig. 1 and detailed in the Appendix (p10), robustness of ML models was evaluated using two methods. We first assessed variability of probability scores calculated by each of the five models in sets of quality control (QC) samples used as part of routine quality assurance for mass spectrometric measurements of plasma catecholamine metabolites within a single laboratory. We also evaluated variability of probability scores calculated by each of the ML models using a dataset of results for measurements of metabolites in 100 identical plasma specimens sent to 12 different laboratories. As described previously,[21] these specimens from different patients covered a range of plasma metabolite concentrations from low, normal and pathological. Each laboratory employed a different mass spectrometric method for measurements.

Final model selection and testing for intended use

To select the final model, we compared the overall performance of all models according to rankings from one to five points for the poorest to the best model for all test steps. The model with the highest summed score was selected as the final model. Further evaluation of

the selected model included testing of its predictive performance in relation to predictions by specialists, as well as testing for improvement in interpretations by specialists after provision of probability scores (Phase 3).

Role of the funding source

The funder had no role in study conception, design, data collection, data analyses, interpretation, or conduct of the study.

Statistical analysis

Continuous variables are shown as medians (interquartiles) for non-normally distributed data and as means (standard deviations) for normally distributed variables. Comparisons of continuous parameters were by the Mann–Whitney U test. Categorical parameters were analysed using the chi-square test. To compare paired data of non-normally distributed parameters, we used the Wilcoxon Signed-Ranks test. Statistical analysis was with JMP pro statistical software package version 17. $P < 0.05$ was considered significant.

Results

Patient characteristics

Of relevance to generalisability of ML models, patients in the training/internal testing cohort showed significant differences in several characteristics compared to those in the external testing cohort (Supplementary Table S1).

Interpretations by specialists

For the combined internal and external testing datasets, diagnostic sensitivity and specificity for the three metabolites were 98.1% and 95.3% respectively. Interpretations of all of the 98.1% true-positive results were correct at phase 1 for four of the ten specialists: A, F, G and I (Supplementary Table S2). Among the other six specialists correct recognition of true-positive results was missed in up to 6.1% of patients. Interpretations of all of the 95.3% true-negative results were correct for two specialists (A&F) followed by C, G and I. Overall recognition of true-negative results according to highly unlikely or unlikely interpretations and of true positive results according to clear, highly likely, likely, or possible interpretations were mainly correct. Nevertheless, there was considerable variation among specialists in numbers of patients assigned to those six interpretative categories (Appendix p12–13). After additional consideration of distributions and concordance of interpretations by specialists, it was determined that four specialists (A, C, F and I) stood out from the others in terms highest accuracy of interpretations (Appendix pp 13–14 Supplementary Fig. S1 and Table S3).

After provision of information about pre-test probability of disease at phase 2, some specialists showed improved identification of both patients with and without PPGL (D,H); however, for others there was either no improvement (A,B,E,F,G,J) or even worse performance (C, I) compared to phase 1 (Appendix p14 Supplementary Table S4).

Model performance

Internal testing identified five ML models with excellent predictive performance (Appendix p15 Supplementary Table S5). External testing established similar and equally excellent predictive performance for all five ML models according to areas under the ROC curves of 0.989 to 0.995 and balanced accuracies of 0.959 to 0.977 (Table 1).

Model concordance

Although the five selected models all offered excellent performance for binary classifications of disease, there was considerable scatter in relationships and discordance between probability scores of different models (Appendix pp15-16 Supplementary Fig. S2). Discordance in probability scores in part reflected variably skewed distributions of probability scores in patients with and without PPGL (Supplementary Fig. S3).

Concordance with human intelligence

The ANN and RF models presented with the highest agreement with both phase 1 and 2 interpretations of the four best performing specialists, followed by the

	SVM	RF	NB	ANN	LR
AUC	0.994 (0.988–0.997)	0.989 (0.980–0.994)	0.991 (0.987–0.995)	0.995 (0.988–0.997)	0.989 (0.981–0.993)
F1-score	0.959 (0.941–0.977)	0.959 (0.942–0.977)	0.940 (0.919–0.960)	0.966 (0.950–0.982)	0.929 (0.906–0.951)
Sensitivity	0.974 (0.965–0.0.982)	0.965 (0.955–0.974)	0.954 (0.943–0.965)	0.967 (0.958–0.976)	0.958 (0.948–0.968)
Specificity	0.978 (0.973–0.983)	0.982 (0.978–0.986)	0.969 (0.964–0.975)	0.986 (0.982–0.990)	0.959 (0.953–0.966)
Accuracy	0.977 (0.969–0.984)	0.977 (0.970–0.985)	0.965 (0.956–0.974)	0.981 (0.974–0.987)	0.959 (0.949–0.969)
Balanced Accuracy	0.976 (0.968–0.983)	0.974 (0.966–0.981)	0.962 (0.952–0.971)	0.977 (0.969–0.984)	0.959 (0.949–0.968)

Data are shown with confidence interval in parentheses. Sensitivities, specificities, F1 scores and accuracies were determined using optimal cut-offs according to Youden indexes: SVM, 0.073; RF, 0.180; NB, 0.888; ANN, 0.119; LR, 0.097. Abbreviations: SVM, support vector machine; RF, random forest; NB, naïve Bayes; ANN, artificial neural networks; LR, logistic regression; AUC, area under ROC curve.

Table 1: **Performance of the top performing ML models established after external testing for each of the five ML algorithms.**

SVM, LR and NB models (Appendix pp18-19 Supplementary Figs. S4–S6).

Model robustness

Overall, the NB model showed the lowest variability in probability scores for the QC samples, followed by the ANN model and then equally by SVM and RF models and finally the RF model (Appendix pp19-20 Supplementary Fig. S7). Nevertheless, despite low variability in probability scores, the NB model showed unusually high probabilities for QC samples with lower metabolite concentrations. The propensity of the NB model towards unusually high probability scores was confirmed from results of between laboratory measurements and derived probability scores (Supplementary Fig. S8). Moreover, for these assessments of between laboratory robustness of ML models, the NB model performed poorly in terms of variability of probability scores compared to the ANN model, which displayed the lowest between laboratory variability in probability scores among all models (Appendix p21 Supplementary Fig. S9).

Final model selection

After completion of performance, concordance and robustness test steps, we selected the best performing model according to overall performance in all steps. As shown in Table 2, the model with the highest performance score was the ANN model.

Specialists versus ANN model predictions of PPGL

Distributions of ANN model probability scores according to phase 1 predictions by specialists of the likelihood of PPGL were highly skewed to lower probability scores for predictions that PPGL was highly unlikely or unlikely compared to higher probability scores for predictions that PPGL was likely, highly likely or clear (Fig. 2A). Probability scores were skewed to both lower and higher values for predictions that PPGL was possible.

Superior performance of the ANN model for interpretations of the likelihood of PPGL compared to interpretations of specialists became apparent after separation of patients into those with and without PPGL (Fig. 2B). Specifically, median probability scores were respectively 70-, 175-, 59-, 15-, 3.5- and 1.7-fold higher ($P < 0.0001$) for patients with than without PPGL for the six predictions by specialists that PPGL was highly unlikely, unlikely, possible, likely, highly likely and clear. Therefore, ML-derived probability scores provide superior indicators of the likelihood of PPGL compared to each of the six predictions of specialists.

Using heat map analyses, the highly variable phase 1 predictions of disease likelihood by specialists were most apparent for patients with minor to moderate increases in plasma metanephrines above upper cut-offs (Fig. 3). Selections based on high standard deviations of digitised predictions of disease likelihood (Fig. 3B and C) facilitated comparisons to ANN probability scores (Supplementary Fig. S10). For 115 patients without PPGL selected for highly variable interpretations by specialists, the ANN model reliably excluded 67% (77/115) of all cases, according to probability scores below 0.03. Among 183 patients with PPGL and similarly variable interpretations by specialists, 97% (178/183) had probability scores above 0.03. The ANN model reliably confirmed 44% (81/183) of these cases according to probability scores above 0.99.

Impact of ANN model probabilities on decision making

Influences of probability scores to improve phase 3 assignments of disease likelihood varied among specialists and were limited in scope (Appendix pp23-28). Provision of ANN probability scores failed to improve predictions by specialists of disease highly unlikely or unlikely (Supplementary Fig. S11). There was, however, a marked change in distributions of probability scores at phase 3 for patients categorised as having highly likely or clear disease; this involved reductions in the number of patients with low probability scores that indicated some improvement for identification of patients at high risk of disease (Supplementary Fig. S12). For the disease likely category, reassignment of disease predictions after provision of ANN probability scores resulted in a 4-fold increase in median probability scores for patients without PPGL and a downward distribution of scores for patients with PPGL (Supplementary Fig. S13). Finally, for patients categorised in the disease possible category there was no relevant difference in probability scores from phase 1 to 3 (Supplementary Fig. S14).

Test steps	SVM	RF	NB	ANN	LR
Step 1: Conventional performance metrics–reproducibility					
AUC	4	1.5	3	5	1.5
F1 Score	3.5	3.5	2	5	1
Balanced Accuracy	4	3	2	5	2
Rank mean[a]	4	3	2	5	1
Step 2: Concordance with human intelligence					
Relationships	2.5	4	2.5	5	1
Digitised agreement	3	4.5	1	4.5	2
Step 3: Robustness					
Within laboratory variation	2.5	1	5	4	2.5
Between laboratory variation	4	2	1	5	3
Summed ranks[a]	**16**	**14.5**	**11.5**	**23.5**	**9.5**

SVM, support vector machine; RF, random forest; NB, naïve Bayes; ANN, artificial neural networks; LR, logistic regression. [a]For conventional performance metrics derived from the data in Table 1 a single rank mean was employed for final evaluation according to the summed fractional ranks from that test step. The higher fractional ranks and final scores indicate stronger performance.

Table 2: **Evaluation of the five ML models according to conventional metrics, concordance with human intelligence and robustness.**

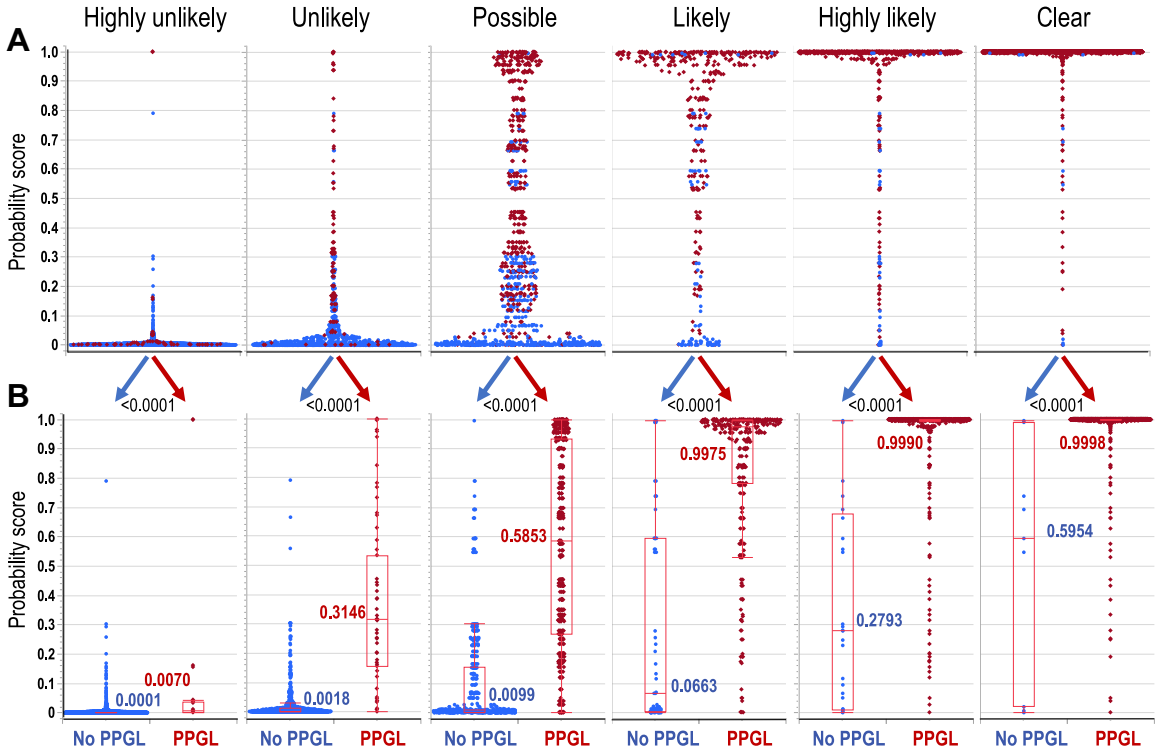

Fig. 2: Dot plot displays of distributions of ANN probability scores according to specialist interpretations of the likelihood of PPGL before (A) and after (B) separation of patients with (◆) and without (●) PPGL. Data reflect phase 1 interpretations of all 10 specialists for combined internal and external validation datasets. Dot plot displays in panel B include box plots with additional inclusion of median values. Data in panel A illustrate the distributions of probability scores for interpretations by specialists of disease likelihood according to the six categories of highly unlikely, unlikely, possible, likely, highly likely and clear. The data in panel B were derived after splitting patients in panel A into two groups with and without PPGL. Findings that ML probability scores were consistently higher for patients with than without PPGL for each of the six predictive categories illustrate superiority of probability scores over each of the six interpretations of specialists for assessment of disease likelihood.

Discussion

This study employed several novel processes for ML model validation to establish a robust and reliable model to improve decision-making in the diagnosis of PPGL. With probability scores that are superior to interpretations of specialists, the selected ANN model has potential to offer guidance to more efficiently direct patients to therapeutic intervention versus exclusion of disease. Nevertheless, our study also shows that simply providing probability scores minimally improves interpretations of test results.

Despite the high accuracy of plasma metanephrines for diagnosis of PPGL, clinicians continue to have difficulty in interpretation of test results. Positive test results are often ignored,[22,23] in part due to high proportions of false-positive versus true-positive results. On the other hand, patients may also undergo unnecessary adrenalectomy due to positive results at screening,[24,25] though without adequate confirmatory tests when these should have been indicated. Difficulty in decision-making is illustrated by the considerable variability in interpretation by clinicians in the present study, which for some results ranged from PPGL unlikely to highly likely.

The above considerations serve to highlight that interpretation of biochemical test results is often challenging, particularly when there is need to consider multiple results and pre-versus post-test probabilities. Provision of ML model-derived probability scores could meet this challenge by enhancing the interpretative framework to guide decision making. For PPGL, ML probability scores may complement information provided by biochemical test results to better guide clinicians about the likelihood of a tumour and the appropriate next steps in patient management.

There are numerous proof-of-concept studies that highlight the potential of ML in health care. The application of ML in clinical settings, however, remains limited. This can be partially attributed to the methodological challenges that these technologies face for translation to the clinical routine.[15] In particular, ML model performance should not be the only decisive

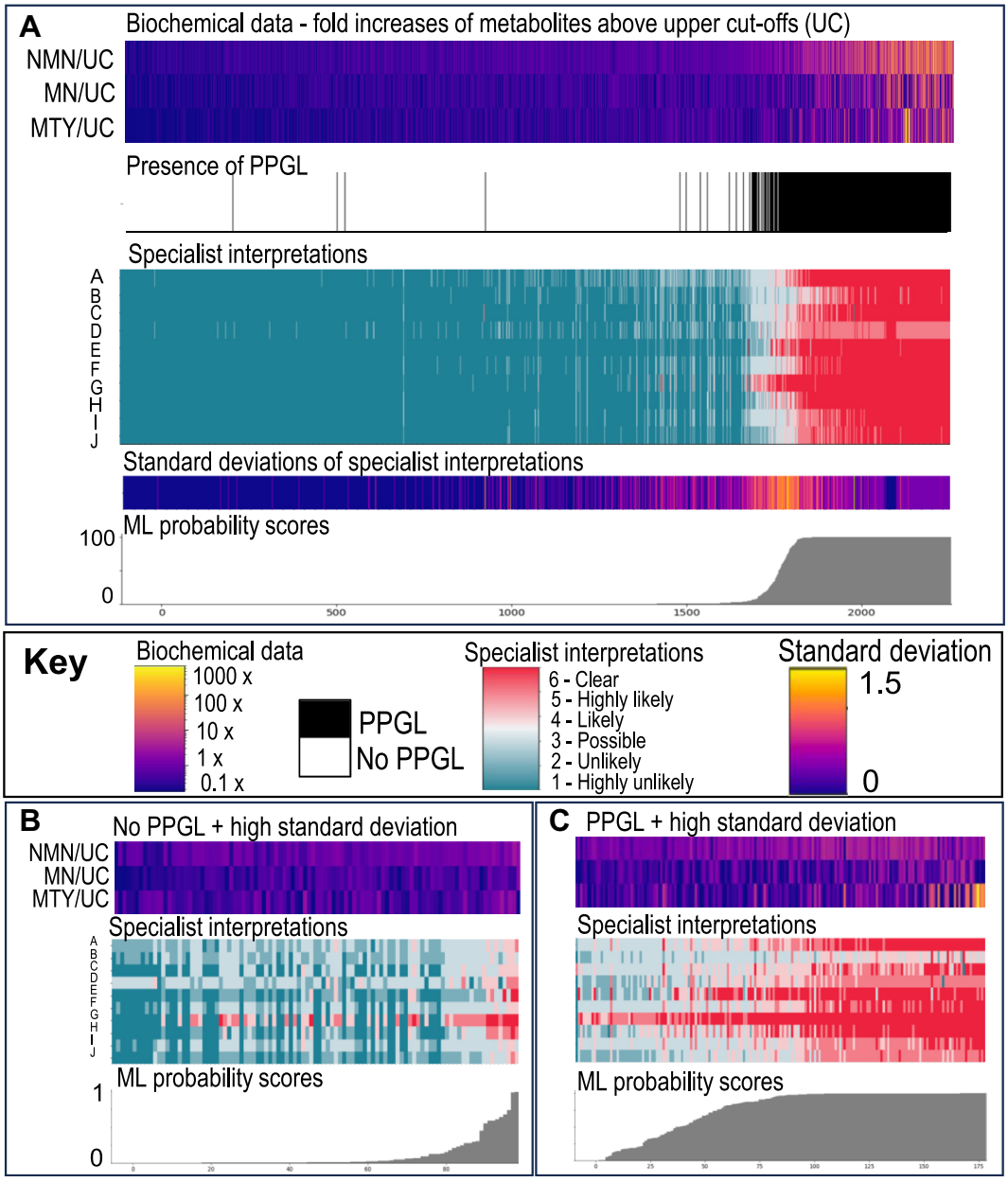

Fig. 3: Heat map displays of biochemical test results and phase 1 interpretations of specialists in relation to ANN model probability scores, presence versus absence of PPGL and according to variability (standard deviations) of interpretations. Panel A displays results for all 2246 patients in internal and external validation series and is arranged in sequence according to probability scores. Panels B and C show results for selections of patients with (C) and without (B) PPGL in who predictions of disease by specialists showed the highest variability according to standard deviations of digitised predictions.

factor to consider before ML is implemented in clinical settings.[26] Our study, which adheres to current technical guidelines and recommendations,[27–29] proposes additional methodological steps to responsibly implement ML models in clinical settings.

External validation in a large population that differs in setting, baseline characteristics or outcomes is a desirable procedure to ensure generalisability of ML models.[30] Although this may be sufficient for binary classification, the present analysis establishes that external validation alone is insufficient to ensure that ML-derived probability scores provide meaningful information for clinicians. In particular, although after external validation all five models offered similarly

excellent performance for disease classification, they also displayed highly variable and discordant probability scores. This is not surprising, as different classification algorithms use different approaches to learn probability distributions according to their capability to produce probabilistic (e.g., ANN, LG, NB) versus non-probabilistic or deterministic predictions (e.g., SVM, RF).[31] Nevertheless, in clinical decision-making, it can be important to have a predictor that correctly quantifies predictive uncertainty instead of simply producing a point estimate.

With consideration that model performance alone cannot ensure utility in clinical settings, we introduced additional validation steps. Concordance of ML models with human intelligence provided a procedure to optimise clinical applicability. Evaluation of variability in probability scores according to within and between laboratory repeated measurements of metabolites in identical plasma samples was employed to determine model robustness, which is also important for generalisability. Using the above metrics, the ANN model was determined to offer highest overall performance.

Importantly, when we examined disease predictions by specialists in relation to ANN-derived probability scores, probability scores were lower for predictions that disease was "high unlikely" or "unlikely" in patients without than with PPGL and higher for predictions of "likely", "highly likely" or "clear" disease in patients with than without PPGL. These findings indicate that the ANN model can better determine relative likelihood of a PPGL than specialists and suggests potential to guide decision-making in routine clinical care.

The fact that the ANN model was trained only on four clinical and biochemical features routinely available in clinical settings, suggested that the model could be directly interpreted by specialists without information loss or error in the explanation process. Thus, the final evaluation of the selected model was to assess the utility of the ANN probability scores in decision-making. Although provision of ANN probability scores to specialists improved identification of patients with a high risk of disease, there was surprisingly no to little improvement for other interpretations.

Even with simplistic binary approaches, interpretation of test results according to pre- and post-test probabilities of disease is difficult for physicians.[6] Add to this the need to consider multiple test results according to their continuous nature, the provision of ML probability scores might serve to magnify interpretative difficulties. Another explanation for failure of the provided probability scores to improve test interpretation could be that physicians do not trust ML technology, particularly when it is unclear how outputs relate to the relevant clinical data.[32,33] This problem may be addressed by interfaces that integrate data and ML models within clinical decision support systems to provide automated patient-specific interpretations to assist clinicians towards a decision.[34,35] Such systems with associated data visualisation tools and narrative reports employed by medical laboratories might then facilitate the smooth and trustworthy integration of ML technologies into the clinical setting. Nevertheless, whether application of ML models in clinical practice can actually improve outcomes for patients requires randomised clinical trials to assess net benefit.

Our study has several strengths, though also limitations; in particular, the study was retrospective in nature. Interpretations by specialists were in part artificial, which is a limitation since in real-life settings physicians incorporate more clinical data and may consult with other physicians for decision-making. Another limitation is that possible intra-reviewer scoring variations between phases were not considered. Among strengths, the separate large sized patient cohorts facilitated generalisability. Use of follow-up minimised misclassification of patients. The developed ML models also utilised widely available biochemical tests that can be applied by clinical laboratories where quality assurance procedures for laboratory tests can also be applied to ML models to ensure continued robustness and reliability of probability score outputs.

Within the progressive and challenging landscape of ML in health care, we established a robust ML model to assist decision-making in diagnosis of PPGL and suggest structures relevant for laboratory medicine to simplify the transition from development to implementation of effective and clinically relevant ML tools. The suggested structures may strengthen the ability of physicians to integrate ML tools within daily routine clinical settings, though this will likely require integration of such tools within clinical decision support systems.

Contributors

C.P., G.P., and G.E. conceived and designed the study, and drafted the manuscript; C.P., M.P., H.R., G.C., A.M.B., M.A.N., F.B., M.F., A.P., K.P., and G.E. contributed to enrolment of patients or analyses of samples and collection and interpretation of clinical data; C.P., G.P., I.A., A.F., and G.E constructed and tested the ML models and have verified the underlying data; all authors contributed to the analyses of data, read and approved the final version of the manuscript.

Data sharing statement

All data generated in this study has been provided for purposes of review and will be made publicly available upon publication of the manuscript.

Declaration of interests

C.P., G.P, A.F. and G.E declare a filed German patent A5914/TUD 017. MN and KP were government employees at the NIH by the time this work was done. MF is an ExCo of the European Society of Endocrinology.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2025.103181.

References

1 Lenders JWM, Duh QY, Eisenhofer G, et al. Pheochromocytoma and paraganglioma: an endocrine society clinical practice guideline. *J Clin Endocrinol Metab*. 2014;99:1915–1942.

2 Eisenhofer G, Prejbisz A, Peitzsch M, et al. Biochemical diagnosis of chromaffin cell tumors in patients at high and low risk of disease: plasma versus urinary free or deconjugated o-methylated catecholamine metabolites. *Clin Chem*. 2018;64:1646–1656.
3 Eisenhofer G, Pamporaki C, Lenders JWM. Biochemical assessment of pheochromocytoma and paraganglioma. *Endocr Rev*. 2023;44:862–909.
4 Manrai AK, Bhatia G, Strymish J, Kohane IS, Jain SH. Medicine's uncomfortable relationship with math: calculating positive predictive value. *JAMA Intern Med*. 2014;174:991–993.
5 Whiting PF, Davenport C, Jameson C, et al. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open*. 2015;5:e008155.
6 Morgan DJ, Pineles L, Owczarzak J, et al. Accuracy of practitioner estimates of probability of diagnosis before and after testing. *JAMA Intern Med*. 2021;181:747–755.
7 Darr R, Pamporaki C, Peitzsch M, et al. Biochemical diagnosis of phaeochromocytoma using plasma-free normetanephrine, metanephrine and methoxytyramine: importance of supine sampling under fasting conditions. *Clin Endocrinol (Oxf)*. 2014;80:478–486.
8 Pamporaki C, Bursztyn M, Reimann M, et al. Seasonal variation in plasma free normetanephrine concentrations: implications for biochemical diagnosis of pheochromocytoma. *Eur J Endocrinol*. 2014;170:349–357.
9 Pommer G, Pamporaki C, Peitzsch M, et al. Preanalytical considerations and outpatient versus inpatient tests of plasma metanephrines to diagnose pheochromocytoma. *J Clin Endocrinol Metab*. 2022;107:e3689–e3698.
10 Constantinescu G, Preda C, Constantinescu V, et al. Silent pheochromocytoma and paraganglioma: systematic review and proposed definitions for standardized terminology. *Front Endocrinol (Lausanne)*. 2022;13:1021420.
11 Rothman B, Leonard JC, Vigoda MM. Future of electronic health records: implications for decision support. *Mt Sinai J Med*. 2012;79:757–768.
12 He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. 2019;25:30–36.
13 Reel PS, Reel S, van Kralingen JC, et al. Machine learning for classification of hypertension subtypes using multi-omics: a multicentre, retrospective, data-driven study. *EBioMedicine*. 2022;84:104276.
14 Pamporaki C, Berends AMA, Filippatos A, et al. Prediction of metastatic pheochromocytoma and paraganglioma: a machine learning modelling study using data from a cross-sectional cohort. *Lancet Digit Health*. 2023;5:e551–e559.
15 Spies NC, Farnsworth CW, Wheeler S, McCudden CR. Validating, implementing, and monitoring machine learning solutions in the clinical laboratory safely and effectively. *Clin Chem*. 2024;70:1334–1343.
16 Zhang J, Whebell S, Gallifant J, et al. An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. *Lancet Digit Health*. 2022;4:e212–e213.
17 Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*. 2020;2:e489–e492.
18 Kwong JCC, Khondker A, Lajkosz K, et al. Appraise-ai tool for quantitative evaluation of ai studies for clinical decision support. *JAMA Netw Open*. 2023;6:e2335377.
19 Lenders JW, Eisenhofer G, Armando I, Keiser HR, Goldstein DS, Kopin IJ. Determination of metanephrines in plasma by liquid chromatography with electrochemical detection. *Clin Chem*. 1993;39:97–103.
20 Peitzsch M, Prejbisz A, Kroiss M, et al. Analysis of plasma 3-methoxytyramine, normetanephrine and metanephrine by ultra-performance liquid chromatography-tandem mass spectrometry: utility for diagnosis of dopamine-producing metastatic phaeochromocytoma. *Ann Clin Biochem*. 2013;50:147–155.
21 Peitzsch M, Novos T, Kaden D, et al. Harmonization of lc-ms/ms measurements of plasma free normetanephrine, metanephrine, and 3-methoxytyramine. *Clin Chem*. 2021;67:1098–1112.
22 Garrahy A, Casey R, Wall D, Bell M, O'Shea PM. A review of the management of positive biochemical screening for phaeochromocytoma and paraganglioma: a salutary tale. *Int J Clin Pract*. 2015;69:802–809.
23 Samsudin IN, Page MM, Hoad K, et al. Annals express: the challenge of improving the diagnostic yield from metanephrine testing in suspected phaeochromocytoma and paraganglioma. *Ann Clin Biochem*. 2018:4563218774590.
24 Dobri GA, Bravo E, Hamrahian AH. Pheochromocytoma: pitfalls in the biochemical evaluation. *Expert Rev Endocrinol Metab*. 2014;9:123–135.
25 Carr JC, Spanheimer PM, Rajput M, et al. Discriminating pheochromocytomas from other adrenal lesions: the dilemma of elevated catecholamines. *Ann Surg Oncol*. 2013;20:3855–3861.
26 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25:1337–1340.
27 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *BMC Med*. 2015;13:1.
28 Wolff RF, Moons KGM, Riley RD, et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51–58.
29 Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. *BMJ*. 2024;386:e078276.
30 Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14:49–58.
31 Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. 2015;521:452–459.
32 Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318:517–518.
33 Ali S, Abuhmed T, El-Sappagh S, et al. Explainable artificial intelligence (xai): what we know and what is left to attain trustworthy artificial intelligence. *Inf Fusion*. 2023;99.
34 Sim I, Gorman P, Greenes RA, et al. Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assoc*. 2001;8:527–534.
35 Cubukcu HC, Topcu DI, Yenice S. Machine learning-based clinical decision support using laboratory data. *Clin Chem Lab Med*. 2024;62:793–823.