Prediction of metastatic pheochromocytoma and paraganglioma: a machine learning modelling study using data from a cross-sectional cohort

Pamporaki, Christina ; Berends, Annika M A ; Filippatos, Angelos ; Prodanov, Tamara ; Meuter, Leah ; Prejbisz, Alexander ; Beuschlein, Felix ; Fassnacht, Martin ; Timmers, Henri J L M ; Nölting, Svenja ; Abhyankar, Kaushik ; Constantinescu, Georgiana ; Kunath, Carola ; de Haas, Robbert J ; Wang, Katharina ; Remde, Hanna ; Bornstein, Stefan R ; Januszewicz, Andrzeij ; Robledo, Mercedes ; Lenders, Jacques W M ; Kerstens, Michiel N ; Pacak, Karel ; Eisenhofer, Graeme

DOI: https://doi.org/10.1016/S2589-7500(23)00094-8

Originally published at:
Pamporaki, Christina; Berends, Annika M A; Filippatos, Angelos; Prodanov, Tamara; Meuter, Leah; Prejbisz, Alexander; Beuschlein, Felix; Fassnacht, Martin; Timmers, Henri J L M; Nölting, Svenja; Abhyankar, Kaushik; Constantinescu, Georgiana; Kunath, Carola; de Haas, Robbert J; Wang, Katharina; Remde, Hanna; Bornstein, Stefan R; Januszewicz, Andrzeij; Robledo, Mercedes; Lenders, Jacques W M; Kerstens, Michiel N; Pacak, Karel; Eisenhofer, Graeme (2023). Prediction of metastatic pheochromocytoma and paraganglioma: a machine learning modelling study using data from a cross-sectional cohort. The Lancet Digital Health, 5(9):e551-e559.
DOI: https://doi.org/10.1016/S2589-7500(23)00094-8

Prediction of metastatic pheochromocytoma and paraganglioma: a machine learning modelling study using data from a cross-sectional cohort

Christina Pamporaki, Annika M A Berends, Angelos Filippatos*, Tamara Prodanov, Leah Meuter, Alexander Prejbisz, Felix Beuschlein, Martin Fassnacht, Henri J L M Timmers, Svenja Nölting, Kaushik Abhyankar, Georgiana Constantinescu, Carola Kunath, Robbert J de Haas, Katharina Wang, Hanna Remde, Stefan R Bornstein, Andrzeij Januszewicz, Mercedes Robledo, Jacques W M Lenders, Michiel N Kerstens, Karel Pacak, Graeme Eisenhofer*

Summary

Background Pheochromocytomas and paragangliomas have up to a 20% rate of metastatic disease that cannot be reliably predicted. This study prospectively assessed whether the dopamine metabolite, methoxytyramine, might predict metastatic disease, whether predictions might be improved using machine learning models that incorporate other features, and how machine learning-based predictions compare with predictions made by specialists in the field.

Methods In this machine learning modelling study, we used cross-sectional cohort data from the PMT trial, based in Germany, Poland, and the Netherlands, to prospectively examine the utility of methoxytyramine to predict metastatic disease in 267 patients with pheochromocytoma or paraganglioma and positive biochemical test results at initial screening. Another retrospective dataset of 493 patients with these tumors enrolled under clinical protocols at National Institutes of Health (00-CH-0093) and the Netherlands (PRESCRIPT trial) was used to train and validate machine learning models according to selections of additional features. The best performing machine learning models were then externally validated using data for all patients in the PMT trial. For comparison, 12 specialists provided predictions of metastatic disease using data from the training and external validation datasets.

Findings Prospective predictions indicated that plasma methoxytyramine could identify metastatic disease at sensitivities of 52% and specificities of 85%. The best performing machine learning model was based on an ensemble tree classifier algorithm that used nine features: plasma methoxytyramine, metanephrine, normetanephrine, age, sex, previous history of pheochromocytoma or paraganglioma, location and size of primary tumours, and presence of multifocal disease. This model had an area under the receiver operating characteristic curve of 0·942 (95% CI 0·894–0·969) that was larger (p<0·0001) than that of the best performing specialist before (0·815, 0·778–0·853) and after (0·812, 0·781–0·854) provision of *SDHB* variant data. Sensitivity for prediction of metastatic disease in the external validation cohort reached 83% at a specificity of 92%.

Interpretation Although methoxytyramine has some utility for prediction of metastatic pheochromocytomas and paragangliomas, sensitivity is limited. Predictive value is considerably enhanced with machine learning models that incorporate our nine recommended features. Our final model provides a preoperative approach to predict metastases in patients with pheochromocytomas and paragangliomas, and thereby guide individualised patient management and follow-up.

Funding Deutsche Forschungsgemeinschaft.

Introduction

Pheochromocytomas and paragangliomas are neuroendocrine tumours with up to a 35% hereditary predisposition[1] and an approximately 20% prevalence of metastatic disease.[2,3] Unlike other tumours, there are no histopathological methods to identify metastatic disease and all pheochromocytomas and paragangliomas must be considered to have variable potential to metastasise.[4] Currently only presence of metastases at sites where no chromaffin tissue should be expected (eg, bones and lymph nodes) establishes a definitive diagnosis of metastatic disease.[4,5] Therefore, long-term follow-up is recommended for all patients with pheochromocytoma or paraganglioma.[6]

Earlier therapeutic intervention in patients with metastatic pheochromocytoma or paraganglioma is expected to reduce morbidity and mortality.[7] Identification of features to reliably predict metastatic potential of pheochromocytoma or paraganglioma at initial tumour diagnosis is therefore crucial. The relation of tumoural dopamine production to metastatic disease in patients with these tumours is established.[2,8,9] Use of

Lancet Digit Health 2023; 5: e551–59

Published Online July 18, 2023 https://doi.org/10.1016/S2589-7500(23)00094-8

*Contributed equally

Department of Medicine III (C Pamporaki PD, G Constantinescu MD, C Kunath, Prof S R Bornstein PhD, Prof J W M Lenders PhD, Prof G Eisenhofer PhD) **and Institute of Clinical Chemistry and Laboratory Medicine** (Prof G Eisenhofer), **University Hospital Carl Gustav Carus, Institute of Lightweight Engineering and Polymer Technology** (Prof A Filippatos PhD, K Abhyankar Dipl-Ing), **TU Dresden, Dresden, Germany; Department of Endocrinology, Medical Centre Groningen** (A M A Berends PhD, M N Kerstens PhD) **and Department of Radiology, Medical Imaging Centre Groningen** (R J de Haas PhD), **University of Groningen, Groningen, Netherlands; Machine Design Laboratory, Department of Mechanical Engineering & Aeronautics, University of Patras, Patras, Greece** (Prof A Filippatos); **Section on Medical Neuroendocrinology, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA** (T Prodanov PhD, L Meuter MD, Prof K Pacak DrSc); **Department of Hypertension, Institute of Cardiology, Warsaw, Poland** (Prof A Prejbisz PhD, Prof A Januszewicz PhD); **Department of Internal Medicine IV, University Hospital LMU, Ludwig Maximilian University of Munich, Munich, Germany** (Prof F Beuschlein PhD,

Prof S Nölting PhD, K Wang Staatsexa); **Department of Endocrinology, Diabetology, and Clinical Nutrition, University Hospital, University of Zurich, Zurich, Switzerland** (Prof F Beuschlein, Prof S Nölting); **Department of Internal Medicine, Division of Endocrinology and Diabetes, University of Würzburg, Würzburg, Germany** (Prof M Fassnacht PhD, H Remde PhD); **Department of Internal Medicine, Radboud University Medical Centre, Nijmegen, Netherlands** (Prof H J L M Timmers PhD, Prof J W M Lenders); **Hereditary Endocrine Cancer Group, Human Cancer Genetics Programme, Spanish National Cancer Reserch Centre, Madrid, Spain** (Prof M Robledo PhD); **Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Madrid, Spain** (Prof M Robledo)

Correspondence to:
Dr Christina Pamporaki, Department of Medicine III, University Hospital Carl Gustav Carus, TU Dresden, Dresden D-01307, Germany
christina.pamporaki@uniklinikum-dresden.de

Research in context

Evidence before this study

We searched PubMed on June 10, 2022, using the search (metastatic[Title] OR malignant[Title]) AND (pheochromocytoma[Title] OR paraganglioma[Title]) AND (predictor OR predict OR diagnose OR diagnosis). We also searched abstracts of the American Society of Clinical Oncology Annual Meeting, European Society for Medical Oncology Congress and American Association for Cancer Research Annual Meeting, European Society of Endocrinology Annual Meeting, Ensat International Adrenal Cancer Symposium, within the past 3 years using the same search terms. We identified several studies on predictors of metastatic disease among patients with pheochromocytomas and paragangliomas. In particular, retrospective observational studies have established that young age, large tumour size, extra-adrenal tumour location, presence of specific pathogenic germline (eg, *SDHB*) or somatic variants (eg, *ATRX, TERT, MAML3*), and specific, long, non-coding RNAs are associated with higher risk of metastatic disease among patients with pheochromocytomas and paragangliomas. Nevertheless, none of the aforementioned features were robust enough alone to predict metastatic disease. We also identified five studies that focused on combining features in scoring systems. Histopathological features were included in most of these scores; however, these have low reproducibility and accuracy. Similarly, a scoring system derived purely from clinical data had an inappropriately low positive predictive value. Machine learning is a new digital approach that could potentially support decision making in health care. We identified studies that established machine learning models to differentiate patients with pheochromocytomas and paragangliomas from patients with other forms of hypertension, using mainly metabolomics, or in the field of radiomics for the differentiation of incidental adrenal masses. However, no studies were identified that introduced machine learning models to predict metastatic pheochromocytomas and paragangliomas.

Added value of this study

This clinician-designed and implemented study introduces robust non-invasive machine learning models to predict metastatic disease in patients with pheochromocytomas and paragangliomas. These models use only routinely available features preoperatively, and can be readily applied and adapted by clinicians not only for pheochromocytomas and paragangliomas, but also for other cancers. High performance and reproducibility of the selected machine learning models was secured by both external validation using a different patient cohort and also through comparisons with interpretations by an international group of clinical care specialists with expertise in the management of patients with pheochromocytomas and paragangliomas. This comparison established that the selected ensemble tree classifier machine learning model provided significantly superior performance over interpretations of all specialists and could reliably predict metastatic disease in most patients with pheochromocytomas and paragangliomas.

Implications of all the available evidence

We expect that clinicians will benefit from the assistance of the selected machine learning models, as they provide suitable prediction of metastatic pheochromocytomas and paragangliomas, and can be easily implemented in digital health-care systems. Overall, our findings support emerging concepts that machine learning will gain traction in oncology for its potential to facilitate robust diagnostic stratification and guide personalised patient management.

methoxytyramine, the O-methylated metabolite of dopamine, as a predictor of metastases offers promise, but has only been evaluated in a single retrospective patient series.[2] Young age,[10,11] large tumour size,[2,11,12] and extra-adrenal location of primary tumours[2,12] represent other established clinical predictors of metastases. Tumours due to pathogenic variants of *SDHB* and somatic genomic alterations such as *ATRX*, *TERT*, or *MAML3* are also associated with higher metastatic potential.[13,14] However, such information is rarely available preoperatively when establishing metastatic risk would be useful.

Despite the association of the features with the development of metastases, there is no robust method to reliably predict metastatic disease in patients with pheochromocytoma or paraganglioma. Some effort has been made to combine different features in scoring systems to predict metastatic pheochromocytoma or paraganglioma, but most involve histopathological parameters,[15,16] which are difficult to standardise in clinical practice[17] and are low in accuracy.[18] An attempt to establish a predictive score using routinely available clinical features similarly did not meet expectations according to a low positive predictive value.[19]

Advances in computational power have led to the introduction of multidimensional digitalised approaches that could potentially support decision making in health care. Machine learning is one such approach for interrogating multidimensional data and an area of artificial intelligence that uses computational algorithms for different tasks; this method is without the need for the explicit programming of previously established mathematical relationships.[20,21] In diagnostics, these tasks principally involve classification.[22,23]

Taking the above into consideration, the present study had three aims: (1) prospectively validate the use of methoxytyramine as a preoperative predictor of metastases in patients with pheochromocytoma or paraganglioma; (2) establish machine learning models that incorporate methoxytyramine with other features to predict metastatic pheochromocytoma or paraganglioma preoperatively; and (3) compare the performance of the selected machine

learning models with the predictions of 12 clinical care specialists with expertise in the management of patients with pheochromocytoma or paraganglioma.

Methods

Study design

This machine learning modelling study, included patients with and without metastatic pheochromocytoma or paraganglioma enrolled at seven tertiary centers (in Germany, the Netherlands, Poland, and the USA). Data were from three sources: the PMT trial, National Institutes of Health (NIH)'s 00-CH-0093, and the PRESCRIPT trial. Patients were enrolled in the PMT trial in centres in Germany, the Netherlands, and Poland between July 5, 2010, and Aug 31, 2019. Patients were enrolled in NIH's 00-CH-0093 in the USA between Aug 25, 2010, and April 18, 2016, and in the PRESCRIPT study in the Netherlands, between Jan 1, 2012, and Dec 31, 2017. The PMT trial had follow-up until June 9, 2021. Clinical protocols were approved by local ethics committees.

Patients

Available clinical information for both the PMT trial and retrospective data included sex, age at initial tumour diagnosis, presence of multifocal and metastatic disease, initial tumour location and size, and plasma concentrations of free normetanephrine, metanephrine, and methoxytyramine.

Recruitment of patients into the PMT trial was based on clinical suspicion or risk of pheochromocytoma or paraganglioma according to four main criteria. First, signs and symptoms of catecholamine excess. Second, hereditary risk of pheochromocytoma or paraganglioma according to syndromic presentation, family history, or an established mutation of a tumour-susceptibility gene. Third, findings of an incidentally discovered mass during imaging studies carried out for reasons unrelated to suspicion of pheochromocytoma or paraganglioma. Finally, a previous history of a resected pheochromocytoma or paraganglioma. Individuals taking norepinephrine reuptake blockers and levodopa, or other medications known to raise plasma concentrations of O-methylated metabolites, were excluded. Registration of patients into electronic case report forms according to these criteria was initiated for all centres.

Procedures

Metastatic disease was defined as the presence of metastases in tissues distant from the primary tumour, where chromaffin cells are normally absent.[4] Metastases were identified by conventional and functional imaging or histopathological examination of resected lymph nodes (appendix p 2). Testing for germline pathogenic variants of *VHL, RET, SDHX, MAX,* and *TMEM127* was done for 708 patients using Sanger sequencing or next generation sequencing, and multiplex ligation-dependent probe amplification or custom array comparative genomic hybridisation for deletion detection. Genetic testing was not available for 80 patients (ie, 60 without metastases and 17 with metastases in the retrospective data, and two without metastases and one with metastases in the PMT trial).

The first outcome of the study was to assess the prospective use of plasma methoxytyramine to predict metastatic disease. This outcome involved 267 patients with pheochromocytoma or paraganglioma under the PMT trial who had positive biochemical test results at initial screening. One objective of the PMT trial was to establish utility of plasma methoxytyramine to predict metastases (appendix pp 2–3). For this purpose, the investigator responsible for biochemical tests provided predictions of metastatic disease that were based primarily on measurements of plasma methoxytyramine, with additional consideration of the other two metabolites. These predictions and others, along with biochemical test results, were provided back to the responsible physicians at each participating centre. Predictions were restricted to patients with positive biochemical tests and were in the form of standardised comments that indicated strong, moderate, possible, or low risk of metastases (appendix p 3).

For the **PMT trial** see https://pmt-study.pressor.org

The second outcome of the study was to generate machine learning models to predict metastatic disease. For this outcome, we retrieved data from 493 patients in the USA and the Netherlands with pheochromocytoma or paraganglioma (ie, the training cohort) to generate and internally test various machine learning models using four different machine learning algorithms. The best candidate machine learning models were then externally validated using the dataset of 295 patients with pheochromocytoma or paraganglioma who were enrolled in the PMT trial (ie, the external validation cohort). After external validation, we compared machine learning models using multiple metrics and selected the final top performing models (figure 1).

Machine learning models were developed after data preparation and normalisation using four supervised machine learning algorithms with all variables included and according to ten cross validations in five folds. The supervised machine learning algorithms included decision tree classifier, Support Vector Machine, Naive Bayes, and AdaBoost ensemble tree classifier. χ^2 feature analysis for classification was done in the training cohort to identify invalid features containing irrelevant or redundant information.

During data preparation, we did feature analysis in the training cohort twice. The first feature analysis included nine features: (1) plasma free methoxytyramine, (2) age at initial tumour diagnosis, (3) sex, (4) previous history of pheochromocytoma or paraganglioma (input as yes or no), (5) primary tumour location, (6) primary tumour size, (7) presence of multifocal disease (input as yes or no), (8) plasma free metanephrine, and (9) plasma free normetanephrine. The second feature analysis included

See **Online** for appendix

the same features, supplemented by presence of *SDHB* pathogenic variants (input as positive or negative), the genetic component with the strongest anticipated metastatic predictive potential.

After feature analysis, multiple rounds of training and internal testing of machine learning models were done to identify the best candidate machine learning models according to areas under the curves (AUC) of the receiver operating characteristic (ROC), and with consideration of Matthew's correlation coefficient (MCC) and balanced accuracy. To confirm reproducibility of our results, we then externally validated the best candidate machine learning models in a separate cohort of patients from the PMT trial (ie, the external validation cohort). The best machine learning models of the external validation were again selected by comparing their predictive performance according to AUC, with consideration of MCC and balanced accuracy. Machine learning was done using MATLAB MathWorks R2020a (appendix pp 3–4).

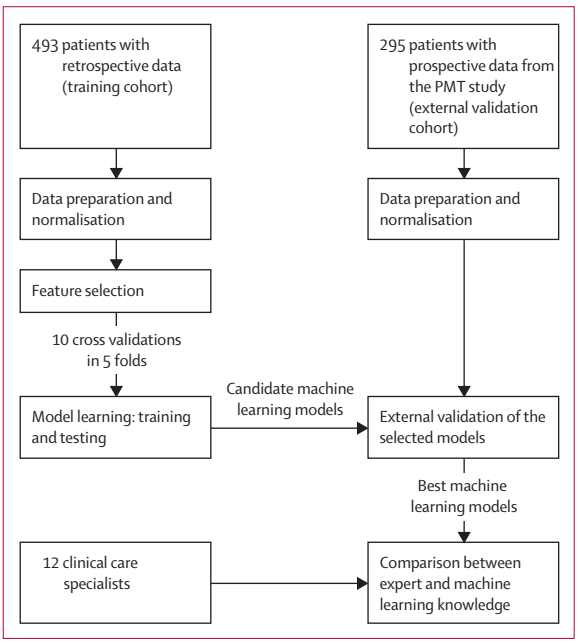

Figure 1: **Workflow for the data analysis**

The third outcome of the study was to assess predictions of metastatic disease by clinical care specialists. We invited 12 clinical care specialists from the USA, Germany, and the Netherlands with expertise in pheochromocytomas and paragangliomas to provide predictions of metastatic disease for the training and external validation cohort. Seven specialists reported experience of more than 10 years, and five reported less than 10 years. Specialists were requested to provide their own estimates of metastatic disease using a classification score of four categories: low, possible, moderate, and strong probability. Before the review process, specialists received detailed definitions for each of the four classification categories, including specific probability intervals for metastatic disease and narrative interpretations for further patient management (appendix p 5). These probabilities were according to the same nine features described for machine learning feature analysis.

Similar to the feature analysis and machine learning training, specialists were instructed to provide probabilities of metastatic disease twice, which included

	Training cohort			External validation cohort		
	Without metastases	With metastases	p value	Without metastases	With metastases	p value
Sex			0·045			0·0100
Male	156/327 (48%)	95/166 (57%)	..	93/238 (39%)	33/57 (58%)	..
Female	171/327 (52%)	71/166 (43%)	..	145/238 (61%)	24/57 (42%)	..
Age at initial diagnosis (years)	39·6 (37·9–41·3)	31·8 (31·9–35·4)	<0·0001	44·7 (43·2–46·2)	40·6 (39·2–42·1)	0·025
Initial tumour size (cm)	2·7 (2·4–2·9)	4·4 (4·3–4·5)	<0·0001	2·8 (2·6–2·9)	5·4 (5·3–5·5)	<0·0001
Location (extra-adrenal)	55/327 (17%)	118/166 (71%)	<0·0001	53/238 (22%)	33/57 (58%)	<0·0001
Multifocal	67/327 (21%)	33/166 (20%)	0·087	41/238 (17%)	13/57 (23%)	0·32
Presence of *SDHB* mutation*	16/267 (6%)	74/149 (50%)	<0·0001	7/236 (3%)	15/56 (27%)	<0·0001
Previous history of pheochromocytomas and paragangliomas†	31/327 (10%)	116/166 (70%)	<0·0001	34/238 (14%)	40/57 (70%)	<0·0001
Biochemistry (pg/mL)						
Normetanephrine	598·3 (594–602)	832·9 (827–838)	0·026	549·5 (523·5–531)	526·1 (521–531)	0·81
Metanephrine	144·8 (139–150)	42·1 (38–45)	<0·0001	124·2 (118–129)	52·5 (48–56)	<0·0001
Methoxytyramine	13·6 (10–16)	46·2 (44–47)	<0·0001	15·1 (12–17)	49·5 (40–58)	<0·0001
Follow-up (months)	83 (79–85)	96 (92–98)	0·14	49 (46–51)	100 (96–103)	<0·0001

Data are n/N (%), geometric mean (95% CI), or p value. *Genetic testing was not available for 60 patients without metastases and for 17 with metastases in the training cohort and for two patients without metastases and one with metastases in the external validation cohort. †Local recurrence or new tumour. The PMT and PRESCRIPT trials were 100% White; the National Institutes of Health 00-CH-093 had White, Black African, and Asian participants.

Table 1: **Baseline characteristics of patients with pheochromocytomas and paragangliomas according to presence versus absence of metastatic disease for the training and external validation (PMT-trial) cohorts**

probabilities according to the same nine features described for machine learning feature analysis. After an interval of 4 weeks, all specialists received a second dataset with the same features, supplemented by *SDHB* variant status (appendix pp 4–5). Specialist's predictions were then compared with those of the top performing machine learning models (figure 1).

Statistical analysis

Continuous parametric variables were calculated as geometric means with 95% CIs. Comparisons of continuous parameters were done with the Mann-Whitney U test. Categorical parameters were analysed using the χ^2 test. Cox hazard regression models were used to identify predictive value of methoxytyramine for metastatic disease. The cut-off for plasma concentrations of methoxytyramine was determined using ROC analysis and the derived Youden index. Comparisons of the diagnostic performance of each participant in the two datasets with versus without the *SDHB* variant status were made using the Wilcoxon signed-rank sum test for paired measurements. Statistical analysis was done using JMP pro statistical software package version 15. No patients were lost to follow-up so no analyses were required to address loss. $p<0·05$ was considered significant.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

The PMT trial had 295 patients with pheochromocytoma or paraganglioma (ie, the external validation cohort). Among these there were 267 patients with positive biochemical test results, in whom predictions of metastatic disease were possible according to the prospective study design (appendix pp 6–8). The remaining patients comprised 28 with negative test results, among whom 23 had head and neck paragangliomas (appendix p 7). The 493 patients with pheochromocytoma or paraganglioma in the training cohort showed some demographic differences from the 295 patients in the PMT trial used for external validation (appendix p 6). Nevertheless, in both datasets, patients with metastases were more often male and younger than those without metastases (table 1). Patients with metastases had larger, extra-adrenal tumours than those without metastases. There was also a higher prevalence of *SDHB* variants and recurrent disease in those with metastases than those without. Finally, patients with metastases had lower metanephrine, but higher methoxytyramine concentrations than those without metastases. All differences were highly significant ($p<0·0001$).

211 patients (79%) of 267 were correctly classified by specialist-based predictions. Specifically, predictions were correct for 186 (85%) of 219 patients without metastases (specificity 85%) and 25 (52%) of 48 patients with metastases (sensitivity 52%; appendix p 7). Low sensitivity largely reflected patients with normal or mildly elevated plasma concentrations of methoxytyramine (appendix p 8). Among 13 patients classified by the specialist with a strong risk for metastases, 11 (85%) had metastases. The higher sensitivity in this particular category reflected high (>678 pg/mL) plasma concentrations of methoxytyramine in all 11 patients.

For the second outcome, feature and machine learning analyses were done in the training cohort twice. The first analysis included nine clinical and biochemical features,

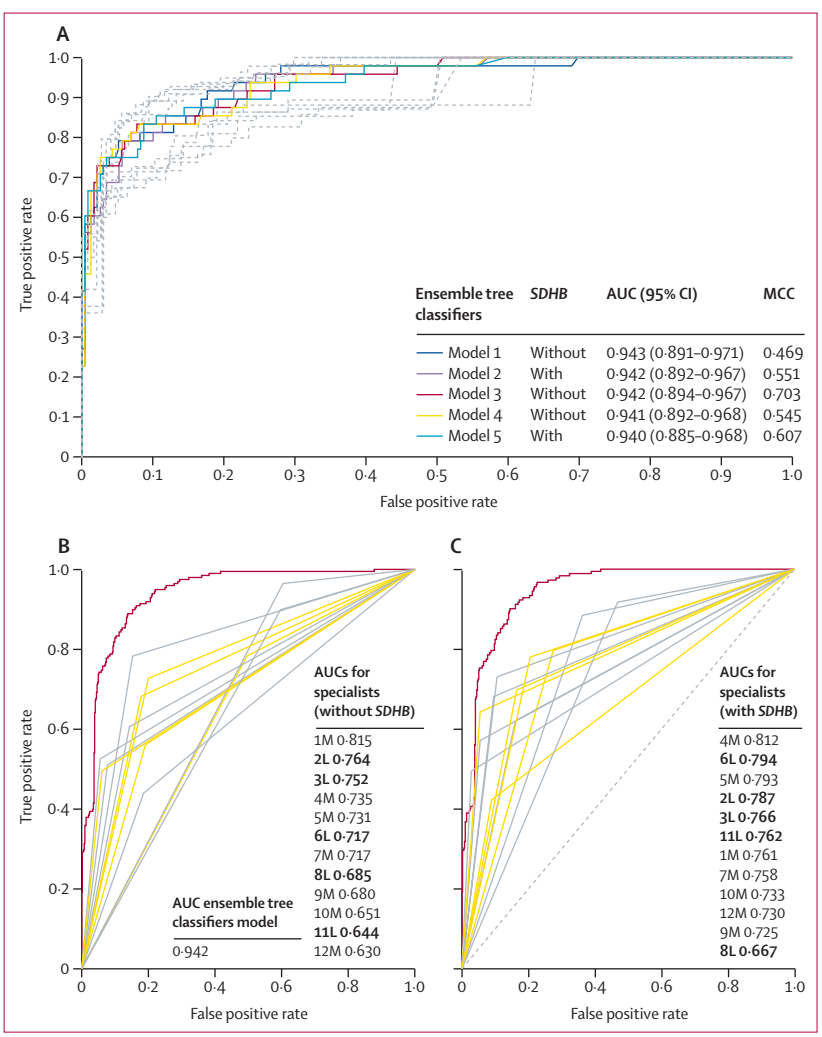

Figure 2: **ROC curves for performance**

(A) Predictive performance of the top five machine learning models after external validation according to ROC curves. AUCs for ROC curves are shown with 95% CIs (dotted curves). All five machine learning models had similar diagnostic performance in terms of AUC. The final selected machine learning model was model 3, which did not use *SDHB* pathogenic variant status and had high MCC and balanced accuracy metrics. (B, C) Comparison of the diagnostic performance of the selected ensemble tree classifiers model with that of the 12 clinical care specialists according to their interpretations of likely presence or absence of metastatic disease. The classification performance of the ensemble tree classifiers model, which was established without requirement of the *SDHB* pathogenic variant status, was significantly better than the performance of all specialists, both before (B) and after (C) provision of information about *SDHB* pathogenic variant status. Specialist ID numbers are indicated by 1–12 M or L. AUC=area under the curve. ROC=receiver operating characteristic. MCC=Matthew's correlation coefficient.

	Decision tree classifier	Support vector machine	Naive Bayes	AdaBoost ensemble tree classifiers
Data set with nine features (without *SDHB* mutation status)				
AUC	0·889 (0·823–0·934)	0·929 (0·889–0·957)	0·839 (0·752–0·891)	0·942 (0·894–0·969)
Matthew's correlation coefficient	0·863 (0·808–0·893)	0·795 (0·735–0·840)	0·710 (0·651–0·771)	0·851 (0·801–0·898)
F1-score	0·774 (0·699–0·863)	0·661 (0·549–0·770)	0·554 (0·417–0·610)	0·755 (0·667–0·833)
Sensitivity	0·854 (0·725–0·939)	0·813 (0·687–0·909)	0·854 (0·757–0·951)	0·833 (0·707–0·929)
Specificity	0·927 (0·894–0·957)	0·866 (0·831–0·914)	0·745 (0·690–0·808)	0·922 (0·893–0·955)
Precision	0·707 (0·599–0·841)	0·557 (0·465–0·691)	0·410 (0·308–0·506)	0·690 (0·568–0·834)
Accuracy	0·914 (0·879–0·939)	0·857 (0·812–0·889)	0·764 (0·723–0·805)	0·907 (0·861–0·932)
Balanced accuracy	0·890 (0·822–0·933)	0·839 (0·770–0·888)	0·799 (0·758–0·840)	0·878 (0·808–0·922)
Data set with ten features (with *SDHB* mutation status)*				
AUC	0·893 (0·823–0·936)	0·924 (0·881–0·953)	0·826 (0·751–0·878)	0·940 (0·886–0·969)
Matthew's correlation coefficient	0·849 (0·782–0·891)	0·795 (0·761–0·841)	0·672 (0·617–0·719)	0·804 (0·741–0·849)
F1-score	0·750 (0·635–0·826)	0·651 (0·533–0·726)	0·559 (0·423–0·695)	0·672 (0·571–0·780)
Sensitivity recall rates	0·750 (0·596–0·841)	0·896 (0·783–0·962)	0·791 (0·636–0·946)	0·854 (0·777–0·939)
Specificity	0·948 (0·908–0·974)	0·821 (0·771–0·869)	0·781 (0·726–0·836)	0·859 (0·801–0·900)
Precision	0·750 (0·578–0·834)	0·512 (0·408–0·602)	0·413 (0·293–0·546)	0·554 (0·454–0·674)
Accuracy	0·914 (0·877–0·942)	0·834 (0·785–0·866)	0·783 (0·749–0·832)	0·856 (0·800–0·897)
Balanced accuracy	0·849 (0·793–0·910)	0·858 (0·808–0·908)	0·786 (0·752–0·820)	0·855 (0·789–0·907)

Data are point estimate (95% CI). AUC=area under the curve of the receiver operating characteristic. *Information regarding the presence or not of *SDHB* mutation was included as an extra feature.

Table 2: **Classification performance of the top performing machine learning models established after external validation for each of the four machine learning algorithms using nine and ten features.**

whereas the second was supplemented with *SDHB* variant status. Among the nine features in the first analysis, the top five that predicted metastases included previous history of pheochromocytoma or paraganglioma, extra-adrenal primary tumour location, large primary tumour size, high plasma methoxytyramine concentration, and low plasma metanephrine concentration (appendix p 8). For the ten-feature analysis, the five most important features were previous history of pheochromocytomas and paragangliomas, extra-adrenal primary tumour location, presence of *SDHB* variants, large primary tumour size, and high plasma methoxytyramine concentration (appendix p 8).

Among the 380 initial machine learning models evaluated in the training cohort (appendix p 9), the 40 best performing machine learning models were selected for external validation. Comparisons of the AUC, with additional considerations of MCC and balanced accuracy after external validation, revealed five top performing machine learning models, all involving ensemble tree classifier algorithms (figure 2A). All five models showed similar diagnostic performance (appendix p 10). The best performing ensemble tree classifier model, which had an AUC of 0·942 (95% CI 0·894–0·969), an MCC of 0·851, and a balanced accuracy of 88%, did not use *SDHB* variant status as a feature (table 2, figure 2A). The second best was an ensemble tree classifier model that used *SDHB* variant status and had an AUC of 0·940 (0·886–0·969), an MCC of 0·804, and a balanced accuracy of 86%.

Three other algorithms (tree classifier, support vector machine, and Naive Bayes) provided machine learning models with predictive performance that approached that of the ensemble tree classifier algorithm-derived models (appendix pp 11–14). For the dataset that did not include *SDHB* variant status, the best tree classifier model had an AUC of 0·889 (95% CI 0·823–0·934), an MCC of 0·863, and a balanced accuracy of 89%. The best support vector machine model had an AUC of 0·929 (0·889–0·957), an MCC of 0·795, and a balanced accuracy of 84%. This model was followed by the the Naive Bayes model, with an AUC of 0·839 (0·752–0·891), an MCC of 0·710, and a balanced accuracy of 80% (table 2).

For the dataset supplemented with *SDHB* variant status, the best tree classifier model had an AUC of 0·893 (95% CI 0·823–0·936), an MCC of 0·849, and a balanced accuracy of 85%. The best support vector machine model had an AUC of 0·924 (0·881–0·953), an MCC of 0·795, and a balanced accuracy of 86%. This model was again followed by the Naive Bayes model with an AUC of 0·826 (0·751–0·878), an MCC of 0·672, and a balanced accuracy of 79% (table 2).

Among the 12 specialists who provided predictions of metastatic risk, predictive performance varied widely according to the nine-feature and ten-feature datasets without and with *SDHB* variant status (table 3). The highest performance among specialists for the dataset without *SDHB* variant status was by specialist 1M (AUC 0·815, 95% CI 0·778–0·853), whereas the highest

performance for the dataset supplemented with *SDHB* variant status was by specialist 4M (0·812, 0·781–0·854).

The diagnostic performance of specialists did not differ among those with more than 10 years' experience versus less than 10 years experience (table 3; appendix p 16). Specifically, for the nine-feature dataset, the specialists with more than 10 years' experience had a mean AUC of 0·708 (95% CI 0·648–0·768), which was similar (p=0·7550) to the AUC of 0·712 (0·662–0·772) for those with less experience. Likewise, for the ten-feature dataset supplemented by the *SDHB* variant status, the specialists with more than 10 years' experience had a mean AUC of 0·758 (0·728–0·788), which again was similar (p=0·6390) to the AUC of 0·755 (0·705–0·805) for those with less experience.

Paired comparisons revealed that only four specialists (4M, 6L, 11L, and 12M) improved their performance after the provision of *SDHB* variant status (table 3). Overall, neither specialists with more (p=0·0630) or less than 10 years' experience (p=0·1380) improved their performance after provision of *SDHB* variant status.

For the third outcome, among the 12 specialists, none attained the diagnostic performance reached by the ensemble tree classifier model (figure 2B). The average performance of specialists (AUC 0·710, 95% CI 0·655–0·765) was less (p<0·0001) than the performance of the ensemble tree classifier model (0·942, 0·894–0·969; figure 2B). After provision of *SDHB* variant status, average performance of specialists (0·756, 0·716–0·796) also remained inferior (p<0·0001) to that of the ensemble tree classifier model (figure 2C; appendix p 16).

	Dataset without *SDHB* status	Dataset with *SDHB* status	Paired comparisons (p value)
Selected ensemble tree classifier model	0·942 (0·894–0·969)	0·940 (0·885–0·968)	..
1M	0·815 (0·778–0·853)	0·761 (0·723–0·799)	0·10
2L	0·764 (0·723–0·805)	0·787 (0·747–0·828)	0·24
3L	0·752 (0·710–0·794)	0·766 (0·723–0·810)	0·11
4M	0·735 (0·689–0·781)	0·812 (0·781–0·854)	0·0001
5M	0·731 (0·687–0·776)	0·793 (0·749–0·836)	0·16
6L	0·717 (0·670–0·764)	0·794 (0·750–0·838)	0·0001
7M	0·717 (0·670–0·763)	0·758 (0·711–0·805)	0·14
8L	0·685 (0·639–0·731)	0·667 (0·617–0·717)	0·20
9M	0·680 (0·642–0·719)	0·725 (0·685–0·764)	0·18
10M	0·651 (0·609–0·694)	0·733 (0·684–0·782)	0·17
11L	0·644 (0·601–0·687)	0·762 (0·720–0·802)	0·0001
12M	0·630 (0·582–0·677)	0·730 (0·684–0·776)	0·0001

Data are AUC (95% CI) or p value. M refers to specialists with more than 10 years of experience, L refers to specialists with 5–10 years of experience. AUC=area under the curve of the receiver operating characteristic.

Table 3: **Performance of 12 specialists for the prediction of metastatic disease in patients with pheochromocytomas and paragangliomas, before versus after provision of the *SDHB* mutation status**

Discussion

This study introduces machine learning models to more accurately predict metastatic disease than previously possible. Importantly, these models allow for predictions at first diagnosis of pheochromocytoma or paraganglioma by use of clinical features that are routinely and preoperatively available. More generally, our findings support emerging concepts that machine learning mathematical processes will gain traction in medicine and oncology for their potential to facilitate robust non-invasive diagnostic stratification and guide personalised patient management.

The initial prospective assessments of plasma methoxytyramine as a predictor of metastatic disease confirmed previous retrospective findings.[2] However, the 52% of all patients correctly predicted by methoxtyramine with metastatic pheochromocytoma or paraganglioma is only a little better than the use of *SDHB* pathogenic variants for the same purpose, with a prevalence of up to 41% among patients with metastatic pheochromocytoma or paraganglioma.[13] This finding was supported in the present study by both training and external validation data, showing an overall prevalence of 40% for *SDHB* pathogenic variants among patients with metastatic disease. Nevertheless, among patients with highly increased levels of methoxytyramine, the post-test probability of metastatic disease was 85%. Overall, however, and similar to *SDHB* variant status, measurements of methoxytyramine alone cannot be used to accurately predict or exclude metastases.

Apart from methoxytyramine and *SDHB* pathogenic variants, several other features have been indicated as predictors of metastases among patients with pheochromocytoma or paraganglioma. Our second objective was to combine routinely available clinical and biochemical features with measurements of methoxytyramine to develop a machine learning tool to predict metastases preoperatively. Large tumour size, extra-adrenal tumour location, previous history of pheochromocytoma or paraganglioma, high plasma methoxytyramine concentration, and low metanephrine concentration were consistently identified to predict metastases according to the feature analysis. Those risk factors probably reflect a more undifferentiated tumour phenotype associated with pseudohypoxia signaling and hypermethylation pathways.[24,25] Those pathways might drive the mesenchymal transition step in metastatic progression.[26,27] Findings that high plasma methoxytyramine, but low plasma metanephrine predict metastases, and that both metabolites are among selected features, emphasise the importance of accurate and reliable biochemical tests carried out according to appropriate preanalytical and analytical procedures.

External validation of the best candidate machine learning models after internal testing identified the five best performing machine learning models with similar diagnostic performance and a mean AUC of 0·942 (95% CI 0·891–0·968). Among those models, the ensemble tree classifier model provided the best MMC and balanced accuracy metrics without requirement for

SDHB variant data, which are often not available at initial diagnosis. The finding that *SDHB* test results were not required for prediction of metastases is explained by the key features of large tumour size, extra-adrenal location, and noradrenergic or dopaminergic tumour phenotype shared between the group of patients with *SDHB*-mutated metastatic pheochromocytoma or paraganglioma and the more than twice larger group of all patients with metastatic pheochromocytoma or paraganglioma.[10,12]

Apart from establishing machine learning models that can be easily applied preoperatively using readily available clinical information, we also validated the models in a separate cohort of patients to establish reproducibility. Furthermore, after external validation we also compared the best performing machine learning models with interpretations by clinical care specialists with expertise in pheochromocytoma or paraganglioma. These comparisons established that the finally selected ensemble tree classifier model provided significantly improved performance over interpretations of all specialists.

Of course, one could argue that, in real life, clinicians incorporate more clinical information into decision making, and that the identified performance of specialists is artificial. In an attempt to partially eliminate this potential confounder, we investigated whether provision of *SDHB* variant status improved the ability of specialists to predict metastasis. Only four specialists showed improved performance; overall performance remained significantly inferior to that of the selected ensemble tree classifier machine learning model, which did not require *SDHB* variant status as a feature. Another argument to be considered is that clinicians focus more on management decisions rather than diagnostic classifications. In this context, we incorporated the management decisions into the study design and provided all specialists before the review process with narrative interpretations for further patient management for each of the four classification categories.

With the aforementioned considerations in mind, our data show that the selected machine learning models provide a suitable tool for prediction of metastases in patients with pheochromocytoma or paraganglioma. Furthermore, these models should be of benefit to clinicians at different levels of training and experience. The models could be implemented in digital systems or smart phone applications and used together with other routinely available data to facilitate individualised diagnostic stratification and patient management. Apart from identifying patients with low probabilities of metastases, who can then be excluded from intensive, long-term and costly follow-up programmes, our machine learning models provide justification for preoperative functional imaging and extensive follow-up in patients with high probability of metastases. In turn, the use of these models provides opportunities for earlier disease detection and interventional strategies for improved patient outcomes.

Despite growing acceptance of the superior predictive power of machine learning compared to conventional statistical scores for oncological staging,[28] many have considered machine learning a black box where connections between features and disease probabilities are invisible to clinicians.[29] These concerns are being addressed by interfaces that integrate data with clinical decision support systems to provide automated patient-specific interpretations and narrative reports to assist clinicians towards a decision.[30] Thus, machine learning-integrated decision-support systems are expected to facilitate further the smooth and trustworthy integration of machine learning technologies into the clinical setting.

Our study has four main limitations. The shorter duration of follow-up among patients without metastases compared with those with metastases in the prospective PMT cohort might have affected the importance of methoxytyramine to predict metastases by underestimating diagnostic sensitivity. We also did not develop machine learning models for patients with head and neck tumours separately from those with abdominal paragangliomas, which are known to have different characteristics. Third, the present data do not establish whether our machine learning models improve decision making and outcomes for patients, which requires a prospective clinical trial. Finally, another apparent study limitation is the omission from the machine learning analyses of histopathological, radiological, and somatic variant features that could have strengthened predictive value of machine learning models (appendix pp 17–18). Heterogeneity in radiological procedures and histopathological interpretations renders retrospective use of such features problematic. However, the higher the complexity of the machine learning models, the lower their applicability in routine clinical practice health-care settings.

Despite the aforementioned limitations, our study is the first to develop accurate machine learning models for the prediction of metastases in patients with pheochromocytoma or paraganglioma using routinely available data, and without need for genetic, imaging, or histopathological data. The high performance of our machine learning models was facilitated by the availability of complete and comprehensive data for the training and external validation cohorts and the long duration of follow-up in the training cohort, which minimised possibilities of misclassifying patients with metastatic risk among those without evidence of metastases. Importantly, the large number of patients included in the study and its multinational, multicentric design, supports the generalisability of our machine learning models. Finally, the reproducibility of the selected machine learning models was secured not only through external validation by a different patient cohort, but also through comparisons with the performance of clinical care specialists with expertise in the care of patients with pheochromocytoma or paraganglioma.

In conclusion, our study shows that, although plasma methoxytyramine provides some utility to predict metastases among patients with pheochromocytoma or paraganglioma, sensitivity is limited. However, incorporation of plasma methoxytyramine in machine learning models, along with other clinical features such as primary tumour location and size, provides a highly accurate, non-invasive approach to predict metastases in patients with pheochromocytomas and paragangliomas, and can thereby guide individualised patient management and follow-up strategies.

Contributors

CP, AF, and GE contributed to the conception and design of the study, and drafted the paper. CP, AMAB, TP, LM, AP, FB, MF, HJLMT, SN, GC, CK, RJdH, KW, HR, and AJ contributed to the enrolment of patients in the study, selection of samples, collection and interpretation of clinical data, and revised the paper. AF, KA, CP, and GE constructed the machine learning models and analysed the data. CP, AMAB, FB, MF, HJLMT, SN, GC, HR, JWML, MNK, KP, and GE provided interpretations with predictions of metastatic disease in the patient cohorts. FB, MF, HJLMT, MNK, SRB, MR, JWML, and KP revised the paper. CP, GE, and AF accessed and verified the data. All authors approved the final version of the manuscript. All authors had full access to the data in the study and were involved in data interpretation and writing of the report. All authors had final responsibility for the decision to submit for publication.

Declaration of interests

CP, AF, and GE declare a filed German patent A5914/TUD 017, with the title "Verfahren zur Vorhersage eines Nebennierentumours sowie eines Metastaserisikos mithilfe klinisch relevanter Parameter", relevant to this manuscript. All other authors declare no competing interests.

Data sharing

The data generated in this study are available on the online platform zenodo.org: https://doi.org/10.5281/zenodo.7749613.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft 314061271-TRR/CRC 205- 1/2 (to CP, FB, MF, SN, GC, CK, HR, SRB, JWML, and GE), by the Free State of Saxony and TU Dresden (to AF and KA), the National Institutes of Health (to TP, LM, and KP) and by the Clinical Research Priority Program of the University of Zurich for the CRPP HYRENE (to FB).

References

1 Tischler AS. Pheochromocytoma and extra-adrenal paraganglioma: updates. *Arch Pathol Lab Med* 2008; **132:** 1272–84.
2 Eisenhofer G, Lenders JW, Siegert G, et al. Plasma methoxytyramine: a novel biomarker of metastatic pheochromocytoma and paraganglioma in relation to established risk factors of tumour size, location and *SDHB* mutation status. *Eur J Cancer* 2012; **48:** 1739–49.
3 Li M, Prodanov T, Meuter L, et al. Recurrent disease in patients with sporadic pheochromocytoma and paraganglioma. *J Clin Endocrinol Metab* 2023; **108:** 397–404.
4 Mete O, Asa SL, Gill AJ, Kimura N, de Krijger RR, Tischler A. Overview of the 2022 WHO classification of paragangliomas and pheochromocytomas. *Endocr Pathol* 2022; **33:** 90–114.
5 Fassnacht M, Assie G, Baudin E, et al. Adrenocortical carcinomas and malignant phaeochromocytomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2020; **31:** 1476–90.
6 Lenders JWM, Kerstens MN, Amar L, et al. Genetics, diagnosis, management and future directions of research of phaeochromocytoma and paraganglioma: a position statement and consensus of the Working Group on Endocrine Hypertension of the European Society of Hypertension. *J Hypertens* 2020; **38:** 1443–56.
7 Pamporaki C, Prodanov T, Meuter L, et al. Determinants of disease-specific survival in patients with and without metastatic pheochromocytoma and paraganglioma. *Eur J Cancer* 2022; **169:** 32–41.
8 McMillan M. Identification of hydroxytyramine in a chromaffin tumour. *Lancet* 1956; **271:** 284.
9 van der Harst E, de Herder WW, de Krijger RR, et al. The value of plasma markers for the clinical behaviour of phaeochromocytomas. *Eur J Endocrinol* 2002; **147:** 85–94.
10 Pamporaki C, Hamplova B, Peitzsch M, et al. Characteristics of pediatric *vs* adult pheochromocytomas and paragangliomas. *J Clin Endocrinol Metab* 2017; **102:** 1122–32.
11 Parasiliti-Caprino M, Lucatello B, Lopez C, et al. Predictors of recurrence of pheochromocytoma and paraganglioma: a multicenter study in Piedmont, Italy. *Hypertens Res* 2020; **43:** 500–10.
12 Ayala-Ramirez M, Feng L, Johnson MM, et al. Clinical risk factors for malignancy and overall survival in patients with pheochromocytomas and sympathetic paragangliomas: primary tumor size and primary tumor location as prognostic indicators. *J Clin Endocrinol Metab* 2011; **96:** 717–25.
13 Brouwers FM, Eisenhofer G, Tao JJ, et al. High frequency of *SDHB* germline mutations in patients with malignant catecholamine-producing paragangliomas: implications for genetic testing. *J Clin Endocrinol Metab* 2006; **91:** 4505–09.
14 Monteagudo M, Martínez P, Leandro-García LJ, et al. Analysis of telomere maintenance related genes reveals *NOP10* as a new metastatic-risk marker in pheochromocytoma/paraganglioma. *Cancers (Basel)* 2021; **13:** 4758.
15 Thompson LD. Pheochromocytoma of the Adrenal gland Scaled Score (PASS) to separate benign from malignant neoplasms: a clinicopathologic and immunophenotypic study of 100 cases. *Am J Surg Pathol* 2002; **26:** 551–66.
16 Kimura N, Takayanagi R, Takizawa N, et al. Pathological grading for predicting metastasis in phaeochromocytoma and paraganglioma. *Endocr Relat Cancer* 2014; **21:** 405–14.
17 Wu D, Tischler AS, Lloyd RV, et al. Observer variation in the application of the Pheochromocytoma of the Adrenal Gland Scaled Score. *Am J Surg Pathol* 2009; **33:** 599–608.
18 Stenman A, Zedenius J, Juhlin CC. The value of histological algorithms to predict the malignancy potential of pheochromocytomas and abdominal paragangliomas – a meta-analysis and systematic review of the literature. *Cancers (Basel)* 2019; **11:** 225.
19 Cho YY, Kwak MK, Lee SE, et al. A clinical prediction model to estimate the metastatic potential of pheochromocytoma/paraganglioma: ASES score. *Surgery* 2018; **164:** 511–17.
20 Obermeyer Z, Lee TH. Lost in thought – the limits of the human mind and the future of medicine. *N Engl J Med* 2017; **377:** 1209–11.
21 Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial intelligence and surgical decision-making. *JAMA Surg* 2020; **155:** 148–58.
22 Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* 2019; **321:** 2003–17.
23 Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology – new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019; **16:** 703–15.
24 Qin N, de Cubas AA, Garcia-Martin R, et al. Opposing effects of HIF1α and HIF2α on chromaffin cell phenotypic features and tumor cell proliferation: insights from MYC-associated factor X. *Int J Cancer* 2014; **135:** 2054–64.
25 Letouzé E, Martinelli C, Loriot C, et al. SDH mutations establish a hypermethylator phenotype in paraganglioma. *Cancer Cell* 2013; **23:** 739–52.
26 Thienpont B, Steinbacher J, Zhao H, et al. Tumour hypoxia causes DNA hypermethylation by reducing TET activity. *Nature* 2016; **537:** 63–68.
27 Morin A, Goncalves J, Moog S, et al. TET-mediated hypermethylation primes SDH-deficient cells for HIF2α-driven mesenchymal transition. *Cell Rep* 2020; **30:** 4551–4566.e7.
28 Shimizu H, Nakayama KI. Artificial intelligence in oncology. *Cancer Sci* 2020; **111:** 1452–60.
29 Wang X, Wang D, Yao Z, et al. Machine learning models for multiparametric glioma grading with quantitative result interpretations. *Front Neurosci* 2019; **12:** 1046.
30 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:** 527–34.