Development and Validation of Prediction Models for Subtype Diagnosis of Patients With Primary Aldosteronism

Jacopo Burrello,[1,*] Alessio Burrello,[2,*] Jacopo Pieroni,[1] Elisa Sconfienza,[1] Vittorio Forestiero,[1] Paola Rabbia,[3] Christian Adolf,[4] Martin Reincke,[4] Franco Veglio,[1] Tracy Ann Williams,[1,4] Silvia Monticone,[1#] and Paolo Mulatero[1,#]

[1]Division of Internal Medicine and Hypertension, Department of Medical Sciences, University of Torino, 10126, Italy; [2]Department of Electrical, Electronic and Information Engineering "Guglielmo Marconi" (DEI), University of Bologna, 40126, Italy; [3]Division of Radiology, University of Torino, 10126, Italy; and [4]Medizinische Klinik und Poliklinik IV, Klinikum der Universität, Ludwig-Maximilians-Universität München, Munich, 80366, Germany

ORCiD numbers: 0000-0001-7884-7314 (J. Burrello); 0000-0002-6215-8220 (A. Burrello); 0000-0002-9206-0953 (C. Adolf); 0000-0002-9817-9875 (M. Reincke); 0000-0002-6654-6247 (F. Veglio); 0000-0002-2388-6444 (T. A. Williams); 0000-0002-7322-2005 (S. Monticone); 0000-0002-5480-1116 (P. Mulatero).

Context: Primary aldosteronism (PA) comprises unilateral (lateralized [LPA]) and bilateral disease (BPA). The identification of LPA is important to recommend potentially curative adrenalectomy. Adrenal venous sampling (AVS) is considered the gold standard for PA subtyping, but the procedure is available in few referral centers.

Objective: To develop prediction models for subtype diagnosis of PA using patient clinical and biochemical characteristics.

Design, Patients and Setting: Patients referred to a tertiary hypertension unit. Diagnostic algorithms were built and tested in a training (N = 150) and in an internal validation cohort (N = 65), respectively. The models were validated in an external independent cohort (N = 118).

Main outcome measure: Regression analyses and supervised machine learning algorithms were used to develop and validate 2 diagnostic models and a 20-point score to classify patients with PA according to subtype diagnosis.

Results: Six parameters were associated with a diagnosis of LPA (aldosterone at screening and after confirmatory testing, lowest potassium value, presence/absence of nodules, nodule diameter, and computed tomography results) and were included in the diagnostic models. Machine learning algorithms displayed high accuracy at training and internal validation (79.1%-93%), whereas a 20-point score reached an area under the curve of 0.896, and a sensitivity/specificity of 91.7/79.3%. An integrated flowchart correctly addressed 96.3% of patients to surgery and would have avoided AVS in 43.7% of patients. The external validation on an independent cohort confirmed a similar diagnostic performance.

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in USA

Received 22 March 2020. Accepted 11 June 2020.
First Published Online 19 June 2020.
Corrected and Typeset 31 August 2020.

[*] J.B. and A.B. contributed equally and should be considered as joint first authors.
[#] S.M. and P.M. contributed equally and should be considered as joint last authors.

Abbreviations: AVS, adrenal venous sampling; BP, blood pressure; BPA, bilateral aldosteronism; CT, computed tomography; DDD, daily defined dose; LDA, linear discriminant analysis; LI, lateralization index; LPA, lateralized primary aldosteronism; MRA, mineral receptor antagonist; OR, odds ratio; PA, primary aldosteronism; RF, random forest; ROC, receiver operating characteristic; SPACE, Subtyping Primary Aldosteronism by Clinical Evaluation.

doi:10.1210/clinem/dgaa379

Conclusions: Diagnostic modelling techniques can be used for subtype diagnosis and guide surgical decision in patients with PA in centers where AVS is unavailable. **(*J Clin Endocrinol Metab* 105: e3706–e3717, 2020)**

Key Words: aldosterone, primary aldosteronism, adrenal venous sampling; machine learning

Primary aldosteronism (PA) accounts for 3% to 13% of primary care hypertensive patients (1-3) and is associated with an increased cardio- and cerebrovascular risk compared with patients affected by essential hypertension (4, 5). The 2 major subtypes of PA are unilateral primary aldosteronism (lateralized [LPA]), mainly from an aldosterone-producing adenoma, and bilateral primary aldosteronism (BPA). The treatments of choice are unilateral adrenalectomy, or medical therapy with a mineralocorticoid receptor antagonist, respectively (6). A timely and accurate subtype diagnosis is critical to recommend the appropriate treatment and improving the outcomes of these patients (7, 8).

Over the past few decades, many procedures have been proposed to differentiate LPA from BPA, including posture testing, functional imaging (using 11-C-metomidate or 68-Ga-pentixafor tracers (9, 10) and steroid profiling (11-13)). Nevertheless, technical issues and/or the lack of sensitivity and specificity hampered the introduction of these tests in the routine management of PA patients (6).

Adrenal venous sampling (AVS) is currently considered the gold standard for subtype diagnosis (6). Nevertheless, several concerns prevent its widespread use: AVS is an invasive, time-consuming, and relatively expensive procedure, requiring a high level of technical skill and is available only in a limited number of referral centers (14). Beside AVS, adrenal computed tomography (CT) scanning is widely available in most centers and performed in all patients with confirmed PA (6). Even if several studies reported unreliable diagnostic performance of CT in PA subtyping (15, 16), score-based algorithms combining imaging findings with clinical and biochemical parameters have been developed (17-23). Küpers et al. first proposed a prediction score to bypass AVS; patients with a potassium < 3.5 mmol/L, an estimated glomerular filtration rate > 100 mL/min and a typical adenoma at CT imaging could avoid AVS. Sensitivity and specificity were 53.1% and 100%, respectively (17). Other clinical scores were subsequently developed in the attempt of differentiating LPA from BPA, with an accuracy ranging from 58.2% to 86.3% (18-23). Only 2 of these scores were also validated in independent cohorts (17, 22), 5 included fewer than 100 patients in the development cohort (17-20, 23), and the majority of these scores was applicable only in selected cohorts of patients with PA (18-21, 23).

Considering the high prevalence of PA and the limited availability of AVS, an alternative method that reduces the number of requested AVS is highly desirable. Our objective was to develop and validate clinical models to discriminate LPA from BPA, which can bypass AVS for bilateral disease, and indicating unilateral adrenalectomy for patients with high probability of LPA who cannot undergo AVS. We propose herein 2 advanced diagnostic models based on supervised machine learning algorithms and a flow-chart integrating our score-system (the Subtyping Primary Aldosteronism by Clinical Evaluation score [SPACE]) in PA patient management. Validation of previously described score-based algorithms is also provided and demonstrates the superiority of our prediction models.

Methods

Data analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request. A detailed description of patient management and data extraction, statistical analyses, and diagnostic modelling is provided as supplemental Data (24) (available at https://github.com/ABurrello/SPACE-score).

Study cohort and data extraction

We retrospectively evaluated a cohort of 215 patients referred to the hypertension unit of the University of Torino between 2008 and 2019, to train and test the diagnostic models. Eligible patients were randomly assigned to the training cohort (N = 150) or to the internal validation cohort (N = 65). An independent cohort of 118 consecutive patients from the Munich Klinikum der Universität treated between 2008 and 2014 was used for external validation. PA was diagnosed according to the Endocrine Society Guideline (6). Inclusion criteria were (1) confirmed diagnosis of PA and (2) successful AVS for subtype diagnosis.

Unilateral PA mainly depends on unilateral aldosterone-producing adenoma. However, a lateralization at AVS may also occur also in the presence of a dominant lesion with asymmetrical autonomous aldosterone production in the context of bilateral adrenal alterations, including aldosterone producing cell clusters, or diffuse/nodular hyperplasia. For this reason, the AVS-based term of "lateralized PA" was used to indicate a prevalently unilateral disease throughout the present study.

Statistical analysis

IBM SPSS Statistics 22 (IBM Corp., Armonk, NY) was used for statistical analyses. Data distribution was assessed by the Kolmogorov-Smirnov test. Normally distributed variables were analyzed by Student *t* test and reported as mean ± SD. Non-normally distributed variables were analyzed by Mann-Whitney test and reported as median [interquartile range]. Categorical variables were analyzed by χ^2 test and reported as absolute number and proportion (%). Univariate logistic regression analysis was used to define the odds ratios (ORs) for each analyzed parameter. Six selected variables were included in the multivariate logistic regression analysis. An OR > 1 is associated with an increased likelihood of the defined outcome (diagnosis of LPA), an OR < 1, a decreased likelihood. A P value < 0.05 was considered significant.

Diagnostic modelling

Python 3.5 (library, scikit-learn) was used for the development and validation of diagnostic models by machine learning techniques. Supervised machine learning algorithms are widely used in clinical research to formulate predictions about possible outcomes based on a predefined set of labeled paired input-output training sample data (13, 25). Supervised machine learning and in particular linear discriminant analysis (LDA) and random forest (RF) algorithms were applied on the combined cohort to develop diagnostic models able to discriminate patients with LPA vs. BPA. Six selected variables were used in both models (aldosterone at screening and after confirmatory testing, lowest potassium recorded in the absence of diuretic therapy, presence/absence of a nodule at CT scanning imaging, nodule diameter, and descriptive CT scanning findings).

LDA maximizes the separation between groups by increasing precision estimates by variance reduction. The algorithm computes a set of coefficients for linear combination of each variable to classify patients according to their diagnosis; a canonical plot was used to represent diagnostic performance of the LDA model. The RF model was composed of 20 classification trees with a maximum number of 8 splits for each tree. The predicted diagnosis was defined on the basis of the outcome of each classification tree of the RF: if at least 11 of 20 trees of the forest predict the diagnosis of lateralized PA, the patient is classified as LPA. The RF model was integrated in a free downloadable tool that allows the application of the algorithm in clinical practice (available at https://github.com/ABurrello/SPACE-score/raw/master/Random_Forest_model.zip).

Performance and generalizability of both LDA and RF models were evaluated by a 10K-cross validation algorithm (see extended methods—supplemental Data) (24).

The 6 variables were used to develop a 20-point score to predict the diagnosis of LPA. Variables were categorized, points were assigned to each reference interval, and cutoffs were derived to achieve the best accuracy in an automated way. The SPACE score was generated using the training cohort and tested with both internal and external validation cohorts. Receiver operating characteristic (ROC) curve analysis was used to assess the area under the curve and derive the best cutoff to discriminate patients with LPA by evaluation of the Younden Index (J = sensitivity + specificity − 1). A second online tool was developed to automatically calculate the score and the predicted diagnosis (available at https://github.com/ABurrello/SPACE-score/raw/master/SPACE%20Score%20Calculator.xlsm).

Results

Patient characteristics

Two hundred and fifteen patients were included in the analyses from the developmental cohort of Torino, 133 with a diagnosis of LPA and 82 with BPA. Clinical and biochemical characteristics are reported in Table 1. The mean age at diagnosis was 49 ± 9.5 years, mean blood pressure (BP) was 164/99 mm Hg, with a duration of hypertension of 68 [27; 128] months. Patients with a diagnosis of LPA were more frequently females (42.1% vs. 23.2%; $P = 0.005$), had a higher daily defined dose (DDD) (3.8 [2.2; 5.7] vs. 3.0 [1.3; 4.7] $P = 0.027$) and a lower potassium level (3.1 ± 0.6 vs. 3.8 ± 0.4; $P < 0.001$). At the diagnostic workup, patients with LPA displayed higher levels of aldosterone, both at screening (38.0 [25.7; 49.7] vs. 28.7 [19.8; 37.9] ng/dL; $P < 0.001$) and postconfirmatory testing (20.5 [13.3; 32.9] vs. 11.5 [8.2; 17.7] ng/dL; $P < 0.001$). To confirm PA diagnosis, 165 patients underwent saline infusion testing (76.7%), and 50 had captopril challenge testing (23.3%). CT scanning demonstrated the presence of a defined nodule in 85.7% of patients with LPA; a nodule was also detected in 41.5% of patients with bilateral disease (unilateral nodule in 29 of 34 cases, bilateral in 5). In patients with BPA, CT scanning was bilaterally normal in 24.4% of patients, bilaterally abnormal in 22%, and with a unilateral abnormality in 53.7% of the cases (see the supplemental Data for details on adrenal CT scanning interpretation and definition of nodule) (24). Among the 37 patients with bilateral abnormalities at CT scanning, 40.6% displayed a unilateral nodule in the context of bilateral adrenal thickening or contralateral thickening, 37.8% bilateral nodules, and 21.6% bilateral hyperplasia. Prevalence of target organ damage (estimated glomerular filtration rate, microalbuminuria, and left ventricular hypertrophy at echocardiography) and prior cardiovascular events was not significantly different between groups. As expected, the lateralization index (LI) at AVS was significantly higher in LPA than BPA patients (12.0 [6.9; 21.3] vs. 1.8 [1.3; 2.6]; $P < 0.001$). According to the Primary Aldosteronism Surgery Outcome criteria (7), after a follow-up of 6 to 12 months from unilateral adrenalectomy, patients with LPA displayed complete clinical and biochemical success in 54.1% and 98.5% of cases, respectively.

Table 1. Patient Characteristics of Study Cohort

Variable	LPA (N = 133)	BPA (N = 82)	*P* Value
Female sex, n (%)	56 (42.1)	19 (23.2)	**0.005**
Age at diagnosis (years)	49 ± 10.5	50 ± 7.7	0.248
Duration of HTN (months)	74 [27; 168]	63 [22; 123]	0.284
Systolic BP (mm Hg)	165 ± 25.0	163 ± 20.5	0.613
Diastolic BP (mm Hg)	99 ± 14.5	99 ± 11.7	0.873
Antihypertensive medication (DDD)	3.8 [2.2; 5.7]	3.0 [1.3; 4.7]	**0.027**
eGFR (mL/min)	96 [81; 109]	94 [80; 102]	0.146
Lowest potassium (mEq/L)	3.1 ± 0.6	3.8 ± 0.4	**<0.001**
PRA at screening (ng/mL/h)	0.30 [0.20; 0.40]	0.20 [0.10; 0.40]	0.554
Aldosterone at screening (ng/dL)	38.0 [25.7; 49.7]	28.7 [19.8; 37.9]	**<0.001**
Confirmatory testing	102 (76.7)	63 (76.8)	0.982
Saline infusion test, n (%)	31 (23.3)	19 (23.2)	
Captopril challenge test, n (%)			
PRA postconfirmatory test (ng/mL/h)	0.15 [0.10; 0.20]	0.15 [0.10; 0.21]	0.850
Aldosterone postconfirmatory test (ng/dL)	20.5 [13.3; 32.9]	11.5 [8.2; 17.7]	**<0.001**
Microalbuminuria, n (%)	42 (31.5)	24 (29.4)	0.800
LVH at echo, n (%)	81 (60.7)	48 (59.1)	0.831
CV events, n (%)	17 (12.6)	15 (18.1)	0.320
Presence of nodule at CT scanning, n (%)	114 (85.7)	34 (41.5)	**<0.001**
Largest nodule at CT scanning (diameter, mm)	14 (10, 20)	12 (10, 19)	0.315
CT scanning findings	5 (3.8)	20 (24.4)	**<0.001**
Bilaterally abnormal	19 (14.2)	18 (22.0)	
Bilaterally abnormal	109 (82.0)	44 (53.7)	
Unilateral abnormality			
AVS protocol	43 (32.3)	37 (45.1)	0.051
Basal, n (%)	51 (38.4)	32 (39.0)	
ACTH continuous infusion, n (%)	39 (29.3)	13 (15.9)	
Both (basal + ACTH), n (%)			
Lateralization Index at AVS	12.0 [6.9; 21.3]	1.8 [1.3; 2.6]	**<0.001**
Clinical outcome: complete, n (%)	72 (54.1)	NA	NA
[only for LPA] Partial, n (%)	55 (41.4)		
Absent, n (%)	6 (4.5)		
Biochemical outcome: complete, n (%)	131 (98.5)	NA	NA
[only for LPA] Partial, n (%)	2 (1.5)		
Absent, n (%)	0 (0.0)		

Clinical characteristics of patients included in the analysis stratified for diagnosis: patients with lateralized PA (LPA; N = 133) vs. bilateral PA (BPA; N = 82). The DDD is the assumed average maintenance dose per day for a drug used for its main indication in adults. Normally and non-normally distributed variables were reported as mean ± standard deviation or median [interquartile range], respectively. Categorical variables were reported as absolute number (n) and proportion (%).

AVS, adrenal venous sampling; BP, blood pressure; CT, computed tomography; CV, cardiovascular; DDD, defined daily dose; Echo, echocardiography; eGFR, estimated glomerular filtration rate; HTN, hypertension; LVH, left ventricular hypertrophy; PRA, plasma renin activity.

Univariate logistic regression analysis was performed including all parameters (supplemental Table 1) (24), showing a significant association with a diagnosis of LPA of female sex (OR 2.41), duration of hypertension (OR 1.01), DDD (OR 1.18), potassium (OR 0.10), aldosterone at screening (OR 1.01) and after confirmatory testing (OR 1.01), presence of a nodule at CT scanning (OR 8.33), nodule diameter (OR 1.12), and CT findings (OR 9.91). Six of these 9 variables were selected considering their discriminative performance and introduced in the multivariate model, which confirmed a highly significant independent association with the diagnosis of LPA for all parameters (Table 2).

Linear discriminant analysis model

The 6 selected variables confirmed by the multivariate regression analysis were used in an LDA model. The linear combination of variables included in the LDA is shown in the canonical plot (Fig. 1A). Each point represents a patient and the clear separation according to their subtype diagnosis indicates that the model can discriminate LPA from BPA with reliable accuracy. In particular, 175 of 215 patients (accuracy 81.4%) were correctly classified, with a sensitivity and specificity for LPA detection of 86.5% and 73.2%, respectively (Fig. 1B). To exclude overfitting bias and assess how the model could generalize in an independent cohort, the LDA was validated by a 10K-cross validation algorithm (see extended methods—supplemental Data) (24). The cross-validation showed a high predictive

Table 2. Selected Discriminant Variables for a Diagnosis of Lateralized PA

Variable	Univariate Analysis		Multivariate Analysis	
(ref. LPA)	OR (CI 95%)	*P* Value	OR (CI 95%)	*P* Value
Aldosterone at screening (ng/dL)	1.04 (1.02–1.07)	**< 0.001**	1.05 (1.01–1.10)	**0.017**
Lowest potassium (mEq/L)	0.10 (0.05–0.21)	**< 0.001**	0.09 (0.03–0.30)	**< 0.001**
Aldosterone post-confirmatory test (ng/dL)	1.09 (1.05–1.12)	**< 0.001**	1.09 (1.02–1.16)	**0.012**
Nodule at CT scanning (ref. presence)	8.33 (4.35–16.67)	**< 0.001**	12.50 (2.94–47.62)	**0.001**
Largest nodule at CT scanning (diameter, mm)	1.12 (1.07–1.16)	**< 0.001**	1.11 (1.06–1.16)	**0.013**
CT scanning findings (ref. unilateral abnormality)	9.91 (3.50–28.05)	**< 0.001**	4.44 (1.30–13.21)	**0.016**

Logistic regression analysis was performed to assess the OR and the 95% CI for each variable. Univariate and multivariate analysis are shown as indicated. An OR > 1 indicates an increased likelihood of lateralized PA (LPA), and an OR less than 1 a decreased likelihood. Aldosterone at screening, lowest potassium, aldosterone post-confirmatory test, and largest nodule at CT were treated as continuous variables. An OR increase of 0.01 represents a 1% increased likelihood of a diagnosis of LPA for each unit of the reference variable. Presence/absence of nodule at CT scanning, and CT scanning findings were treated as categorical variables.

CI, confidence interval; CT, computed tomography; OR, odds ratio; PA, primary aldosteronism.

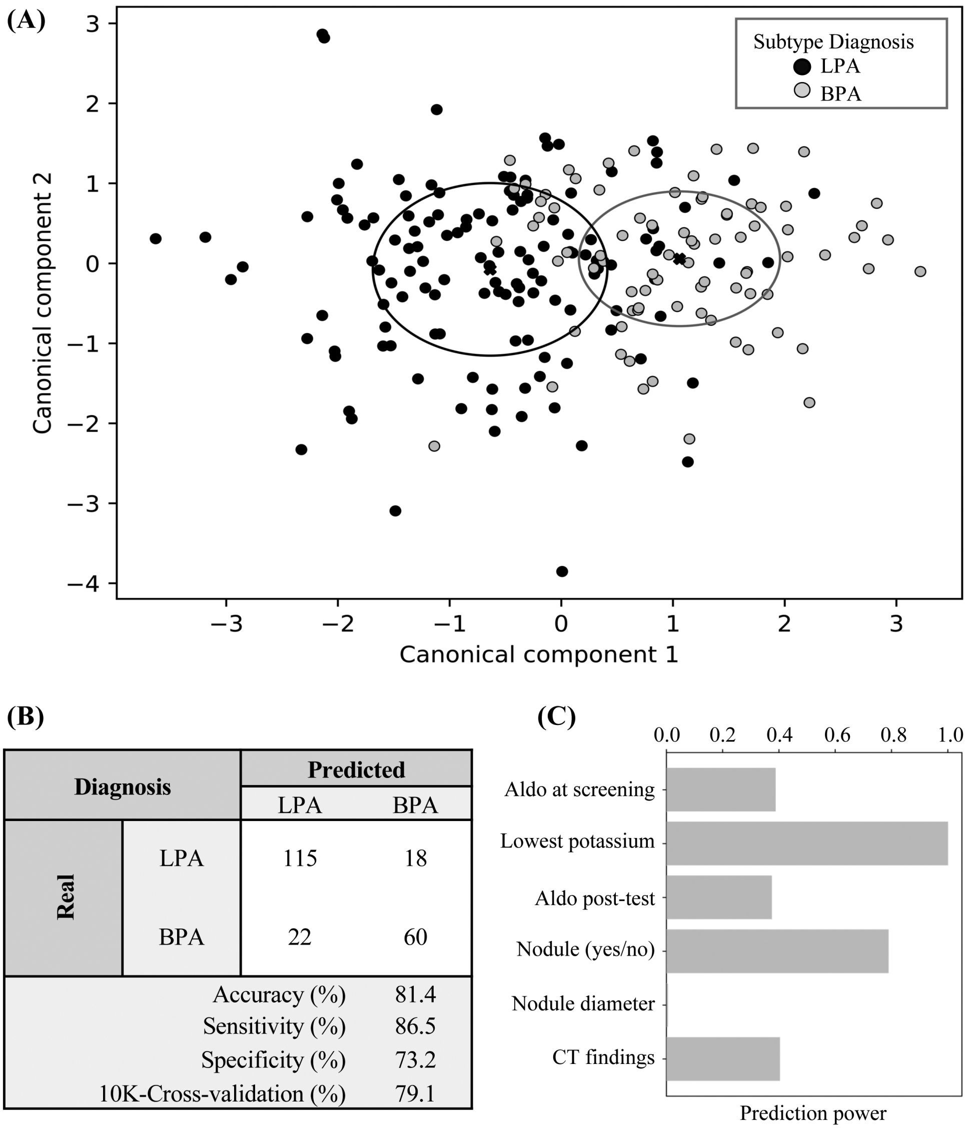
performance with an accuracy of 79.1%, compared with the 81.4% (at training), thus confirming a negligible overfitting bias (overfitting effect = 2.3%). In the LDA model, the stronger predictor was the lowest potassium level (normalized LDA coefficient = 1.0), followed by presence of a defined nodule and CT findings (0.8 and 0.4, respectively; Fig. 1C and supplemental Table 2) (24).

Random forest model

Besides LDA, we also developed a nonlinear classification model, exploiting RF classification algorithms. The same 6 selected variables were combined in an RF model comprising 20 classification trees (a representative tree is reported in Fig. 2A), and were able to correctly discriminate 132 of 133 patients with LPA (sensitivity 99.2%), and 68 of 82 patients with BPA (specificity 82.9%), resulting in an overall accuracy of 93% at the training and of 87% after 10K-cross validation (overfitting effect 6%; Fig. 2B). In this case, the stronger predictor was nodule diameter, followed by the lowest potassium level and by the presence of a nodule at CT scanning (Fig. 2C).

Prediction score

Patients included in the described models (combined cohort; N = 215) were randomly assigned to a training cohort (N = 150) or internal validation cohort (N = 65). No differences were found for all evaluated parameters between the 2 groups (supplemental Table 3) (24). The same 6 variables used in the LDA and RF models were then used to develop the SPACE score, a 20-point score to discriminate patients with LPA vs. BPA. The SPACE score was developed in the training cohort and then tested in the internal validation cohort. Fig. 3A and 3C report the categorization of the 6 different variables and assignment of points. The analysis of the ROC curve demonstrated a high diagnostic performance (Fig. 3B). The area under the curve was 0.896 (95% confidence interval, 0.852-0.940) and the cutoff with the higher accuracy was 12. At the training, a score > 12 correctly identified a diagnosis of LPA in 87 of 93 patients (sensitivity 93.5%), whereas a score ≤ 12 identified a diagnosis of BPA in 47 of 57 patients (specificity 82.5%), with an overall accuracy of 89.3%. Of note, the prediction score displayed a very high performance with an accuracy of 81.5% at validation, and a sensitivity and specificity of 87.5% and 72%, which was not significantly different from the machine learning models (accuracy at validation of 79.1% and 87% for the LDA and the RF model, respectively). Confusion matrix for training, internal validation, and combined developmental cohort are reported in Fig. 3D. The difference between the accuracy of the prediction score in the training cohort compared with the internal validation cohort, revealed a modest bias due to an expected overfitting effect (7.8%), which did not affect the reliability of the model. A cutoff of > 8 or of > 16 optimized sensitivity or specificity, respectively (supplemental Table 4) (24). With a cutoff of 8, sensitivity was increased to 97.8% and 95%, correctly classifying 91 of 93, and 38 of 40 patients with LPA, in the training cohort and in the validation cohorts, respectively. With a cutoff of 16, specificity was increased to 98.2% and 92%, correctly classifying 56 of 57, and 23 of 25 patients with BPA, in the training cohort and in the validation cohort, respectively.



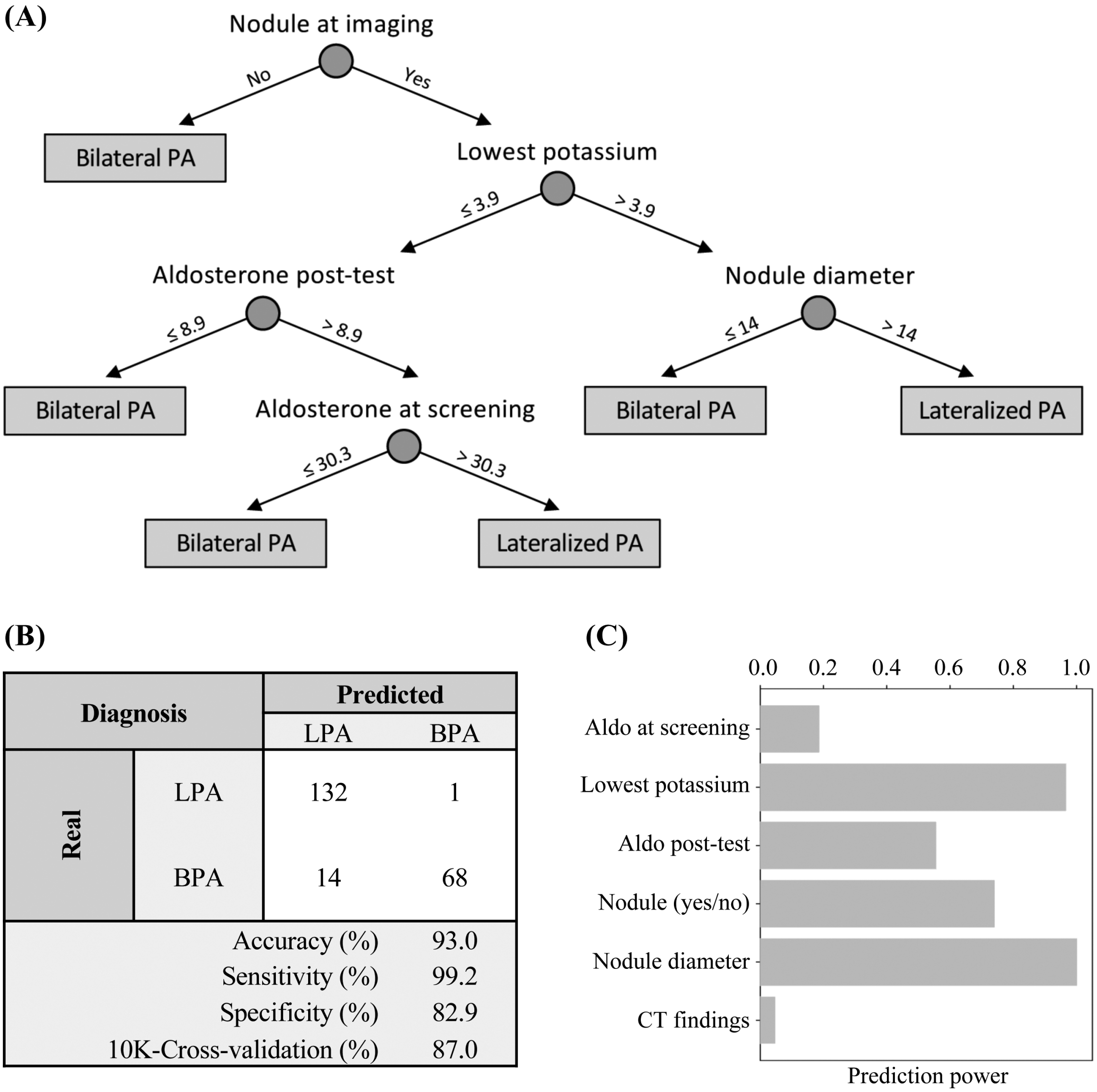
Diagnosis		Predicted	
		LPA	BPA
Real	LPA	115	18
	BPA	22	60
	Accuracy (%)		81.4
	Sensitivity (%)		86.5
	Specificity (%)		73.2
	10K-Cross-validation (%)		79.1

Figure 1. Diagnostic modelling: LDA. The LDA model included the 6 variables with the highest classification power for subtype diagnosis in the combined cohort (N = 215). (A) Canonical plot representing diagnostic performance of LDA; each patient is indicated by a point and subtype diagnosis are reported by color (LPA, lateralized PA, black; BPA, bilateral PA, gray). The axes (canonical component 1 and 2) are calculated by weighted linear combination of the 6 variables included in the model to maximize the separation between groups. The crosses indicate the means of (canonical 1; canonical 2) for patients with LPA or BPA, the ellipse included patients with a linear combination coefficient that falls within the mean ± SD. (B) Confusion matrix reporting real and predicted diagnosis, accuracy, sensitivity, specificity, and 10K-cross validation. (C) Histogram representing normalized LDA coefficients for each variable included in the model. CT, computed tomography; LDA, linear discriminant analysis; PA, primary aldosteronism.

To evaluate further the diagnostic performance of the SPACE score, 7 previously published scores (17-23) were tested on our combined cohort (supplemental Table 5) (24). The accuracy of our prediction score (89.3% and 81.5% at training and internal validation analysis, respectively) was superior to all available scores (accuracy ranging from 58.2% to 86.3% at training and from 67.3% to 78% at validation). Of note, the RF classification algorithm outperformed all other models with an accuracy of 87% at validation, higher than all score evaluated at training.

External validation

LDA, RF model, and the SPACE score were validated on an external independent cohort from Munich of 118 patients, 57 with LPA and 61 with BPA (supplemental Table 6) (24). Compared with the developmental cohort, the prevalence of LPA was significantly lower in the external validation cohort (48.3% vs. 61.9%; $P = 0.017$) and mean BP (153/94 mm Hg vs. 164/99 mm Hg), DDD (2.5 [1.0; 4.0] vs. 3.3 [2.0; 5.0]), potassium levels (3.1 ± 0.5 mEq/L vs. 3.4 ± 0.7 mEq/L) were also significantly lower. PRA at screening (0.29 vs. 0.25 ng/mL/h) and after confirmatory testing



Diagnosis		Predicted	
		LPA	BPA
Real	LPA	132	1
	BPA	14	68
	Accuracy (%)		93.0
	Sensitivity (%)		99.2
	Specificity (%)		82.9
	10K-Cross-validation (%)		87.0

Figure 2. Diagnostic modelling: RF. The RF algorithm included the 6 variables with the highest classification power for subtype diagnosis in the combined cohort (N = 215). (A) The first classification tree of the forest is shown for the prediction of LPA (lateralized PA) vs. BPA (bilateral PA). (B) Confusion matrix reporting real and predicted diagnosis, accuracy, sensitivity, specificity, and 10K-cross validation. (C) Histogram representing normalized predictive coefficients for each variable included in the model. CT, computed tomography; PA, primary aldosteronism; RF, random forest.

(0.21 vs 0.15 ng/mL/h) was significantly higher and aldosterone levels at screening (17.9 vs 33.4 ng/dL) and after confirmatory testing (11.2 and 16.4 ng/dL) were significantly lower ($P < 0.05$ for all comparisons) in the validation compared with the developmental cohort. The reliability of the diagnostic performance of our prediction models was confirmed at external validation. The accuracy was 78.8%, 80.5%, and 78.8%, respectively for LDA, RF, and the score system (supplemental Figure 1) (24), with a minimum overfitting bias compared with the internal validation on the developmental cohort (range between 0.3% and 6.5%).

Management of PA patient

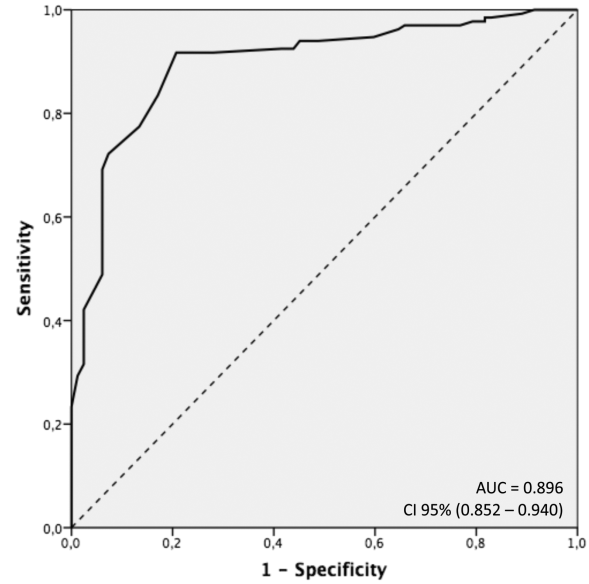
The SPACE score was directly correlated with the proportion of patients with a diagnosis of LPA (supplemental Table 7) (24) and with the LI at the AVS (supplemental Table 8) (24). Figure 4A clearly illustrates the stratification of patients with a diagnosis of LPA vs. BPA for the prediction score and graphically confirmed the cutoffs of 8, 12, and 16, which maximize sensitivity, accuracy, and specificity, as defined by ROC curve analysis. In addition, all patients with a score > 18 had LPA, whereas all patients with a score ≤ 2 had BPA.

Finally, our score was integrated in a flowchart for PA management (Fig. 4B). Patients with a score ≤ 8 were classified as "probable BPA" and treated with

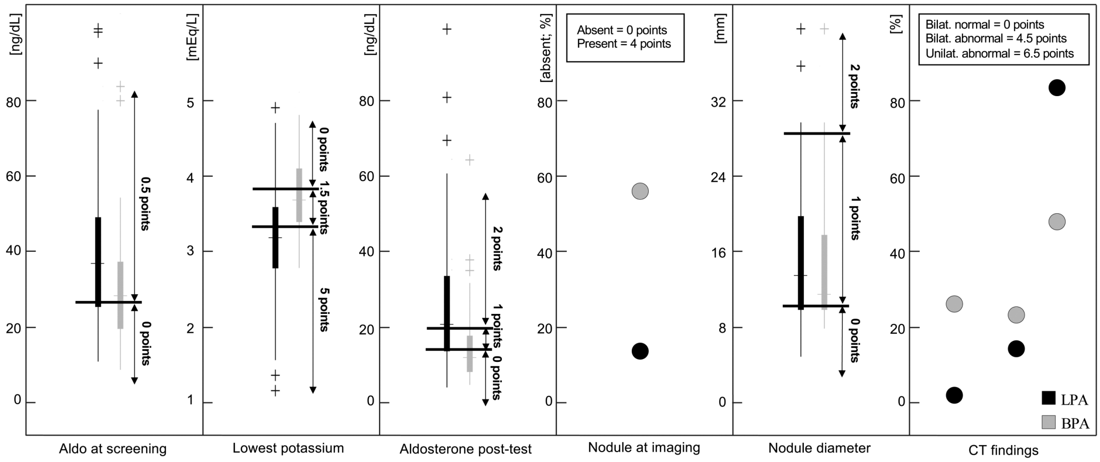
(A)

Variable		Label	Points
Aldosterone at screening (ng/dL)		</= 25	0
		> 25	0.5
Lowest Potassium (mEq/L)		< 3.4	5
		3.4 - 3.9	1.5
		>/= 4	0
Aldosterone post-confirmatory test (ng/dL)		</= 15	0
		15.1 – 19.9	1
		>/= 20	2
Nodule at CT scanning		Yes	4
		No	0
Largest nodule at CT scanning (diameter, mm)		</= 10	0
		11-30	1
		> 30	2
CT scanning Findings	Bilaterally Normal		0
	Bilaterally Abnormal		4.5
	Unilateral Abnormality		6.5

(B)



(C)



(D)

	AVS Score Accuracy	Predicted Diagnosis		Performances	
Real Diagnosis (Cut-off > 12)	**Training cohort** (N = 150)	LPA	BPA	Accuracy (%)	89.3
	LPA	87	6	Sensitivity (%)	93.5
	BPA	10	47	Specificity (%)	82.5
	Validation cohort (N = 65)	LPA	BPA	Accuracy (%)	81.5
	LPA	35	5	Sensitivity (%)	87.5
	BPA	7	18	Specificity (%)	72.0
	Combined cohort (N = 215)	LPA	BPA	Accuracy (%)	87.0
	LPA	122	11	Sensitivity (%)	91.7
	BPA	17	65	Specificity (%)	79.3

Figure 3. Score development and validation. Univariate/multivariate regression analyses and coefficients from the LDA and RF models were used to assign points to each variable according to stratification level. The score was developed in the training cohort (N = 150) and tested on the validation cohort (N = 65). (A) Table showing included variables and final point system used for the score. (B) Receiver operating characteristics (ROC) curve to assess AUC (area under the curve) and the best cutoff for the score in the combined cohort (N = 215). (C) Representation of cutoffs and assigned points for each variable after categorization: subtype diagnosis is represented by colors (LPA, lateralized PA, black; BPA, bilateral PA, gray); the bars indicate median and interquartile range for each group. (D) Confusion matrix representing real and predicted subtype diagnosis, accuracy sensitivity, specificity for the training cohort (N = 150), the validation cohort (N = 65), and the combined cohort (N = 215). CI, confidence interval; CT, computed tomography; PA, primary aldosteronism.

mineral receptor antagonist (MRA) (N = 32), thus resulting in 28 patients with true bilateral disease correctly managed, and 4 patients with an LPA (1.9%) that missed the possibility to undergo adrenalectomy. Patients with a score > 16 were classified as "probable LPA," with indication to unilateral adrenalectomy (N = 62). Accordingly, 3 patients with bilateral disease would undergo inappropriate surgery (1.4%), and 1 patient with LPA would have resection of the wrong adrenal (0.5%). All remaining patients (N = 121), with a score between 8.5 and 16 would undergo AVS with management according to the result of the procedure. Sensitivity, specificity, and positive and negative predictive values are reported in the confusion matrix (Fig. 4C).

We combined patients from the developmental and external validation cohorts (N = 333) and stratified these patients into 3 groups according to the points of the SPACE score (score ≤ 8 vs. 8.5 to 16 vs. > 16): the median LI displayed a gradual increase in the 3 groups of patients (supplemental Table 8) (24). Moreover, clinical and biochemical outcomes in patients with LPA misclassified as BPA were worse than patients with a correct prediction of LPA (83.3% vs. 48.8% partial + absent clinical success, and 5.6% vs. 0.6% partial + absent biochemical success; supplemental Table 9) (24).

After stratification for the confirmatory test performed during the diagnostic workup (supplemental Table 10) (24), the SPACE score confirmed its applicability both for patients diagnosed by saline infusion testing (accuracy 84%) or captopril challenge testing (accuracy 84.6%).

The application of the prediction score in our clinical context would result in the correct management of 207 of 215 patients (96.3%) with a reduction of 43.7% (94 of 215) of AVS procedures in the developmental combined cohorts. Notably, the accuracy of the flowchart for patient management at external validation remained high (94.9%), with a reduction of 66.1% (78 of 118) of AVS procedures (supplemental Figure 1) (24).

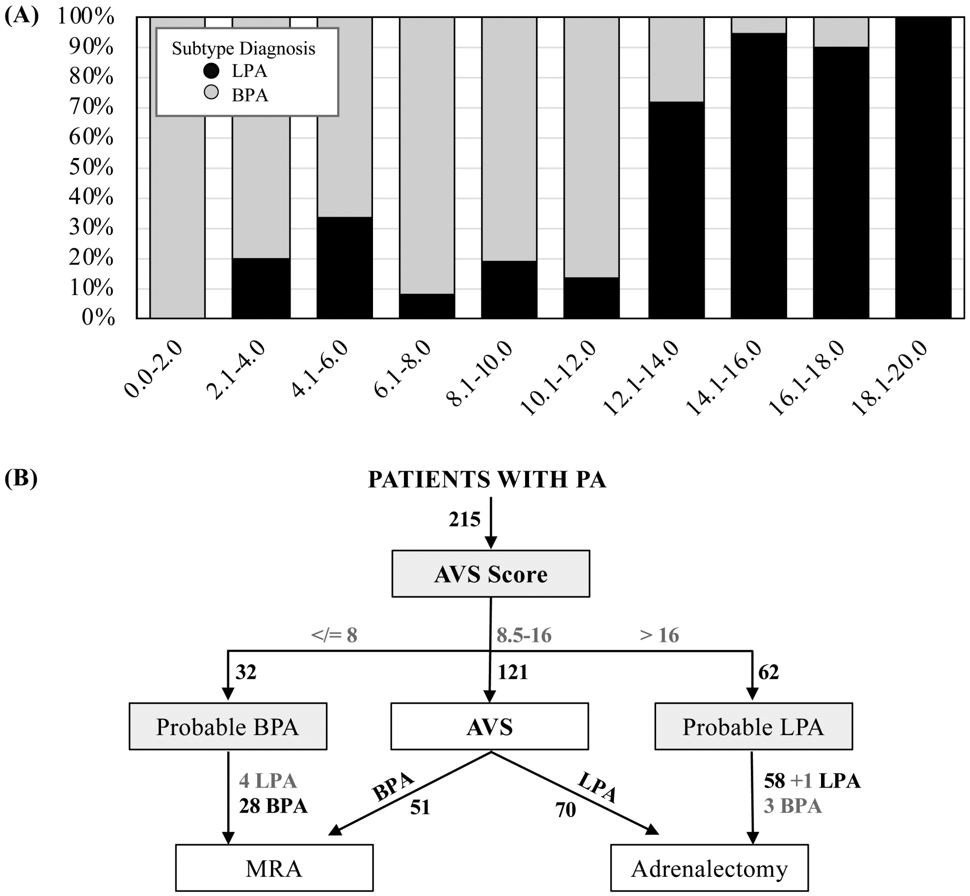
Discussion

In our study, we developed and validated 2 different prediction models based on supervised machine learning algorithms and a clinical score for the subtype diagnosis of PA. An online tool was developed to allow the application of the RF algorithm to clinical practice. Moreover, we proposed a flowchart for patient management that integrates our score system in a second user-friendly downloadable tool.

Küpers et al. proposed for the first time a clinical score to diagnose lateralized PA; the major advantages were easy applicability and a very high positive predictive value, resulting in the correct classification of all patients predicted as LPA (17). However, this score displays very low sensitivity, misclassifying 43% of LPA patients, who would miss the chance of potentially curative adrenalectomy. In addition, validation on independent cohorts did not confirm its diagnostic performance with a low accuracy, between 56.0% and 72.7% (26-29). Six other score systems were proposed (supplemental Table 11) (24). Two of them (18, 23) used only biochemical or demographic features, thus applicable before imaging. However, these scores were useful only to detect patients that could avoid AVS resulting from BPA. The other scores (19, 20) combined biochemical parameters with radiological findings and displayed a high negative predictive value (82.2% to 100%), with the identification of patients with BPA to be allocated to medical treatment. Limitations of these studies were the low number of enrolled patients, the absence of an internal or external validation, and the applicability only to patients undergoing captopril challenge (18, 20) or IV saline loading (19, 23) for confirmatory testing. The score proposed by Kamemura et al. was developed in more than 200 patients but was applicable only to patients without evidence of an adrenal mass at CT scanning, which represent a minority of patients with PA (21). Finally, Kobayashi et al. proposed and validated a score on more than 1000 patients, reporting a negative predictive value of 92.5% (22), but with insufficient accuracy. The application of this score in our patients resulted in an accuracy of 67.4% to 72.7%.

In our diagnostic models, the highest performance was reached by the RF algorithm, which identified 132 of 133 patients with LPA and correctly classified 68 of 69 patients with BPA, resulting in a sensitivity of 99.2% and a negative predictive value of 98.5%. The model accuracy was 93.0% and 87.0% at training and internal validation, respectively. Our SPACE score displayed an equally high performance with an overall accuracy of 89.3% in the training cohort and 81.5% in the internal validation cohort (using 12 as cutoff), outperforming all previously proposed clinical scores. A cutoff of 8 maximized sensitivity and negative predictive value (97% and 87.5%, respectively, in the combined cohort), correctly identifying 28 of 32 patients with BPA, whereas a cutoff of 16 maximized specificity and positive predictive value (96.3% and 95.2%, respectively, in the combined cohort), correctly identifying 59 of 62 patients with LPA.

All previously proposed score systems were applied in our cohort, to assess their generalizability. The accuracy at validation was not suitable for clinical use, ranging between 67.3% and 78% and suggesting a moderate



(C)

Diagnosis		Predicted		Sensitivity (%)	97.0
		LPA	BPA	Specificity (%)	96.3
Real	LPA	128 +1	4	PPV (%)	97.7
	BPA	3	79	NPV (%)	95.2

Figure 4. Score performance and management of PA patients. Flowchart for PA patient management using our prediction score. (A) Histogram showing the proportion of patients (y-axis, %) for each subtype diagnosis (LPA, lateralized PA, black; BPA, bilateral PA, gray), stratified by score points (x-axis) on the combined cohort (N = 215). The total number of patients (N) for each AVS score level and their proportion (%) are reported in supplemental Table 7 (24). (B) PA patient management using our score; the number of patients is indicated in bold; cutoffs and misclassified patients are indicated in gray. (C) Confusion matrix representing real and predicted subtype diagnosis, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). AVS, adrenal venous sampling; MRA, mineral receptor antagonist; PA, primary aldosteronism.

overfitting bias (up to 19%). Conversely, the overfitting effect was low in our models (from 2.3% to 7.8%) with a high accuracy at validation (from 81.5% to 93%).

To exclude selection bias and further assess the generalizability of our diagnostic models, we performed an external validation on an independent cohort of patients. LDA, RF model, and the SPACE score confirmed a high diagnostic performance (accuracy range 78.8%-94.4%), with a minimum overfitting bias.

We combined 3 biochemical variables with 3 imaging-related parameters associated with subtype diagnosis. These parameters were selected considering the results of univariate and multivariate regression analysis, and then used for the LDA, the RF model, and the score system. Potassium levels and aldosterone levels at screening and after confirmatory test are clinical criteria associated with a high probability of LPA and reflect the severity of disease (30). Imaging-related parameters resulted to be crucial for subtype diagnosis; in our cohort, only 5 of 133 LPA patients (3.8%) displayed a bilaterally normal CT scanning, whereas 85.7% had a defined nodule.

The SPACE score was integrated in a flowchart for the management of patients with PA, resulting in the correct classification of 96.3% of patients, potentially reducing almost half of the AVS. The lower cutoff identifies patients with BPA to address to MRA treatment: 28/32 patients with BPA were correctly classified, whereas 4 patients with LPA would be diagnosed as BPA and treated with MRA, therefore missing the chance of treatment by adrenalectomy. These 4 patients displayed bilaterally normal adrenals at CT scanning and are thus at high risk of partial/absent clinical success after surgery according with the recently proposed prognostic Primary Aldosteronism Surgery Outcome score (7, 25). The higher cutoff identifies patients with LPA, who could undergo unilateral adrenalectomy in centers where AVS is not available. With this strategy, 58 patients with LPA would be correctly adrenalectomized, whereas 4 patients would receive inappropriate surgery (3 patients with BPA and 1 patient with lateralization on the other side). The 3 BPA patients would also be misclassified by all other previously published scores.

The external validation resulted in similar performance, with correct management of 94.9% of patients and a potential reduction of 66.1% in the number of AVS, thus excluding a significant inter-center variability. The assessment of clinical and biochemical outcomes of patients with a correct prediction of LPA compared with those misclassified by the SPACE score, reinforced our findings. Finally, unlike previous models, our score system was applicable both to patients with PA diagnosed by saline infusion testing and by captopril challenge testing, with a similar accuracy (84.0% vs. 84.6%, respectively).

The present score is expected to be of interest to hypertension and endocrine centers and in particular for those that perform systematic screening of patients with hypertension, therefore having a high rate of diagnosis of BPA (31). With our score, a high proportion of BPA patients can avoid unnecessary AVS with a significant reduction of costs and potential complications.

The failure to define with certainty the side of aldosterone hypersecretion represents the main limit of our score and of all others previously proposed. A second limit is the retrospective inclusion of the patients with PA: a prospective validation in a large number of patients is warranted to confirm and further validate our prediction models. Moreover, our score cannot be applied to patients with PA diagnosed by the furosemide upright posture test or the oral saline loading test. Finally, dichotomization into LPA and BPA reflects the need to address patients to surgical vs. medical treatment and does not represent the complexity of the disease. Many patients with unilateral disease are cases with bilateral but asymmetrical aldosterone production displaying a high LI at AVS. These cases should benefit from adrenalectomy and are therefore considered as patients with unilateral or lateralized PA (32). The strengths of our study include the reliable accuracy of our diagnostic models after internal and external validation, using both machine learning algorithms, or a simple scoring system, with a potential impact on clinical practice for centers where AVS is not available. In addition, we proposed 2 user-friendly downloadable tools that integrate the RF model and the flowchart based on the SPACE score, allowing their application for the management of PA patients.

Conclusions

We developed and validated 2 prediction model and an easy applicable scoring system for the subtype diagnosis of PA. Our findings support the integration of clinical, biochemical, and imaging parameters by advanced computational approaches, to define PA subtype diagnosis, potentially reducing the number of AVS for patients with confirmed PA and guiding surgical decision in centers where AVS is not available.

Acknowledgments

Financial Support: This research did not receive any specific grant. C.A. and M.R. are supported by the Else Kröner-Fresenius Stiftung (2019_A104) and by the Deutsche Forschungsgemeinschaft within the CRC/Transregio 205/1 "The Adrenal: Central Relay in Health and Disease."

Additional Information

Correspondence and Reprint Requests: Paolo Mulatero, Division of Internal Medicine and Hypertension Unit, Department of Medical Sciences, University of Torino, Città della Salute e della Scienza, Via Genova 3, 10126 Torino, Italy. E-mail: paolo.mulatero@unito.it.

Disclosure Summary: The authors have nothing to disclose.

Data Availability: The datasets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

References

1. Käyser SC, Dekkers T, Groenewoud HJ, et al. Study Heterogeneity and estimation of prevalence of primary aldosteronism: a systematic review and meta-regression analysis. *J Clin Endocrinol Metab*. 2016;**101**(7):2826-2835.
2. Buffolo F, Monticone S, Burrello J, et al. Is primary aldosteronism still largely unrecognized? *Horm Metab Res*. 2017;**49**(12):908-914.
3. Monticone S, Burrello J, Tizzani D, et al. Prevalence and clinical manifestations of primary aldosteronism encountered in primary care practice. *J Am Coll Cardiol*. 2017;**69**(14):1811-1820.

4. Mulatero P, Monticone S, Bertello C, et al. Long-term cardio- and cerebrovascular events in patients with primary aldosteronism. *J Clin Endocrinol Metab.* 2013;**98**(12):4826-4833.
5. Monticone S, D'Ascenzo F, Moretti C, et al. Cardiovascular events and target organ damage in primary aldosteronism compared with essential hypertension: a systematic review and meta-analysis. *Lancet Diabetes Endocrinol.* 2018;**6**(1):41-50.
6. Funder JW, Carey RM, Mantero F, et al. The management of primary aldosteronism: case detection, diagnosis, and treatment: an Endocrine Society Clinical Practice Guideline. *J Clin Endocrinol Metab.* 2016;**101**(5):1889-1916.
7. Williams TA, Lenders JWM, Mulatero P, et al.; Primary Aldosteronism Surgery Outcome (PASO) investigators. Outcomes after adrenalectomy for unilateral primary aldosteronism: an international consensus on outcome measures and analysis of remission rates in an international cohort. *Lancet Diabetes Endocrinol.* 2017;**5**(9):689-699.
8. Hundemer GL, Curhan GC, Yozamp N, Wang M, Vaidya A. Cardiometabolic outcomes and mortality in medically treated primary aldosteronism: a retrospective cohort study. *Lancet Diabetes Endocrinol.* 2018;**6**(1):51-59.
9. Burton TJ, Mackenzie IS, Balan K, et al. Evaluation of the sensitivity and specificity of (11)C-metomidate positron emission tomography (PET)-CT for lateralizing aldosterone secretion by Conn's adenomas. *J Clin Endocrinol Metab.* 2012;**97**(1):100-109.
10. Heinze B, Fuss CT, Mulatero P, et al. Targeting CXCR4 (CXC chemokine receptor type 4) for molecular imaging of aldosterone-producing adenoma. *Hypertension.* 2018;**71**(2):317-325.
11. Mulatero P, di Cella SM, Monticone S, et al. 18-hydroxycorticosterone, 18-hydroxycortisol, and 18-oxocortisol in the diagnosis of primary aldosteronism and its subtypes. *J Clin Endocrinol Metab.* 2012;**97**(3):881-889.
12. Williams TA, Peitzsch M, Dietz AS, et al. Genotype-specific steroid profiles associated with aldosterone-producing adenomas. *Hypertension.* 2016;**67**(1):139-145.
13. Yang Y, Burrello J, Burrello A, et al. Classification of microadenomas in patients with primary aldosteronism by steroid profiling. *J Steroid Biochem Mol Biol.* 2019;**189**:274-282.
14. Monticone S, Viola A, Rossato D, et al. Adrenal vein sampling in primary aldosteronism: towards a standardised protocol. *Lancet Diabetes Endocrinol.* 2015;**3**(4):296-303.
15. Kempers MJ, Lenders JW, van Outheusden L, et al. Systematic review: diagnostic procedures to differentiate unilateral from bilateral adrenal abnormality in primary aldosteronism. *Ann Intern Med.* 2009;**151**(5):329-337.
16. Williams TA, Burrello J, Sechi LA, et al. Computed tomography and adrenal venous sampling in the diagnosis of unilateral primary aldosteronism. *Hypertension.* 2018;**72**(3):641-649.
17. Küpers EM, Amar L, Raynaud A, Plouin PF, Steichen O. A clinical prediction score to diagnose unilateral primary aldosteronism. *J Clin Endocrinol Metab.* 2012;**97**(10):3530-3537.
18. Nanba K, Tsuiki M, Nakao K, et al. A subtype prediction score for primary aldosteronism. *J Hum Hypertens.* 2014;**28**(12):716-720.
19. Kocjan T, Janez A, Stankovic M, Vidmar G, Jensterle M. A new clinical prediction criterion accurately determines a subset of patients with bilateral primary aldosteronism before adrenal venous sampling. *Endocr Pract.* 2016;**22**(5):587-594.
20. Kobayashi H, Haketa A, Ueno T, et al. Scoring system for the diagnosis of bilateral primary aldosteronism in the outpatient setting before adrenal venous sampling. *Clin Endocrinol (Oxf).* 2017;**86**(4):467-472.
21. Kamemura K, Wada N, Ichijo T, et al. Significance of adrenal computed tomography in predicting laterality and indicating adrenal vein sampling in primary aldosteronism. *J Hum Hypertens.* 2017;**31**(3):195-199.
22. Kobayashi H, Abe M, Soma M, et al.; JPAS Study Group. Development and validation of subtype prediction scores for the workup of primary aldosteronism. *J Hypertens.* 2018;**36**(11):2269-2276.
23. Leung HT, Woo YC, Fong CHY, et al. A clinical prediction score using age at diagnosis and saline infusion test parameters can predict aldosterone-producing adenoma from idiopathic adrenal hyperplasia. *J Endocrinol Invest.* 2020;**43**(3):347-355.
24. Burrello J, Burrello A, Pieroni J, et al. Data from: Development and validation of prediction models for subtype diagnosis of patients with primary aldosteronism. *J Clin Endocrinol Metab* 2020. Deposited 6 May 2020. https://github.com/ABurrello/SPACE-score
25. Burrello J, Burrello A, Stowasser M, et al. The primary aldosteronism surgical outcome score for the prediction of clinical outcomes after adrenalectomy for unilateral primary aldosteronism [Published online ahead of print January 18, 2019]. *Ann Surg.* 2019. doi: 10.1097/SLA.0000000000003200.
26. Sze WC, Soh LM, Lau JH, et al. Diagnosing unilateral primary aldosteronism - comparison of a clinical prediction score, computed tomography and adrenal venous sampling. *Clin Endocrinol (Oxf).* 2014;**81**(1):25-30.
27. Riester A, Fischer E, Degenhart C, et al. Age below 40 or a recently proposed clinical prediction score cannot bypass adrenal venous sampling in primary aldosteronism. *J Clin Endocrinol Metab.* 2014;**99**(6):E1035-E1039.
28. Venos ES, So B, Dias VC, Harvey A, Pasieka JL, Kline GA. A clinical prediction score for diagnosing unilateral primary aldosteronism may not be generalizable. *BMC Endocr Disord.* 2014;**14**:94.
29. Zhang Y, Niu W, Zheng F, et al. Identifying unilateral disease in Chinese patients with primary aldosteronism by using a modified prediction score. *J Hypertens.* 2017;**35**(12):2486-2492.
30. Young WF Jr, Klee GG. Primary aldosteronism. Diagnostic evaluation. *Endocrinol Metab Clin North Am.* 1988;**17**(2):367-395.
31. Mulatero P, Stowasser M, Loh KC, et al. Increased diagnosis of primary aldosteronism, including surgically correctable forms, in centers from five continents. *J Clin Endocrinol Metab.* 2004;**89**(3):1045-1050.
32. Vaidya A, Mulatero P, Baudrand R, Adler GK. The expanding spectrum of primary aldosteronism: implications for diagnosis, pathogenesis, and treatment. *Endocr Rev.* 2018;**39**(6):1057-1088.