












Harnessing feature extraction capacities from a pre-trained convolutional neural network (VGG-16) for the unsupervised distinction of aortic outflow velocity profiles in patients with severe aortic stenosis

Mark Lachmann ^{1†}, Elena Rippen ^{1†}, Daniel Rueckert ^{2,3}, Tibor Schuster⁴, Erion Xhepa ^{5,6}, Moritz von Scheidt ^{5,6}, Costanza Pellegrini⁵, Teresa Trenkwalder ⁵, Tobias Rheude⁵, Anja Stundl ¹, Ruth Thalmann¹, Gerhard Harmsen⁷, Shinsuke Yuasa⁸, Heribert Schunkert ^{5,6}, Adnan Kastrati ^{5,6}, Michael Joner^{5,6}, Christian Kupatt ^{1,6}, and Karl-Ludwig Laugwitz ^{1,6*}

¹First Department of Medicine, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Straße 22, 81675 Munich, Germany; ²Institute for AI and Informatics in Medicine, Faculty of Informatics and Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany; ³Department of Computing, Imperial College London, London, UK; ⁴Department of Family Medicine, McGill University, Montreal, Quebec, Canada; ⁵Department of Cardiology, German Heart Centre Munich, Technical University of Munich, Munich, Germany; ⁶DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany; ⁷Department of Physics, University of Johannesburg, Auckland Park, South Africa; and ⁸Department of Cardiology, Keio University School of Medicine, Minato, Tokyo, Japan

Received 31 August 2021; revised 14 October 2021; accepted 1 February 2022; online publish-ahead-of-print 22 April 2022

Aims

Hypothesizing that aortic outflow velocity profiles contain more valuable information about aortic valve obstruction and left ventricular contractility than can be captured by the human eye, features of the complex geometry of Doppler tracings from patients with severe aortic stenosis (AS) were extracted by a convolutional neural network (CNN).

Methods and results

After pre-training a CNN (VGG-16) on a large data set (ImageNet data set; 14 million images belonging to 1000 classes), the convolutional part was employed to transform Doppler tracings to 1D arrays. Among 366 eligible patients [age: 79.8 ± 6.77 years; 146 (39.9%) women] with pre-procedural echocardiography and right heart catheterization prior to transcatheter aortic valve replacement (TAVR), good quality Doppler tracings from 101 patients were analysed. The convolutional part of the pre-trained VGG-16 model in conjunction with principal component analysis and k-means clustering distinguished two shapes of aortic outflow velocity profiles. Kaplan–Meier analysis revealed that mortality in patients from Cluster 2 ($n = 40$, 39.6%) was significantly increased [hazard ratio (HR) for 2-year mortality: 3; 95% confidence interval (CI): 1–8.9]. Apart from reduced cardiac output and mean aortic valve gradient, patients from Cluster 2 were also characterized by signs of pulmonary hypertension, impaired right ventricular function, and right atrial enlargement. After training an extreme gradient boosting algorithm on these 101 patients, validation on the remaining 265 patients confirmed that patients assigned to Cluster 2 show increased mortality (HR for 2-year mortality: 2.6; 95% CI: 1.4–5.1, P -value: 0.004).

* Corresponding author. Tel: +49 89 4140 2350, Fax: +49 89 4140 4900, Email: KL.Laugwitz@mri.tum.de

[†]The first two authors contributed equally to the study.

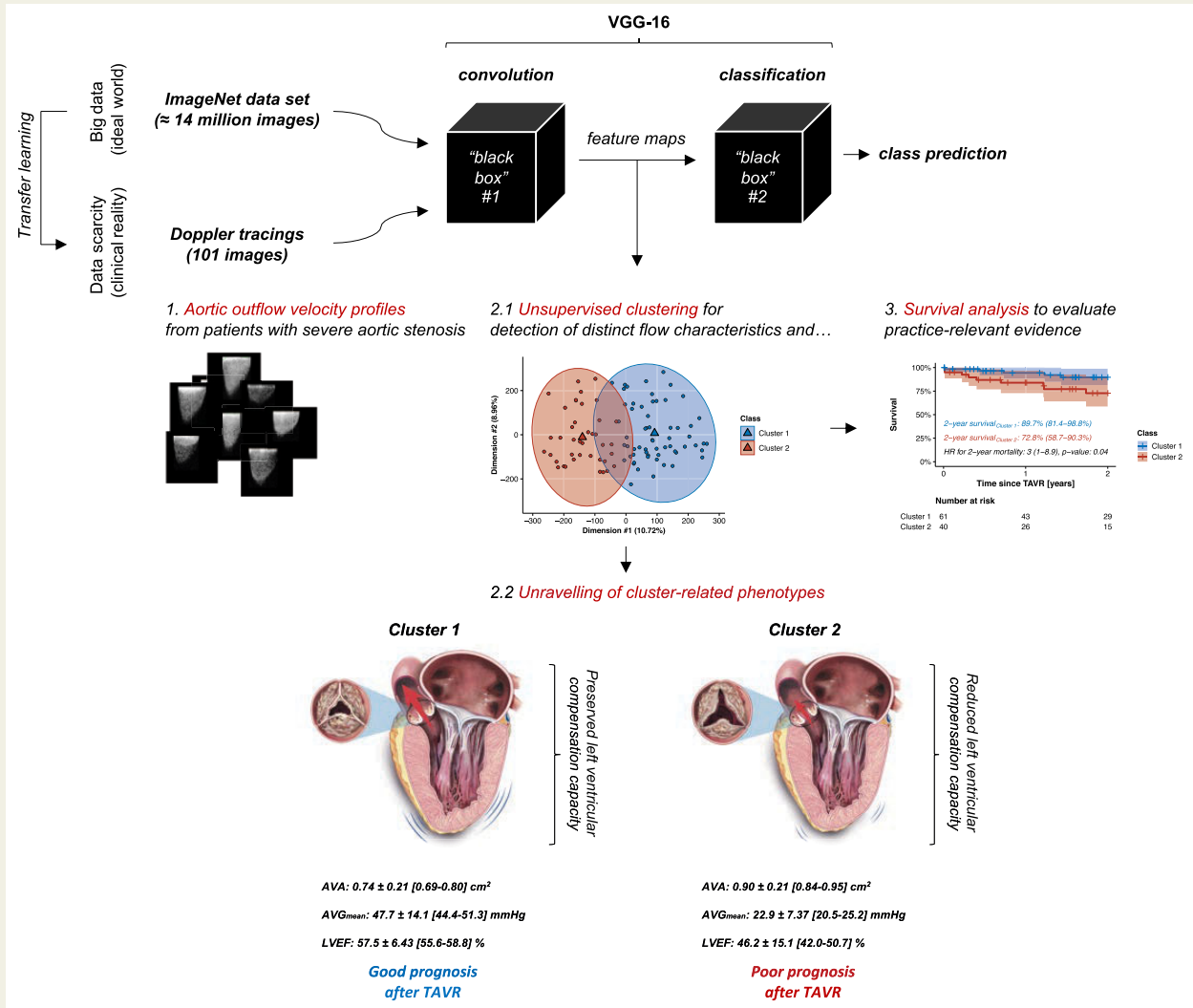
© The Author(s) 2022. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Conclusion

Transfer learning enables sophisticated pattern recognition even in clinical data sets of limited size. Importantly, it is the left ventricular compensation capacity in the face of increased afterload, and not so much the actual obstruction of the aortic valve, that determines fate after TAVR.

Graphical Abstract



Keywords

Severe aortic stenosis • Transcatheter aortic valve replacement • Aortic outflow velocity profile • Convolutional neural network • Transfer learning

Introduction

In the dawning age of artificial intelligence, machine learning technology is progressively implemented into medical research and clinical decision support.¹⁻⁶ Without the constraint of any *a priori*

assumption, machine learning algorithms iteratively learn from data, typically requiring a myriad of information. Bedside research embedded in clinical practice, however, is commonly based on modest-sized patient cohorts, e.g. in the setting of rare diseases or novel treatment strategies. Transfer learning holds the promise to

alleviate the bottleneck of insufficient training data by acquiring feature extraction capacity from a large data set and applying it to a related problem in the same domain.

Severe aortic stenosis (AS), which can trigger a deleterious cascade including left heart dysfunction, pulmonary hypertension (PH), and eventually right heart failure,⁷ and which is associated with a 2-year mortality of up to more than 50% unless valve replacement is performed promptly,⁸ is typically diagnosed by transthoracic echocardiography.⁹ Besides an increasing gradient due to progressive narrowing of the aortic valve, the aortic outflow velocity profile in patients with severe AS changes from a triangular shape with an early peak to a much more rounded form with a later peak.¹⁰ Furthermore, acceleration of flow velocity will eventually deteriorate due to left ventricular decompensation, resulting in a 'low flow, low gradient AS'.¹¹

Hypothesizing that the aortic outflow velocity profile contains more valuable information about aortic valve obstruction and left ventricular contractility than can be captured by human cognition, this study sought to extract features of the complex geometry of Doppler tracings by employing a convolutional neural network (CNN). VGG-16 is a CNN with state-of-the-art feature extraction capacity, which achieved 92.7% Top 5 test accuracy in a data set of over 14 million regular natural images belonging to 1000 classes (ImageNet data set).¹² Adopting the concept of transfer learning, the convolutional part of the pre-trained VGG-16 model was employed to transform Doppler tracings from a small, but well-characterized cohort of patients with severe AS to 1D arrays. After principal component analysis (PCA) and k-means clustering of those 1D arrays, practice-relevant evidence was assessed by relating cluster assignment with all-cause 2-year mortality after transcatheter aortic valve replacement (TAVR).

Methods

Patient recruitment

This is a *post hoc* analysis of prospectively and systematically collected data from patients undergoing TAVR for severe AS at two tertiary care centres in Munich, Germany, between January 2014 and December 2020. The study was approved by the respective local ethics committees in conformity with the Declaration of Helsinki, and all patients enrolled provided written informed consent. In total, the joint registry listed 2575 patients. Among 366 completely characterized patients with pre-procedural echocardiography and right heart catheterization prior to TAVR, good quality Doppler tracings with sharp, well-defined borders from the intentionally small number of 101 patients were analysed, constituting the derivation cohort (Figure 1A). The validation cohort was consequently represented by the remaining 265 patients with complete data from pre-procedural echocardiography and right heart catheterization, yet without good quality Doppler tracings (or no available records at all). As an elderly patient population approaching the end of life was studied, post-procedural 2-year all-cause mortality was defined as a clinically meaningful primary outcome measure. Survival data were obtained from the German Civil Registry in case of patients being registered in Germany ($n = 354$; 96.7%), or from general practitioners, hospitals, and practice cardiologists for patients from foreign countries.

Transthoracic echocardiography

All echocardiographic studies were performed by experienced institutional cardiologists during clinical routine using a commercially available echocardiography system equipped with a 2.5-MHz multifrequency phased-array transducer. The continuous wave Doppler-derived aortic outflow velocity profiles were obtained from the apical four-chamber view (Figure 2A). Only 101 aortic outflow velocity profiles were selected for clustering depending on image quality—meaning that Doppler tracings with insufficient contrast or with labelling within the aortic outflow velocity profile were excluded (Figure 2B).

Statistical analysis

All statistical analyses were performed using R version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria; see [Supplementary material online, Table S1](#) for a complete list of employed R packages). The pre-trained VGG-16 model on the ImageNet data set was loaded from the Keras deep learning library (R package 'keras'), and the classifier part of the VGG-16 model was omitted. After pre-processing (Figure 2A), scaled Doppler tracings as input images were thus converted to a feature tensor of shape as the output of the last layer of the convolutional part (R packages 'magick' and 'imager'). After transformation from feature tensor of shape (7, 7, 512) to 1D array with $7 \times 7 \times 512$ values per instance, PCA and k-means clustering were applied (R packages 'FactoMineR', 'factoextra', and 'NbClust'). Notably, we expected two clusters to be segregated. Survival was illustrated using Kaplan–Meier method, and a Cox proportional hazards model was used to estimate hazard ratios (HRs) between identified clusters (R packages 'survival', 'survminer', and 'ggforest'). Missing values among variables that were identified as significant predictors for 2-year mortality in initial univariate analysis were imputed by a random forest algorithm (R package 'missForest')¹³ before proceeding with multivariate analysis, but were not used hereinafter, e.g. for cluster comparisons. Because the derivation cohort was unbalanced with regards to cluster assignments, a technique to synthetically over-sample the minority class was applied (synthetic minority over-sampling technique; SMOTE) (R package 'DMwR').¹⁴ After balancing, an extreme gradient boosting algorithm (R package 'xgboost') was selected as the machine learning technique of choice for cluster assignment in future patients, and it was trained on a comprehensive set of functional and structural parameters from echocardiography and right heart catheterization. Again, missing values were imputed by a random forest algorithm. SHAP (SHapley Additive exPlanations) values were calculated to compare the contribution of input variables to the model prediction (R package 'SHAPforxgboost').¹⁵ To sum up, this study was designed as a two-step experiment:

- (1) In a first step, we aimed to decipher meaningful echocardiographic signatures and related cardiac phenotypes by analysing aortic outflow velocity profiles (Doppler tracings) from 101 patients with severe AS undergoing TAVR by using a pre-trained CNN in conjunction with PCA and k-means clustering (unsupervised machine learning experiment).
- (2) Since the first experiment did not allow to assign future patients to the just defined clusters, we additionally sought to employ an extreme gradient boosting algorithm, which was trained on functional and structural parameters of cardiopulmonary conditions from the 101 patients with good quality Doppler tracings (hereinafter referred to as derivation cohort) to predict cluster assignments as stemming from the first experiment, and which was then validated on the remaining 265 patients (hereinafter referred to as validation

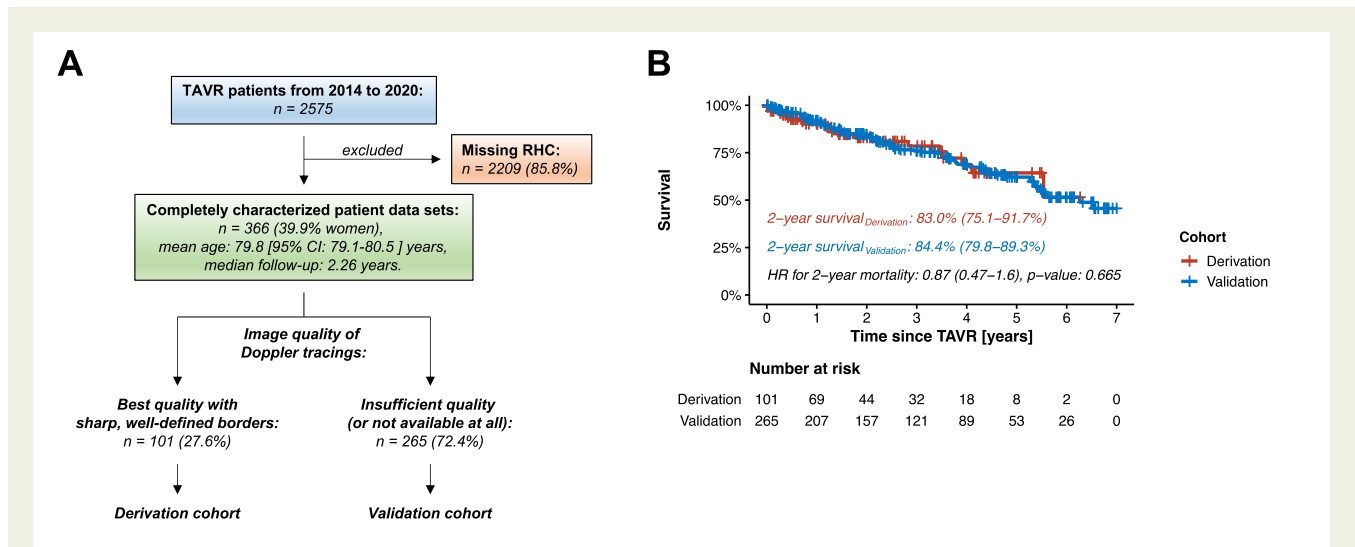


Figure 1 General information about the study population from recruitment to follow-up. (A) A flowchart for patient recruitment in order to select 101 patients with best quality Doppler tracings. Notably, 56 out of 366 patients (15.3%) had no Doppler tracings as raw data available. (B) Kaplan–Meier survival plot testing for differences in survival between derivation and validation cohorts. RHC, right heart catheterization; TAVR, transcatheter aortic valve replacement.

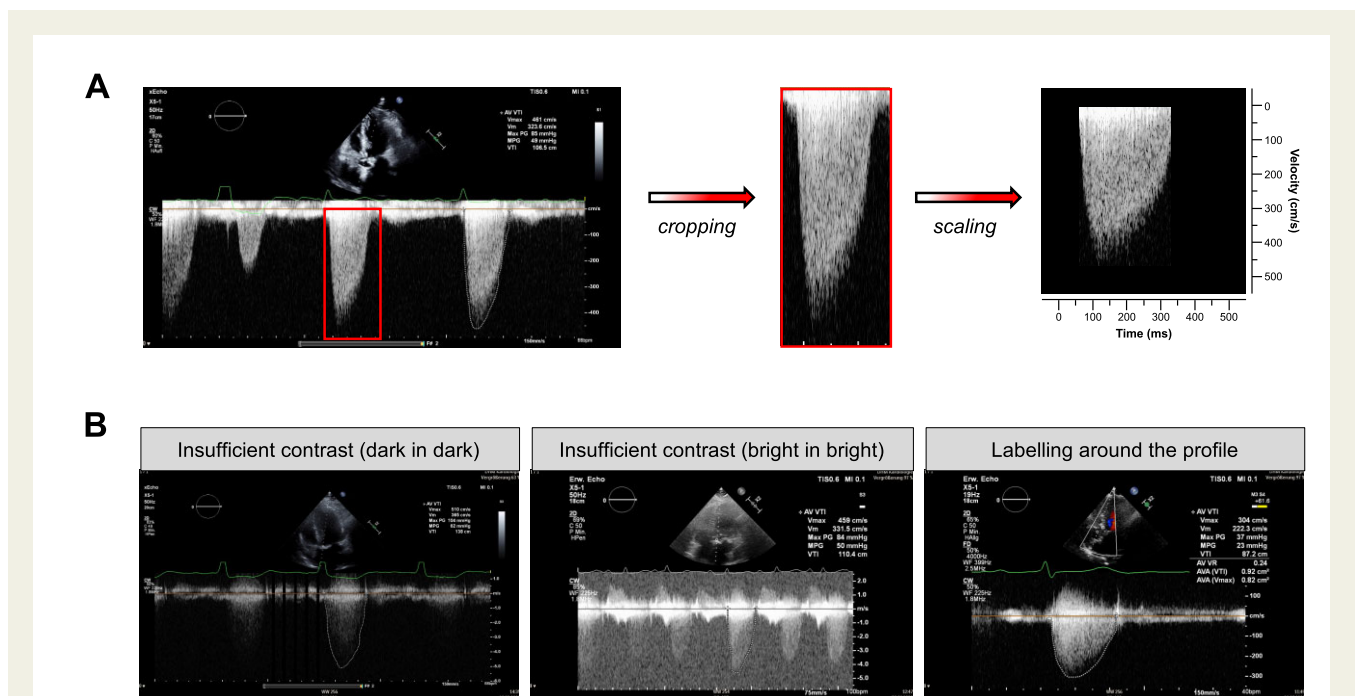


Figure 2 Pre-processing of aortic outflow velocity profiles from patients with severe AS data input for the convolutional neural network. (A) Schematic of the image pre-processing pipeline. One representative aortic outflow velocity profile per patient was extracted from records. Cropping of the region of interest, i.e. during systole, was done manually. Since original echocardiographic images were recorded at different scales, but homogeneity of data input had to be provided, Doppler tracings were further manually scaled according to uniform time and velocity axes (see also [Supplementary material online, Figure S1](#) for a standard operation procedure explaining additional details to create the desired normalized profiles). Neither image normalization nor histogram equalization was applied during pre-processing. Re-sizing to 224×224 pixel format as the default input size of the VGG-16 model was already part of the image processing R code after loading the folder with 101 scaled Doppler tracings. (B) Representative Doppler tracings that were excluded due to insufficient contrast, or due to labelling within the aortic outflow velocity profile.

cohort) with regards to cluster-related survival after TAVR (supervised machine learning experiment).

Categorical variables are presented as numbers and frequencies (%), whilst continuous variables are given as mean \pm standard deviation (SD) and 95% confidence interval (CI). Chi-square or Fisher's exact test was used to evaluate the association between categorical variables, and independent-samples Wilcoxon test was used for comparison of continuous variables. For analysis of collinearity, Pearson's correlation coefficients were calculated. A P -value ≤ 0.05 was considered to indicate statistical significance.

Results

One hundred and one patients with good quality Doppler tracings illustrate the problem of data scarcity in clinical research

Importantly, aortic outflow velocity profiles from only 101 out of 366 patients were initially analysed (Figure 1A) to emphasize the problem of data scarcity in clinical research. Therefore, to confirm the representative nature of the small-sized derivation cohort at hand [mean age: 79.3 ± 6.78 ; 95% CI: 78.0–80.7 years; 49 (48.5%) women], derivation and validation cohorts were initially compared with regards to demographic, clinical, echocardiographic, and haemodynamic characteristics (Supplementary material online, Tables S2 and S3). No significant differences were found with regards to age, symptomatic burden expressed as New York Heart Association (NYHA) functional class, obstruction of the aortic valve expressed as aortic valve area (AVA), left ventricular systolic function, mean pulmonary artery pressure (mPAP), and right ventricular dysfunction. In fact, a difference was detected in the proportion of female patients, which were significantly more often represented in the derivation cohort (48.5% vs.

36.6%, P -value: 0.0499). Presenting with a mean AVA of 0.804 ± 0.223 (95% CI: 0.760–0.848) cm^2 , and predominantly suffering from dyspnoea corresponding to NYHA functional Class III (56.4%) (Tables 1 and 2) 2-year survival after TAVR among patients from the derivation cohort ranged at 83.0% (95% CI: 75.1–91.7), which was statistically indifferent compared to patients from the validation cohort (P -value: 0.665) (Figure 1B).

Two distinct clusters of aortic outflow velocity profiles can be distinguished, reflecting different phenotypes with subsequently differing mortality

The convolutional part of the pre-trained VGG-16 model (Figure 3A) in conjunction with PCA and k-means clustering of the abstractions of Doppler tracings enabled to distinguish two shapes of aortic outflow velocity profiles (Figure 3B). Interestingly, all patients from Cluster 2 presented with a mean aortic valve gradient (AVG_{mean}) below 40 mmHg, whilst AVG_{mean} from patients in Cluster 1 ranged between 20 and 102 mmHg (Figure 3C). Kaplan–Meier analysis revealed that mortality in patients from Cluster 2 ($n = 40$, 39.6%) was significantly increased (HR for 2-year mortality: 3; 95% CI: 1–8.9) (Figure 3D). Besides reduced cardiac output (4.57 ± 1.42 ; 95% CI: 4.17–5.04 L/min) and signs of PH (mPAP: 31.9 ± 12.2 ; 95% CI: 28.5–35.7 mmHg), patients from Cluster 2 also presented with more severe impairment of right ventricular function [tricuspid annular plane systolic excursion (TAPSE): 18.1 ± 3.82 ; 95% CI: 17.3–19.2 mm] and right atrial enlargement [right atrial (RA) area: 22.0 ± 8.28 ; 95% CI: 19.5–24.6 cm^2] in comparison to patients from Cluster 1 (Figure 3E and Table 2). Contrarily to the initial expectation, patients from Cluster 1 with seemingly less extensive cardiac damage were diagnosed with a more severe obstruction of the aortic valve than patients from Cluster 2 (AVA: 0.739 ± 0.211 ; 95% CI: 0.685–0.793 cm^2 vs. 0.903 ± 0.205 ; 95% CI: 0.837–0.968 cm^2 , P -value: 0.0001).

Table 1 Demographic and clinical characteristics in accordance with cluster assignment (derivation cohort)

	Class			P-value
	All ($n = 101$)	Cluster 1 ($n = 61$)	Cluster 2 ($n = 40$)	
Age (years), mean \pm SD [95% CI]	79.3 ± 6.78 [78.0–80.7]	79.4 ± 5.88 [77.8–80.8]	79.3 ± 8.03 [76.8–81.6]	0.4067
Women, N (%)	49 (48.5%)	31 (50.8%)	18 (45.0%)	0.7123
BMI (kg/m^2), mean \pm SD [95% CI]	26.8 ± 4.28 [26.0–27.6]	26.9 ± 4.25 [25.8–28.0]	26.7 ± 4.36 [25.5–28.0]	0.8758
Arterial hypertension, N (%)	88 (87.1%)	51 (83.6%)	37 (92.5%)	0.3166
Diabetes mellitus, N (%)	23 (22.8%)	14 (23.0%)	9 (22.5%)	1
NYHA functional class, mean \pm SD [95% CI]	2.61 ± 0.71 [2.47–2.75]	2.54 ± 0.72 [2.36–2.72]	2.72 ± 0.68 [2.53–2.93]	0.1896
NYHA functional Class III	57 (56.4%)	32 (52.5%)	25 (62.5%)	0.4294
NYHA functional Class IV	6 (5.9%)	3 (4.9%)	3 (7.5%)	0.9152
EuroSCORE (%), mean \pm SD [95% CI]	17.1 ± 14.3 [14.3–20.0]	13.9 ± 8.94 [11.7–16.3]	22.1 ± 19.0 [16.5–27.6]	0.0694
eGFR (mL/min), mean \pm SD [95% CI]	60.7 ± 21.4 [56.4–64.9]	65.4 ± 19.5 [60.1–70.2]	53.3 ± 22.3 [46.8–59.3]	0.0214
CAD, N (%)	85 (84.1%)	49 (80.3%)	36 (90.0%)	0.3061
COPD, N (%)	12 (11.9%)	7 (11.5%)	5 (12.5%)	1
Atrial fibrillation and/or flutter, N (%)	42 (41.6%)	19 (31.1%)	23 (57.5%)	0.0155

BMI, body mass index; CAD, coronary artery disease; CI, confidence interval; COPD, chronic obstructive pulmonary disease; eGFR, estimated glomerular filtration rate; NYHA, New York Heart Association; SD, standard deviation.

Table 2 Comparison of echocardiographic and haemodynamic characteristics in accordance with cluster assignment (derivation cohort)

	All (n = 101)	Class		P-value
		Cluster 1 (n = 61)	Cluster 2 (n = 40)	
AVA (cm ²), mean ± SD [95% CI]	0.804 ± 0.223 [0.760–0.848]	0.739 ± 0.211 [0.685–0.793]	0.903 ± 0.205 [0.837–0.968]	0.0001
AVG _{mean} (mmHg), mean ± SD [95% CI]	37.9 ± 17.0 [34.5–41.2]	47.7 ± 14.1 [44.4–51.3]	22.9 ± 7.37 [20.5–25.2]	4.6 × 10 ⁻¹⁵
Cardiac output (L/min), mean ± SD [95% CI]	5.08 ± 1.33 [4.82–5.34]	5.41 ± 1.17 [5.11–5.70]	4.57 ± 1.42 [4.17–5.04]	0.0006
LVEF (%), mean ± SD [95% CI]	53.0 ± 12.0 [50.6–55.4]	57.5 ± 6.43 [55.6–58.8]	46.2 ± 15.1 [42.0–50.7]	0.0001
LVEDD (mm), mean ± SD [95% CI]	47.2 ± 9.44 [45.4–49.1]	45.7 ± 7.70 [43.8–47.7]	49.7 ± 11.4 [46.2–52.8]	0.1028
mPCWP (mmHg), mean ± SD [95% CI]	17.8 ± 9.01 [16.1–19.6]	15.5 ± 8.40 [13.3–17.6]	21.4 ± 8.83 [18.6–24.2]	0.0007
mPAP (mmHg), mean ± SD [95% CI]	27.6 ± 11.5 [25.3–29.8]	24.7 ± 10.1 [22.1–27.2]	31.9 ± 12.2 [28.5–35.7]	0.0019
RV-RA gradient (mmHg), mean ± SD [95% CI]	34.0 ± 14.7 [30.9–37.1]	32.2 ± 13.8 [28.5–35.9]	37.0 ± 16.0 [31.3–42.6]	0.1302
PVR (WU), mean ± SD [95% CI]	2.10 ± 1.36 [1.83–2.37]	1.80 ± 0.997 [1.55–2.06]	2.56 ± 1.69 [2.02–3.10]	0.0050
TAPSE (mm), mean ± SD [95% CI]	19.8 ± 4.05 [18.9–20.6]	20.8 ± 3.89 [19.8–21.7]	18.1 ± 3.82 [17.3–19.2]	0.0014
Right midventricular diameter (mm), mean ± SD [95% CI]	29.0 ± 6.46 [27.6–30.3]	27.4 ± 5.82 [25.8–29.0]	31.1 ± 6.77 [28.8–33.3]	0.0088
LA area (cm ²), mean ± SD [95% CI]	25.8 ± 8.21 [24.2–27.5]	24.8 ± 8.11 [22.9–27.3]	27.4 ± 8.21 [24.8–30.1]	0.1017
RA area (cm ²), mean ± SD [95% CI]	19.5 ± 6.89 [18.2–20.9]	17.8 ± 5.17 [16.6–18.9]	22.0 ± 8.28 [19.5–24.6]	0.0133
Low gradient (AVG _{mean} < 40 mmHg), N (%)	54 (53.5%)	14 (23.0%)	40 (100%)	1.5 × 10 ⁻¹³
LV dysfunction (LVEF ≤ 45%), N (%)	22 (21.8%)	5 (8.2%)	17 (42.5%)	0.0001
PH (mPAP ≥ 25 mmHg), N (%)	52 (51.5%)	25 (41.0%)	27 (67.5%)	0.0162
RV dysfunction (TAPSE ≤ 16 mm), N (%)	20 (20.4%)	6 (9.8%)	14 (23.0%)	0.0021
MR ≥ III/IV°, N (%)	10 (9.90%)	4 (6.56%)	6 (15.0%)	0.1882
TR ≥ III/IV°, N (%)	7 (6.93%)	2 (3.28%)	5 (12.5%)	0.1101

AVA, aortic valve area; AVG_{mean}, mean aortic valve gradient; CI, confidence interval; LA area, left atrial area; LV dysfunction, left ventricular dysfunction; LVEDD, left ventricular end-diastolic diameter; LVEF, left ventricular ejection fraction; mPAP, mean pulmonary artery pressure; mPCWP, mean postcapillary wedge pressure; MR, mitral regurgitation; PH, pulmonary hypertension; PVR, pulmonary vascular resistance; RA area, right atrial area; RV dysfunction, right ventricular dysfunction; SD, standard deviation; TAPSE, tricuspid annular plane systolic excursion; TR, tricuspid regurgitation.

Comorbidities, such as arterial hypertension, coronary artery disease, and chronic obstructive pulmonary disease, were similarly prevalent between Clusters 1 and 2. Yet, patients from Cluster 2 with failing hearts showed a higher prevalence of atrial fibrillation and/or flutter (57.5% vs. 31.1%, *P*-value: 0.0155), and parallelly suffered from reduced renal function (estimated glomerular filtration rate: 53.3 ± 22.3; 95% CI: 46.8–59.3 mL/min vs. 65.4 ± 19.5; 95% CI: 60.1–70.2 mL/min, *P*-value: 0.0214). Notably, no general association between deteriorating cardiac output and worsening of renal function could be described by correlation analysis (*R*: 0.10, *P*-value: 0.3117),

nor did patients with reduced cardiac output generally display impairments of renal function (Supplementary material online, Figure S2A and B). An illustration of 20 exemplifying profiles per cluster is provided in Figure 3F.

Conventional dichotomization according to AVG_{mean} results in loss of prognostic resolution

In order to compare unsupervised clustering of aortic outflow velocity profiles with a traditional approach of hand-crafted

categorization, the derivation cohort was conventionally dichotomized according to AVG_{mean} (Figure 4A). Survival analysis confirmed that patients with $AVG_{mean} < 40$ mmHg ($n = 54$, 53.5%) died earlier, but no statistical significance was reached (HR for 2-year mortality: 1.8; 95% CI: 0.59–5.2) (Figure 4B). Apart from identifying well-established predictors for mortality, such as deteriorating renal function, male sex, and EuroSCORE, univariate Cox regression analysis also confirmed the prognostic value of left ventricular ejection fraction, mPAP, and TAPSE. At the same time, no significant association between AVG_{mean} or AVG_{max} , on the one hand, and 2-year all-cause mortality, on the other hand, could be detected by regression analysis (Table 3).

An extreme gradient boosting algorithm enables cluster assignment in future patients and confirms that left ventricular compensation capacity rather than the actual obstruction of the aortic valve determines fate after transcatheter aortic valve replacement

To test whether the cluster-related phenotypes as detected by the convolutional part of the pre-trained VGG-16 model in conjunction with PCA and k-means clustering could also be found among the remaining 265 patients with either poor quality or no available Doppler tracings [56 (15.3%) of 366 patients had no Doppler tracings as raw data available], an extreme gradient boosting algorithm was

trained on a comprehensive set of functional and structural parameters from pre-procedural echocardiography and right heart catheterization. In total, 12 variables, ideally covering all stages of cardiac and pulmonary circulatory conditions as previously described,¹⁶ served as input data. Moreover, the actual obstruction of the aortic valve expressed as AVA was included as a thirteenth input variable (Supplementary material online, Figure S3 for a complete list of input variables). Since the derivation cohort was predominantly composed of patients assigned to cluster 1 (60.4%), a minority class oversampling technique (SMOTE) was applied to create a balanced data set (Figure 5). After application of SMOTE, a training and a test set were randomly defined using a 0.75:0.25 split ratio, meaning that 120 ‘patients’ were assigned to the training set and 40 ‘patients’ were assigned to the test set. As a holdout data set, this test set was designated to finally assess the extreme gradient boosting algorithm’s performance, before eventually using the trained algorithm for patient-to-cluster assignment in the validation cohort. The purpose of the validation cohort was to evaluate cluster-related survival differences as they were observed for the clusters that have been segregated during the first, unsupervised machine learning experiment among the derivation cohort. In total, 2.44% of the 1313 data points related to 101 patients from the derivation cohort had missing values for those 13 variables (Supplementary material online, Figure S3A), and the largest proportion of missing values was found for measurements of right midventricular diameter (12.9% of values missing) (Supplementary material online, Figure 3B). After imputing missing values, initially observed and later imputed values for right

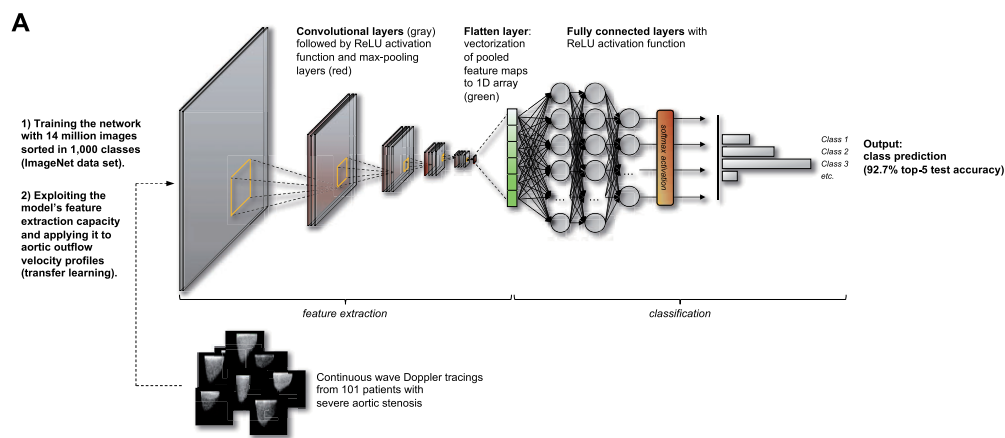


Figure 3 A convolutional neural network followed by PCA and unsupervised k-means clustering provides the proof-of-principle that two subgroups of patients with severe AS can be distinguished according to the aortic outflow velocity profile. (A) VGG-16 network architecture (schematic). The VGG-16 network can be split into two parts: 13 convolutional layers constitute the first part, through which each image is passed through for feature extraction. The convolutional layers are followed by three fully connected layers for classification, and the last layer uses a softmax activation function for final class prediction. Since the aortic outflow velocity profiles were no established class within the ImageNet data set, the classification part of VGG-16 was omitted after pre-training, and hence only the model’s feature extraction capacity was exploited in order to transform aortic outflow velocity profiles to 1D arrays (flatten layer), which were subsequently used for unsupervised clustering. (B) PCA of 1D arrays from 101 aortic outflow velocity profiles. (C) Scatter plot including 95% confidence ellipse in order to illustrate cardiac output and mean aortic valve gradient in accordance with cluster assignment. (D) Kaplan–Meier survival analysis in accordance with cluster assignment. (E) Bee swarm plots for comparison of baseline echocardiographic and haemodynamic data. (F) Representative aortic outflow velocity profiles in accordance with cluster assignment. AVA, aortic valve area; AVG_{mean} , mean aortic valve gradient; LA area, left atrial area; LVEDD, left ventricular end-diastolic diameter; mPAP, mean pulmonary artery pressure; RA area, right atrial area; ReLU, Rectified Linear Unit; TAPSE, tricuspid annular plane systolic excursion.

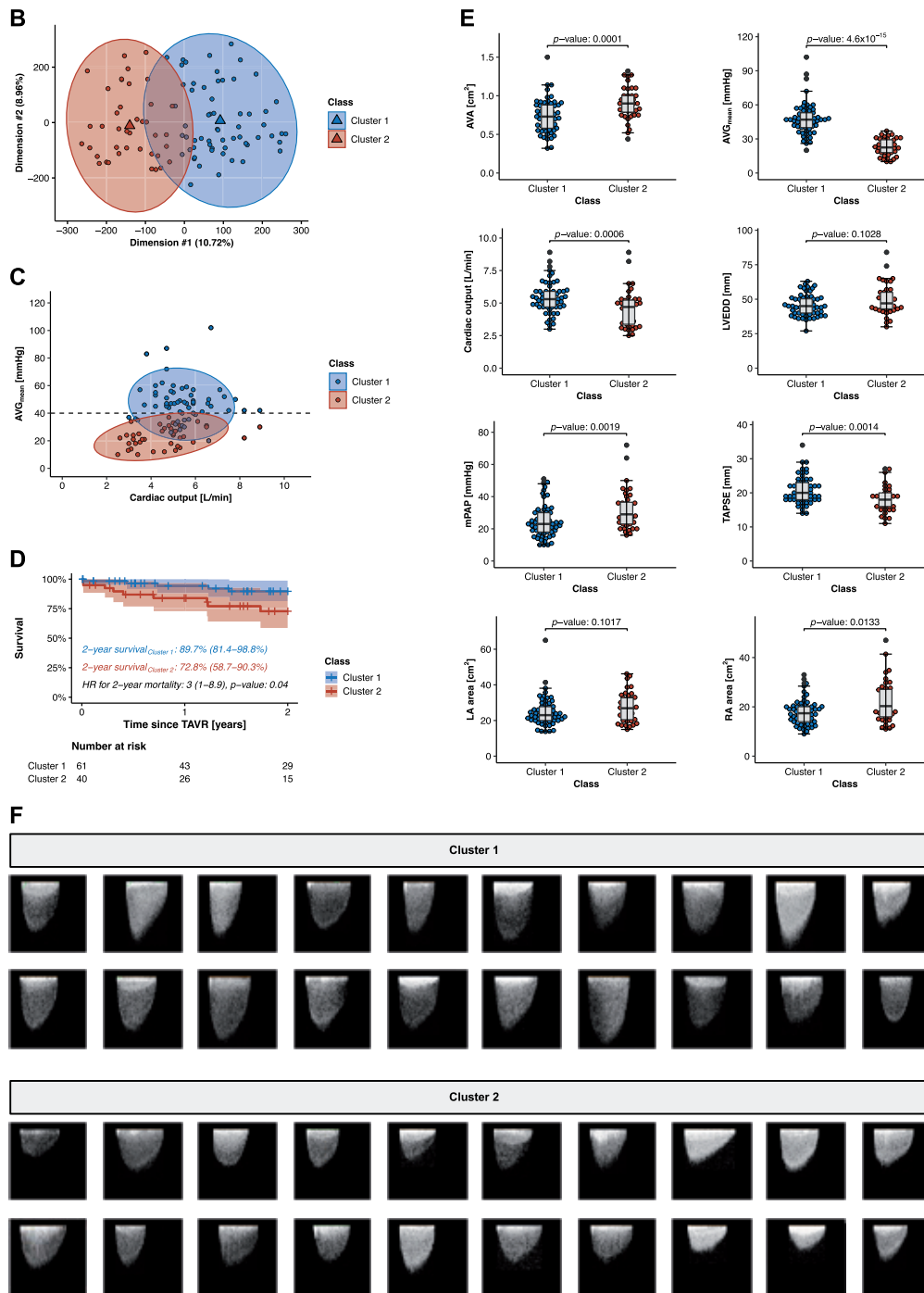


Figure 3 (Continued)

midventricular diameter displayed a similar distribution (29.0 ± 6.46 ; 95% CI: 27.6–30.3 mm vs. 27.6 ± 2.45 ; 95% CI: 26.1–29.1 mm, P -value: 0.5256) (Supplementary material online, Figure S3C and D). Importantly, the main characteristics of Clusters 1 and 2 in terms of cardiac output, AVG_{mean} and AWA were preserved after over-sampling (Figure 6A). An extreme gradient boosting algorithm for cluster assignment was hereinafter trained on 58 instances for ‘Cluster 1’ and on 62 instances ‘Cluster 2’, respectively, and it reached an accuracy of 97.5%, significantly outperforming the no information

rate (P -value: 1.4×10^{-9}), as demonstrated in the test set of 40 ‘patients’ (Figure 6B). Notably, AVG_{mean} showed by far the highest global feature importance for cluster prediction as determined by SHAP values (Figure 6C). Applying the trained extreme gradient boosting algorithm to the validation cohort of 265 patients (Figure 5) enabled identification of patients belonging to high-risk Cluster 2. Again, those patients were characterized by a functionally and structurally failing left heart in conjunction with PH and right heart impairment (Table 4). Compared to patients from Cluster 1, survival was

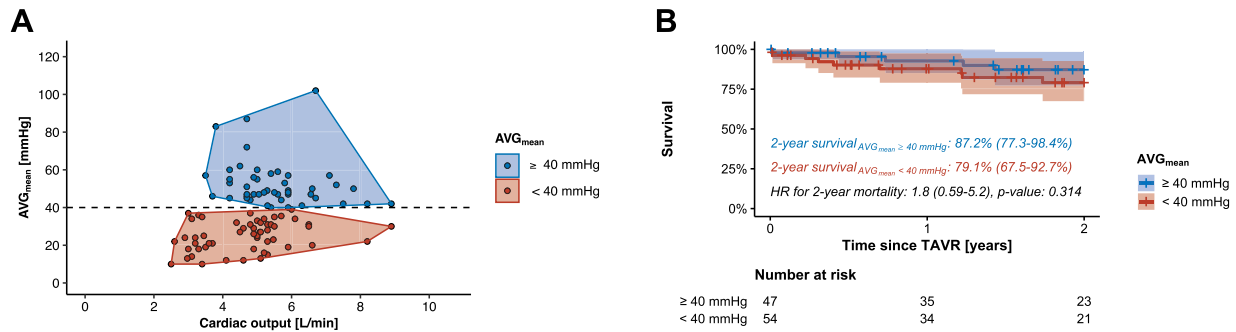


Figure 4 Conventional dichotomization of the study population in accordance with elevation in mean aortic valve gradient. (A) Scatter plot illustrating cardiac output and mean aortic valve gradient after dichotomization in accordance with elevation in mean aortic valve gradient. (B) Kaplan–Meier survival analysis in accordance with elevation in mean aortic valve gradient. AVG_{mean}, mean aortic valve gradient.

subsequently reduced (Figure 6D), and the hazard ratio for 2-year mortality after TAVR was significantly increased (2.6, 95% CI: 1.4–5.1, P -value: 0.004). Importantly, a less severe obstruction of the aortic valve was found again in patients assigned to high-risk cluster 2 (AVA: 0.839 ± 0.219 , 0.789–0.889 in Cluster 2 vs. 0.742 ± 0.186 , 0.715–0.769 in Cluster 1, P -value: 0.0007) (Table 4), confirming the initially surprising finding from the derivation cohort (Figure 6E).

Discussion

Transfer learning exploiting big data could be key to overcome the obstacle of data scarcity as commonly encountered in clinical reality, and learning from a related problem aids in gaining novel insights into phenotypic presentations of patients with severe aortic stenosis

Identifying patients at risk is a core element in the practice of medicine, but risk stratification for patients with severe AS in contemporary clinical practice is often limited by hypothesis-driven selection of a few factors typically regarded in isolation, by suggesting a model of orderly progression of accumulated pathologies upstream of the causative AS, or by the assumption of a parametric linear relationship between predictor variable and outcome. This study demonstrates that prognostic resolution of survival in patients with severe AS undergoing TAVR can be refined by harnessing the intriguing feature extraction capacity from an established CNN pre-trained on big data in order to subsequently recognize complex geometries in aortic outflow velocity profiles, which integrate crucial information about left ventricular contractility and aortic valve obstruction. Thus, two major phenotypes with important clinical implications

could be unravelled. The main messages from our study are therefore as follows (Graphical Abstract):

- (1) Transfer learning has the potential to unearth hidden gems even in clinical data sets of limited size.
- (2) Not so much the actual stenosis of the aortic valve expressed as AVA determines the prognosis after TAVR, but the left ventricular compensation capacity and subsequent development of PH and right heart failure stratify patients into low-risk and high-risk cohorts.

On the drawbacks of traditional methods for risk assessment—and how machine learning technology can pave the way to personalized risk stratification prior to transcatheter aortic valve replacement

In order to illustrate the almost ubiquitous problem of data scarcity in medical research on the one hand, and the vast potential of transfer learning, on the other hand, the number of aortic outflow velocity profiles to be analysed was intentionally kept small. Possibly, differences in survival after dichotomization according to AVG_{mean} would have become statistically significant, if more patients were included. Nonetheless, dichotomization of continuous variables is prone to reducing statistical power without notable benefit (oversimplification), and physicians in a real-world scenario therefore rarely rely on a single variable's dichotomy for decision-making or prognostic assessment, but rather prefer context-specific interpretation of extensive (raw) data. Aiming to detect predictors of mortality among a similar cohort of patients with severe AS undergoing TAVR, Weber *et al.*¹⁷ analysed a set of echocardiographic and haemodynamic data, and identified presence of combined pre- and post-capillary PH, and a lower AVG_{mean} as independent predictors by using multivariate Cox regression analysis. However, traditional regression models assume a parametric linear function relating the predictor variables with the response. This assumption might not hold true in the natural

Table 3 Univariate and multivariate cox regression analysis with 2-year mortality as a dependent variable (derivation cohort)

	Univariate analysis		Multivariate analysis	
	HR (95% CI)	P-value	HR (95% CI)	P-value
Age	0.96 (0.89–1) per year	0.21		
Sex (female)	0.26 (0.074–0.95)	0.042	0.20 (0.04–1.02)	0.0531
BMI	0.93 (0.81–1.1) per kg/m ²	0.26		
Arterial hypertension	1.9 (0.24–14)	0.55		
Smoking	1 (0.33–3)	1		
Diabetes mellitus	1.9 (0.62–5.5)	0.27		
NYHA functional class	1.4 (0.57–3.2) per class	0.49		
EuroSCORE	1 (1–1.1) per %	0.0017	1.03 (0.99–1.08) per %	0.1374
eGFR	0.97 (0.95–1) per mL/min	0.024	0.97 (0.94–1.00) per mL/min	0.0831
Hb	1 (0.75–1.4) per g/dL	0.95		
CAD	0.91 (0.2–4.1)	0.91		
COPD	2.1 (0.6–7.7)	0.24		
Atrial fibrillation and/or flutter	3.7 (1.2–12)	0.026	1.71 (0.35–8.50)	0.5099
AVA	1.4 (0.16–13) per cm ²	0.75		
AVG _{max}	0.98 (0.96–1) per mmHg	0.13		
AVG _{mean}	0.97 (0.94–1) per mmHg	0.11		
Cardiac output	0.66 (0.42–1) per L/min	0.064		
LVEF	0.94 (0.9–0.97) per %	0.0004	0.96 (0.84–1.09) per %	0.4913
LVEDD	1.1 (1–1.1) per mm	0.0056	0.95 (0.87–1.05) per mm	0.3149
mPAP	1 (1–1.1) per mmHg	0.04	1.00 (0.94–1.08) per mmHg	0.9269
mPCWP	1 (1–1.1) per mmHg	0.073		
PVR	1.3 (1–1.6) per WU	0.045	0.81 (0.47–1.40) per WU	0.4539
TAPSE	0.85 (0.73–1) per mm	0.045	0.85 (0.67–1.09) per mm	0.1967
Right midventricular diameter	1.1 (1–1.2) per mm	0.02	1.06 (0.96–1.18) per mm	0.2344
LA area	1.1 (1–1.1) per cm ²	0.011	1.03 (0.93–1.14) per cm ²	0.5449
RA area	1.1 (1.1–1.2) per cm ²	7.7×10^{-5}	1.04 (0.94–1.16) per cm ²	0.4110
Low gradient (AVG _{mean} < 40 mmHg)	1.8 (0.59–5.2)	0.314		
LV dysfunction (LVEF ≤ 45%)	4.7 (1.6–14)	0.0052	0.44 (0.03–7.33)	0.5663
PH (mPAP ≥ 25 mmHg)	1.6 (0.53–4.7)	0.42		
RV dysfunction (TAPSE ≤ 16 mm)	2.3 (0.77–6.8)	0.14		
Assignment to Cluster 2	3 (1–8.9)	0.04	1.12 (0.29–4.37)	0.8676

AVA, aortic valve area; AVG_{max}, maximum aortic valve gradient; AVG_{mean}, mean aortic valve gradient; BMI, body mass index; CAD, coronary artery disease; CI, confidence interval; COPD, chronic obstructive pulmonary disease; GFR, glomerular filtration rate; HR, hazard ratio; IVS, interventricular septum thickness; LA area, left atrial area; LVEDD, left ventricular end-diastolic diameter; LVEF, left ventricular ejection fraction; LVESD, left ventricular end-systolic diameter; mean RV pressure, mean right ventricular pressure; mPAP, mean pulmonary artery pressure; mPCWP, mean postcapillary wedge pressure; NYHA, New York Heart Association; PVR, pulmonary vascular resistance; PW, posterior wall thickness; RA area, right atrial area; RA pressure, right atrial pressure; TAPSE, tricuspid annular plane systolic excursion.

course of AS, since the AVG_{mean} initially increases with progressive narrowing of the AVA, but later decreases as the left ventricle decompensates ('low flow, low gradient AS'). If patients on the transition from moderate to severe AS were analysed by logistic regression analysis, an increasing AVG_{mean} would clearly be interpreted as an indicator for disease progression, and hence serve as a marker for worsened prognosis.¹⁸ The beauty of the hereby established approach lays in the improvement to identify and segregate patients with similar characteristics firstly without applying any *a priori* assumption and secondly without restricting the analysis to human-selected patient characteristics as data features. At the same time, many Doppler tracings were not suitable to be analysed by a CNN due to poor echocardiographic acquisition or due to suboptimal alignment

with the jet and hence inadequate recording of the true transvalvular gradient. So, how can our assignment to distinct clusters and its clinical implication be generalized to the majority of patients? Ideally, this study is not only perceived as a proof-of-concept valid for selected patients, but as yet another step along the road to implementation of artificial intelligence in clinical decision-making. We have therefore decided to additionally train an extreme gradient boosting algorithm on functional and structural data from pre-procedural echocardiography and right heart catheterization, thus opening the avenue for other cardiologists to stratify their patients according to our beforehand created classification generated by transfer learning. Upon loading the trained extreme gradient boosting algorithm and adding the requested input data into

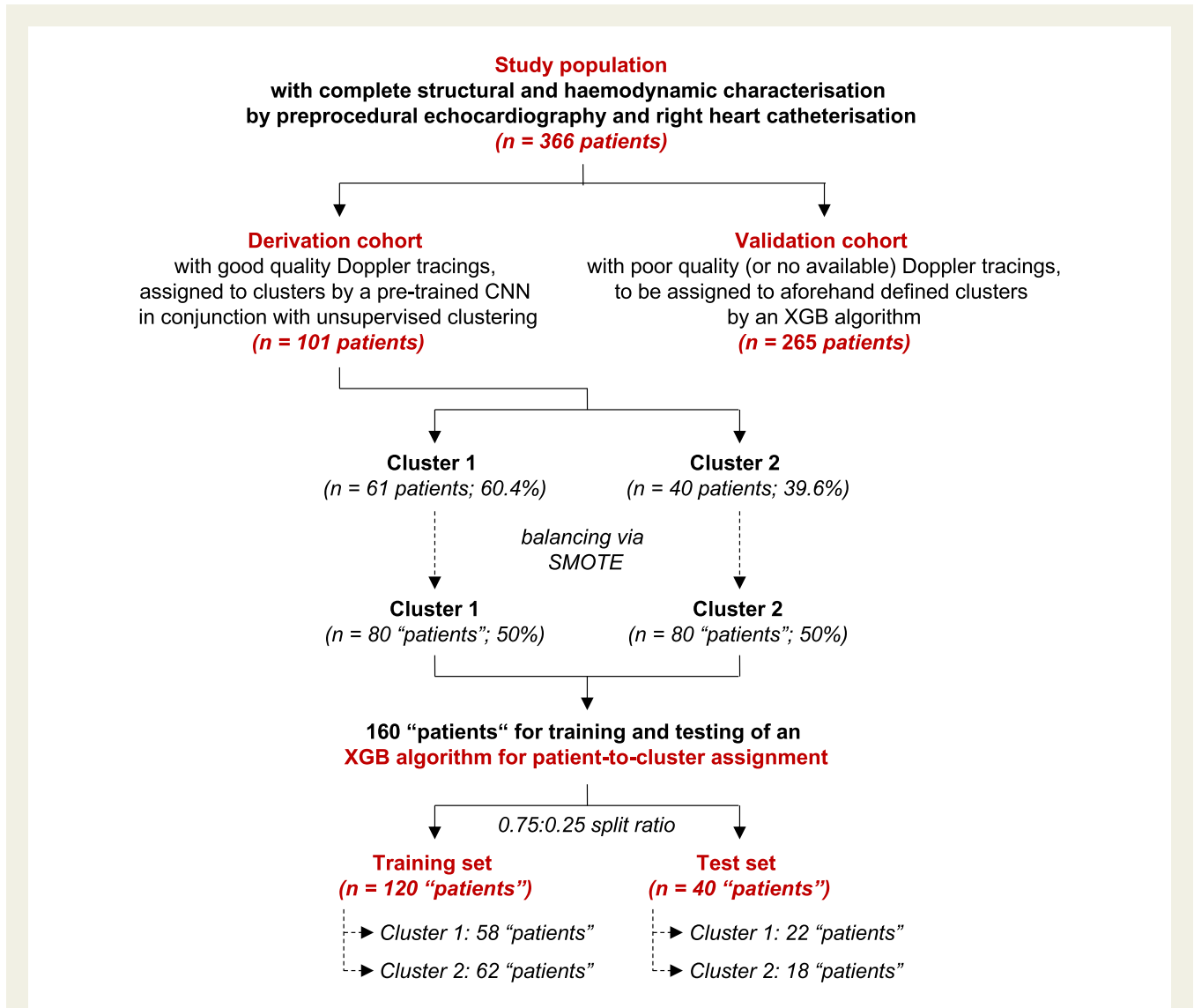


Figure 5 A flowchart illustrating the application of SMOTE to create a balanced data set for training of the extreme gradient boosting algorithm. CNN, convolutional neural network; SMOTE, synthetic minority over-sampling technique; XGB algorithm, extreme gradient boosting algorithm.

the corresponding R code (both available from the corresponding author; see [Supplementary material online, Figure S4](#) for a preview of the R code), future patients can be assigned to either Cluster 1 (good prognosis) or Cluster 2 (poor prognosis).

The extent of cardiac damage is already mirrored in the aortic outflow velocity profile, and it is the left ventricular response to the increased afterload that determines fate in patients with severe aortic stenosis

Capturing the complexity of cardiac damage subsequent to severe AS is key to sophisticated risk stratification prior to TAVR. This is particularly true, as PH and right ventricular dysfunction can persist in a substantial number of cases after TAVR, and persistence

translates into distressing mortality.^{19–21} Généreux *et al.*⁷ therefore established a staging classification, which considers disease progression beyond the compensation capacity of the left ventricle. This divisive, top-down staging classification is driven by the hypothesis that extravalvular damages to the heart and pulmonary circulation subsequent to severe AS occur in a sequential order of left heart failure, PH, and right heart dysfunction. Despite its simplicity, this staging classification cannot be easily implemented into clinical practice, as clinicians commonly encounter disparities between AS-induced haemodynamic burden and extravalvular damages (possibly influenced by comorbidities, such as atrial fibrillation and chronic obstructive pulmonary disease, or by genetic predisposition).^{16,22,23} Failure of left ventricular compensation capacity and subsequent backwards transmission of elevated left-sided filling pressures was more frequently observed in Cluster 2 than in Cluster 1 (mean post-capillary wedge pressure: 21.4 ± 8.83 , 95% CI:

18.6–24.2 mmHg vs. 15.5 ± 8.40 , 95% CI: 13.3–17.6 mmHg, P -value: 0.0007), whilst patients in Cluster 1 presented with a more severe obstruction of the aortic valve, indicating a longer disease progression, yet resulting in less cardiopulmonary impairments. This insight into disease progression in patients with severe AS emphasizes the importance of the complex pathophysiologic valvular-ventricular interactions, which obviously vary among individuals. In the contemporary ‘one-size-fits-all’ practice of medicine, the timing of intervention mainly focuses on the aortic valve. Our study may alter the perception of ideal timing of intervention as well as it may facilitate the development of individualized treatment, as an earlier intervention in patients from Cluster 2 might have had prevented further aggravation of left heart decompensation, PH, and right heart dysfunction. Addressing a similar issue, a study investigating the benefit of early intervention in patients with moderate AS and impaired left ventricular function has already been initialized [TAVR UNLOAD (Transcatheter Aortic Valve Replacement to Unload the Left Ventricle in Patients with Advanced Heart Failure) trial].²⁴ Moreover, it will be interesting to analyse future echocardiographic follow-up studies in accordance with cluster assignment, since recovery from cardiopulmonary damages that cannot be totally attributed to the obstruction of the aortic valve seems questionable. Thus, suspected persistence of PH and right heart dysfunction despite

correction of severe AS by TAVR could emerge as an unmodifiable (?) driver for increased mortality in patients from Cluster 2.

Unsupervised clustering could reveal diversity of aortic stenosis phenotypes with unprecedented precision, but extensive quality control is mandatory before unleashing machine learning algorithms in clinical practice

Extending this proof-of-principle study based on good quality Doppler tracings from 101 patients to a larger cohort could reveal even more diversity in aortic outflow velocity profiles by unravelling additional clusters. Even nowadays, AS with discordant markers of severity, such as severely reduced AVA and low AVG_{mean} , but preserved left ventricular ejection fraction (‘paradoxical low-gradient AS’)²⁵ remains a conundrum in diagnosis and treatment.²⁶ It will be interesting to see whether contemporary classifications of AS phenotypes will be mirrored by unsupervised clustering, or if distinct clinical presentations will emerge. Admittedly, involvement of artificial intelligence in clinical decision-making is still frowned upon due to the ‘black box’ nature, and the potential for a flawed machine learning algorithm to induce iatrogenic harm is vast. The opaqueness in the

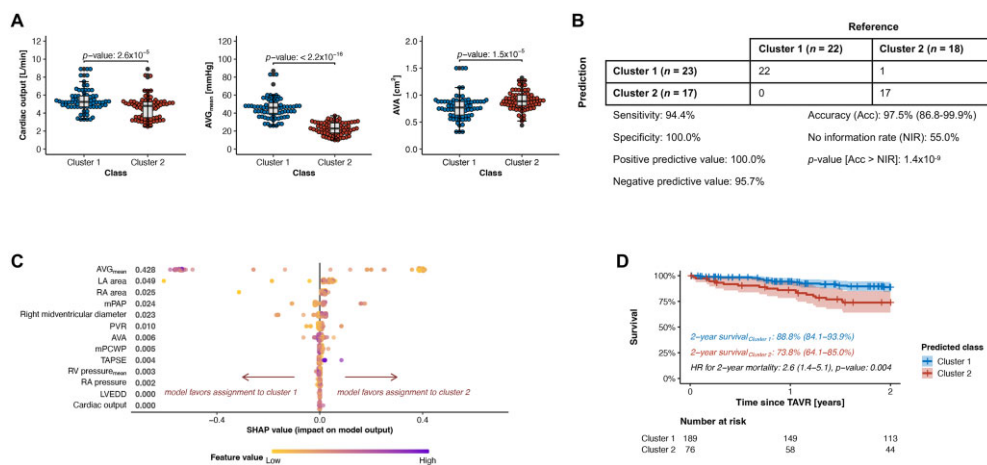


Figure 6 An extreme gradient boosting algorithm opens the perspective to assign patients to beforehand defined clusters by a comprehensive set of functional and structural parameters of cardiac and pulmonary circulatory conditions. (A) Bee swarm plots for comparison of key characteristics between clusters after over-sampling (SMOTE). (B) Confusion matrix (test set). (C) Shedding light on the black box of extreme gradient boosting algorithm-mediated cluster assignment by calculating SHAP (SHapley Additive exPlanations) values for its input variables. The y-axis represents the input variables in descending order of global feature importance, whilst the x-axis indicates the adjustment to the predicted cluster. Moreover, each dot in this sina plot represents an observation, i.e. a patient from the derivation cohort, and the gradient colour denotes the value of the respective input variable. Therefore, if the dots on one side of the central line are increasingly yellow or purple, that suggests that increasing values or decreasing values, respectively, move the predicted cluster in the respective direction (left: Cluster 1; right: Cluster 2). For instance, higher values of AVG_{mean} (purple dots) are associated with assignment to Cluster 1. (D) Kaplan–Meier survival analysis in accordance with extreme gradient boosting-algorithm-mediated cluster assignment (validation cohort). (E) Comparison of clusters as defined by the CNN in conjunction with PCA and k-means clustering (derivation cohort; red) or as determined by the trained extreme gradient boosting algorithm (validation cohort; blue). The central line in each box plot denotes the median value, while the box contains all values ranging between the 25th and 75th percentiles of the data set. The black whiskers mark the 5th and 95th percentiles, and values falling beyond these upper and lower bounds are considered outliers, plotted as black dots. AVA, aortic valve area; AVG_{mean} , mean aortic valve gradient; LA area, left atrial area; LVEDD, left ventricular end-diastolic diameter; mPAP, mean pulmonary artery pressure; mPCWP, mean postcapillary wedge pressure; PVR, pulmonary vascular resistance; RA area, right atrial area; RA pressure, right atrial pressure; RV pressure_{mean}, mean right ventricular pressure; TAPSE, tricuspid annular plane systolic excursion.

Continued

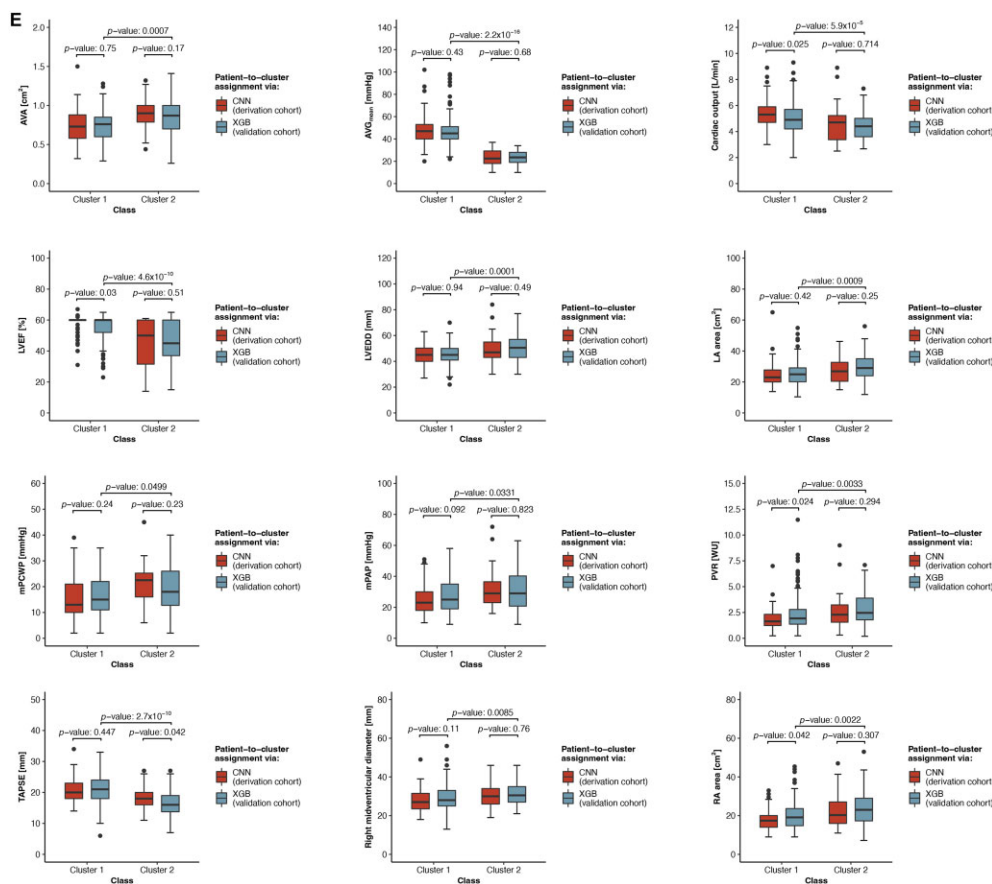


Figure 6 (Continued)

determination of output has therefore fuelled demands for explainability as expressed in the European Union's General Data Protection Regulation.²⁷ Gradient-weighted class activation mapping (GRAD-CAM) visualizations from the tool box of explainable artificial intelligence are typically applied in order to inspect images and to get insights into CNN decisions.²⁸ Yet, GRAD-CAM visualizations connecting the raw image to the decision of a classifier were not used in this study for two reasons:

- (1) Patient-to-cluster assignment was based on PCA and k-means clustering of Doppler tracings, which represents a form of unsupervised learning, and which is different from assignment by means of a trained classifier, which would have represented a form of supervised learning.
- (2) Training a classifier in terms of fully connected layers following the convolutional part of the pre-trained VGG-16 network would have required thousands of Doppler tracings, which could have only been collected in a labour-intensive, multicentric effort.

It, therefore, remains enigmatic which characteristics of the aortic outflow velocity profile would result in assignment to either Cluster 1 or 2. As shown by the PCA (Figure 3B), there cannot be a single 'most important' feature that defines the echocardiographic signature of patients assigned to Cluster 1 or 2, as the first two dimensions of the PCA explain only 10.72% and 8.96% of the variation among all

transformed aortic outflow velocity profiles, respectively. This gap in mechanistic inference can be perceived as a limitation to this study, but it also demonstrates the strengths of neural networks, which enable to identify novel relationships in complex and finely nuanced data sets and which therefore go beyond (simplified) stratification in accordance with human-selected features.²⁹ To explain at least partially which feature within the aortic outflow velocity profiles drive the differences between Clusters 1 and 2, the 20 most distant Doppler tracings (hereinafter referred to as 'top 10' and 'bottom 10' Doppler tracings) along PCA dimension #1 were identified (Supplementary material online, Figure S5A–C) and related echocardiographic and haemodynamic characteristics were compared (Supplementary material online, Figure S5D and Table S4); among the studied characteristics, the strongest difference in terms of statistical significance expressed as the respective *P*-value level was found for AVG_{mean} (59.5 ± 15.6 ; 95% CI: 48.3–70.7 mmHg among Top 10 Doppler tracings vs. 17.9 ± 7.62 ; 95% CI: 12.4–23.4 mmHg among bottom 10 Doppler tracings, *P*-value: 0.0002). This finding was confirmed by direct comparison of top 10 and bottom 10 Doppler tracings (Supplementary material online, Figure S5E).

Scrutinizing the generalizability of our findings as generated on 101 selected patients, we decided to test if cluster-related phenotypes could also be detected among the initially excluded 265 patients due

Table 4 Comparison of echocardiographic and haemodynamic characteristics in accordance with cluster assignment (validation cohort)

	All (n = 265)	Class		P-value
		Cluster 1 (n = 189)	Cluster 2 (n = 76)	
AVA (cm ²), mean ± SD [95% CI]	0.770 ± 0.201 [0.746–0.794]	0.742 ± 0.186 [0.715–0.769]	0.839 ± 0.219 [0.789–0.889]	0.0007
AVG _{mean} (mmHg), mean ± SD [95% CI]	40.4 ± 15.3 [38.5–42.2]	47.1 ± 12.3 [45.3–48.9]	23.5 ± 6.00 [22.1–24.9]	<2.2 × 10 ⁻¹⁶
Cardiac output (L/min), mean ± SD [95% CI]	4.86 ± 1.19 [4.72–5.01]	5.04 ± 1.23 [4.86–5.22]	4.42 ± 0.970 [4.19–4.64]	5.9 × 10 ⁻⁵
LVEF (%), mean ± SD [95% CI]	52.6 ± 10.8 [51.3–53.9]	55.5 ± 8.05 [54.3–56.7]	45.4 ± 13.1 [42.4–48.4]	4.6 × 10 ⁻¹⁰
LVEDD (mm), mean ± SD [95% CI]	46.9 ± 8.16 [45.8–47.9]	45.4 ± 7.41 [44.3–46.5]	50.4 ± 8.86 [48.2–52.5]	0.0001
LA area (cm ²), mean ± SD [95% CI]	26.5 ± 8.33 [25.4–27.6]	25.3 ± 7.85 [24.1–26.5]	29.3 ± 8.82 [27.2–31.4]	0.0009
mPCWP (mmHg), mean ± SD [95% CI]	17.3 ± 8.28 [16.3–18.3]	16.6 ± 7.56 [15.5–17.6]	19.2 ± 9.66 [17.0–21.4]	0.0499
mPAP (mmHg), mean ± SD [95% CI]	28.5 ± 11.4 [27.1–29.9]	27.5 ± 10.8 [25.9–29.0]	31.1 ± 12.5 [28.2–34.0]	0.0331
PVR (WU), mean ± SD [95% CI]	2.48 ± 1.59 [2.29–2.67]	2.34 ± 1.58 [2.11–2.56]	2.83 ± 1.57 [2.47–3.19]	0.0033
TAPSE (mm), mean ± SD [95% CI]	19.6 ± 5.36 [18.9–20.3]	21.0 ± 5.10 [20.2–21.7]	16.3 ± 4.53 [15.3–17.4]	2.7 × 10 ⁻¹⁰
Right midventricular diameter (mm), mean ± SD [95% CI]	29.7 ± 6.65 [28.8–30.5]	29.0 ± 6.89 [28.0–30.1]	31.3 ± 5.81 [29.9–32.6]	0.0085
RA area (cm ²), mean ± SD [95% CI]	21.0 ± 7.68 [20.0–22.0]	19.9 ± 6.95 [18.9–21.0]	23.5 ± 8.75 [21.4–25.6]	0.0022

AVA, aortic valve area; AVG_{mean}, mean aortic valve gradient; CI, confidence interval; LA area, left atrial area; LVEDD, left ventricular end-diastolic diameter; LVEF, left ventricular ejection fraction; mPAP, mean pulmonary artery pressure; mPCWP, mean postcapillary wedge pressure; PVR, pulmonary vascular resistance; RA area, right atrial area; RV dysfunction, right ventricular dysfunction; SD, standard deviation; TAPSE, tricuspid annular plane systolic excursion.

to poor or missing Doppler tracings. The cluster-related clinical implications could be confirmed by an extreme gradient boosting algorithm, and calculation of SHAP values as a state-of-the-art metric to quantify the contribution of input variables to model prediction highlighted the importance of transvalvular gradients, incorporating information about both aortic valve obstruction and left ventricular contractility. Notably, continuous wave Doppler echocardiography in combination with the Bernoulli equation to assess transvalvular pressure gradients is based on oversimplification of human haemodynamics, as for instance a column of flow with uniform velocity distribution is assumed, which is clearly not the case in patients with severe AS.³⁰ The analysis of the spatio-temporal pattern of the ejection jet, e.g. by three-dimensional cardiovascular magnetic resonance imaging, could therefore reveal novel insights into AS phenotypes.

Limitations: on the prohibitive costs of poor image quality, and why you should not trust artificial intelligence implicitly

Machine learning algorithms *per se* learn from data, meaning that insufficient data quality or systematic bias during data collection would hamper the algorithm to identify any consistent and

generalizable patterns. The accuracy of a CNN therefore strictly relies on the input data quality. Physicians in a real-world scenario yet commonly encounter difficulties in examining patients with severe AS, as they typically present dyspnoeic and are hence less suited for optimal positioning for echocardiography. It was, therefore, important to demonstrate that the subset of 101 patients with good quality Doppler tracings was representative of the entire study population of 366 patients (Supplementary material online, Tables S2 and S3). Moreover, we had to ensure by cumbersome manual cropping that the Doppler tracings serving as input images contain no other information than the aortic outflow velocity profile of interest (Figure 2). An example of a seemingly high-performance machine learning algorithm flawed by shortcuts in the training set is a model that is supposed to distinguish a wolf from a husky by animal characteristics but eventually reveals to derive its performance from the simple, but undesired identification of patches of snow on the photograph.³¹ Moreover, the unscrutinized synthesis of training data from separate data sets of COVID-19-negative and COVID-19-positive images was demonstrated to introduce near worst-case confounding and thus abundant opportunity for machine learning algorithms to learn shortcuts due to variations in image acquisition and radiographic projection.³²

Claiming to have found a reasonable echocardiographic signature among patients presenting with severe AS, it was therefore of paramount importance to us to validate (and finally confirm) the clinical implications of related phenotypes in a second cohort by yet another machine learning algorithm. Obviously, we cannot guarantee that the algorithms employed in this study would outperform all other algorithms in clustering patients (unsupervised learning experiment) and in assigning patients to clusters (supervised learning experiment). In the commonly accepted absence of any *a priori* guarantee that one machine learning technique is superior to all others,³³ the only way to determine which algorithms works best for the given data structure is to evaluate them all. However, this is practically impossible, innovative and more powerful algorithms might emerge in the future, and it ultimately also relies on the programmer's ability to tune the hyperparameter of respective models to perfection. As a matter of fact, we applied hierarchical agglomerative clustering to the transformed aortic outflow velocity profiles, hence testing yet another popular clustering algorithm equivalent to k-means clustering. Hierarchical agglomerative clustering also facilitated to identify a cluster with significantly reduced AGV_{mean} ; however, the two segregated clusters were vastly overlapping as demonstrated by the first two dimensions of a PCA as well as by a correlation plot depicting AGV_{mean} and cardiac output, and subsequently, 2-year survival differences did not reach statistical significance (Supplementary material online, Figure S6). Importantly, AS represents a progressive disease with a continuous transition of stages of disease severity. Unlike clustering of e.g. bone marrow cells of distinct haematopoietic lineages (where you would expect clearly defined clusters of e.g. erythrocytes and lymphatic cells based on their gene expression profiles), it is practically impossible to distinguish any clearly separated clusters among patients with severe AS and their respective aortic outflow velocity profiles. This is reflected by the silhouette diagram (Supplementary material online, Figure S7) revealing a mean silhouette coefficient of only 0.0689 ± 0.0459 among the clusters as defined by k-means clustering. Moreover, we acknowledge that the accuracy of the extreme gradient boosting algorithm was evaluated on a test set that was at least partially composed of synthetic data. We have therefore added an alternative experimental design with a test set containing only real and unseen patients (explicitly no synthetic data) (Supplementary material online, Figure S8A). Thus, we could confirm the satisfying accuracy of the extreme gradient boosting algorithm for patient-to-cluster assignment based on 13 variables from pre-procedural echocardiography and right heart catheterization (accuracy: 92.0%; 95% CI: 74.0–99.9%) (Supplementary material online, Figure S8B). Applying the algorithm trained under the alternative experimental design to the validation cohort of patients with poor quality or no available Doppler tracings also confirmed the increased risk of mortality for patients assigned to Cluster 2 in comparison to Cluster 1 (HR for 2-year mortality: 2.1; 95% CI: 1.1–4.1, *P*-value: 0.022) (Supplementary material online, Figure S8C). Again, patients assigned to Cluster 2 were characterized by a relatively larger AVA (0.824 ± 0.214 ; 95% CI: 0.779–0.870 cm²) and also by a reduced left ventricular function (left ventricular ejection fraction: 47.5 ± 13.3 ; 95% CI: 44.7–50.4%) and by a lower AVG_{mean}

(25.6 ± 8.14 , 95% CI: 23.9–27.4 mmHg) (Supplementary material online, Figure S8D and Table S5).

Conclusion

In summary, this is the first study to demonstrate the usefulness of transfer learning for unsupervised clustering of aortic outflow velocity profiles in patients with severe AS. Since the perception of patients presenting with severe AS is in a state of flux from a valve-centred perspective to a personalized comprehensive view covering all aspects of co-developed cardiopulmonary impairments, the unravelled phenotypes in this study hold the promise to better stratify patients into low-risk and high-risk cohorts. Importantly, it is the left ventricular response to the increased afterload, not so much the actual obstruction of the aortic valve, that determines fate after TAVR. As a new arrow in the quiver from interventional cardiologists to refine prognostic assessment prior to TAVR, the trained extreme gradient boosting algorithm for individual cluster assignment in future patients can be requested from the corresponding author.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health* online.

Funding

M.L. has received funding from the Technical University of Munich (clinician scientist grant) and from the Else Kröner-Fresenius Foundation (clinician scientist grant).

Conflict of interest: none declared.

Data availability

The pre-processed (deidentified) aortic outflow velocity profiles from those 101 patients with good quality Doppler tracings are attached as Supplementary material online. Moreover, data concerning baseline echocardiographic and haemodynamic characteristics (which were used for training of the extreme gradient boosting algorithm) as well as the respective survival status are given for those 101 patients (including a code book). Furthermore, the complete R code and the trained extreme gradient boosting algorithm for patient-to-cluster assignment can be requested from the corresponding author.

References

1. Raghunath S, Ulloa Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med* 2020;**26**: 886–891.
2. Fries JA, Varma P, Chen VS, et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nat Commun* 2019;**10**: 3111.
3. Diller G-P, Kempny A, Babu-Narayan SV, et al. Machine learning algorithms estimating prognosis and guiding therapy in adult congenital heart disease: data from a single tertiary centre including 10 019 patients. *Eur Heart J* 2019;**40**:1069–1077.
4. Perez MV, Mahaffey KW, Hedlin H, et al.; Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med* 2019;**381**:1909–1917.

5. Kwak S, Lee Y, Ko T, et al. Unsupervised cluster analysis of patients with aortic stenosis reveals distinct population with different phenotypes and outcomes. *Circ Cardiovasc Imaging* 2020;**13**.
6. Sengupta PP, Shrestha S, Kagiya N, et al.; Artificial Intelligence for Aortic Stenosis at Risk International Consortium. A machine-learning framework to identify distinct phenotypes of aortic stenosis severity. *JACC Cardiovasc Imaging* 2021;**14**:1707–1720.
7. Généreux P, Pibarot P, Redfors B, et al. Staging classification of aortic stenosis based on the extent of cardiac damage. *Eur Heart J* 2017;**38**:3351–3358.
8. Makkar RR, Fontana GP, Jilaihawi H, et al. Transcatheter aortic-valve replacement for inoperable severe aortic stenosis. *N Engl J Med* 2012;**366**:1696–1704.
9. Baumgartner H, Falk V, Bax JJ, et al.; ESC Scientific Document Group. 2017 ESC/EACTS Guidelines for the management of valvular heart disease. *Eur Heart J* 2017;**38**:2739–2791.
10. Bing R, Gu H, Chin C, et al. Determinants and prognostic value of echocardiographic first-phase ejection fraction in aortic stenosis. *Heart* 2020;**106**:1236–1243.
11. Chambers J. Low 'gradient', low flow aortic stenosis. *Heart* 2006;**92**:554–558.
12. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *ArXiv14091556 Cs* 2015.
13. Stekhoven DJ, Buhlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;**28**:112–118.
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002;**16**:321–357.
15. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;**2**:56–67.
16. Lachmann M, Rippen E, Schuster T, et al. Subphenotyping of patients with aortic stenosis by unsupervised agglomerative clustering of echocardiographic and hemodynamic data. *JACC Cardiovasc Interv* 2021;**14**:2127–2140.
17. Weber L, Rickli H, Haager PK, et al. Haemodynamic mechanisms and long-term prognostic impact of pulmonary hypertension in patients with severe aortic stenosis undergoing valve replacement: Impact of PH in severe aortic stenosis. *Eur J Heart Fail* 2019;**21**:172–181.
18. Slimani A, Roy C, de Meester C, et al. Structural and functional correlates of gradient-area patterns in severe aortic stenosis and normal ejection fraction. *JACC Cardiovasc Imaging* 2021;**14**:525–536.
19. Masri A, Abdelkarim I, Sharbaugh MS, et al. Outcomes of persistent pulmonary hypertension following transcatheter aortic valve replacement. *Heart* 2018;**104**:821–827.
20. Cremer PC, Zhang Y, Alu M, et al. The incidence and prognostic implications of worsening right ventricular function after surgical or transcatheter aortic valve replacement: insights from PARTNER IIA. *Eur Heart J* 2018;**39**:2659–2667.
21. Asami M, Stortecky S, Praz F, et al. Prognostic value of right ventricular dysfunction on clinical outcomes after transcatheter aortic valve replacement. *JACC Cardiovasc Imaging* 2019;**12**:577–587.
22. Guzzetti E, Annabi M-S, Pibarot P, Clavel M-A. Multimodality imaging for discordant low-gradient aortic stenosis: assessing the valve and the myocardium. *Front Cardiovasc Med* 2020;**7**:570689.
23. Little SH, O'Gara PT. Considering the hazards of aortic valve stenosis. *JACC Cardiovasc Imaging* 2021;**14**:1738–1741.
24. Spitzer E, Van Mieghem NM, Pibarot P et al. Rationale and design of the Transcatheter Aortic Valve Replacement to Unload the Left ventricle in patients with Advanced heart failure (TAVR UNLOAD) trial. *Am Heart J* 2016;**182**:80–88.
25. Hachicha Z, Dumesnil JG, Bogaty P, Pibarot P. Paradoxical low-flow, low-gradient severe aortic stenosis despite preserved ejection fraction is associated with higher afterload and reduced survival. *Circulation* 2007;**115**:2856–2864.
26. Clavel M-A, Magne J, Pibarot P. Low-gradient aortic stenosis. *Eur Heart J* 2016;**37**:2645–2657.
27. Goodman B, Flaxman S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Mag* 2017;**38**:50–57.
28. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020;**128**:336–359.
29. Altes A, Thellier N, Bohbot Y, et al. Relationship between the ratio of acceleration time/ejection time and mortality in patients with high-gradient severe aortic stenosis. *J Am Heart Assoc* 2021;**10**:e021873.
30. Donati F, Myerson S, Bissell MM, et al. Beyond Bernoulli: improving the accuracy and precision of noninvasive estimation of peak pressure drops. *Circ Cardiovasc Imaging* 2017;**10**.
31. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* San Francisco, CA, USA: ACM; 2016. p.1135–1144.
32. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell* 2021;**3**:610–619.
33. Wolpert DH. The lack of a priori distinctions between learning algorithms. *Neural Comput* 1996;**8**:1341–1390.