Diagnosis and prognosis of abnormal cardiac scintigraphy uptake suggestive of cardiac amyloidosis using artificial intelligence: a retrospective, international, multicentre, cross-tracer development and validation study

Clemens P Spielvogel, David Haberl*, Katharina Mascherbauer, Jing Ning, Kilian Kluge, Tatjana Traub-Weidinger, Rhodri H Davies, Iain Pierce, Kush Patel, Thomas Nakuz, Adelina Göllner, Dominik Amereller, Maria Starace, Alice Monaci, Michael Weber, Xiang Li, Alexander R Haug, Raffaella Calabretta, Xiaowei Ma, Min Zhao, Julia Mascherbauer, Andreas Kammerlander, Christian Hengstenberg, Leon J Menezes, Roberto Sciagra, Thomas A Treibel, Marcus Hacker, Christian Nitsche*

Summary

Background The diagnosis of cardiac amyloidosis can be established non-invasively by scintigraphy using bone-avid tracers, but visual assessment is subjective and can lead to misdiagnosis. We aimed to develop and validate an artificial intelligence (AI) system for standardised and reliable screening of cardiac amyloidosis-suggestive uptake and assess its prognostic value, using a multinational database of ^{99m}Tc-scintigraphy data across multiple tracers and scanners.

Methods In this retrospective, international, multicentre, cross-tracer development and validation study, 16 241 patients with 19 401 scans were included from nine centres: one hospital in Austria (consecutive recruitment Jan 4, 2010, to Aug 19, 2020), five hospital sites in London, UK (consecutive recruitment Oct 1, 2014, to Sept 29, 2022), two centres in China (selected scans from Jan 1, 2021, to Oct 31, 2022), and one centre in Italy (selected scans from Jan 1, 2011, to May 23, 2023). The dataset included all patients referred to whole-body ^{99m}Tc-scintigraphy with an anterior view and all ^{99m}Tc-labelled tracers currently used to identify cardiac amyloidosis-suggestive uptake. Exclusion criteria were image acquisition at less than 2 h (^{99m}Tc-3,3-diphosphono-1,2-propanodicarboxylic acid, ^{99m}Tc-hydroxymethylene diphosphonate, and ^{99m}Tc-methylene diphosphonate) or less than 1 h (^{99m}Tc-pyrophosphate) after tracer injection and if patients' imaging and clinical data could not be linked. Ground truth annotation was derived from centralised core-lab consensus reading of at least three independent experts (CN, TT-W, and JN). An AI system for detection of cardiac amyloidosis-associated high-grade cardiac tracer uptake was developed using data from one centre (Austria) and independently validated in the remaining centres. A multicase, multireader study and a medical algorithmic audit were conducted to assess clinician performance compared with AI and to evaluate and correct failure modes. The system's prognostic value in predicting mortality was tested in the consecutively recruited cohorts using cox proportional hazards models for each cohort individually and for the combined cohorts.

Findings The prevalence of cases positive for cardiac amyloidosis-suggestive uptake was 142 (2%) of 9176 patients in the Austrian, 125 (2%) of 6763 patients in the UK, 63 (62%) of 102 patients in the Chinese, and 103 (52%) of 200 patients in the Italian cohorts. In the Austrian cohort, cross-validation performance showed an area under the curve (AUC) of 1·000 (95% CI 1·000–1·000). Independent validation yielded AUCs of 0·997 (0·993–0·999) for the UK, 0·925 (0·871–0·971) for the Chinese, and 1·000 (0·999–1·000) for the Italian cohorts. In the multicase multireader study, five physicians disagreed in 22 (11%) of 200 cases (Fleiss' kappa 0·89), with a mean AUC of 0·946 (95% CI 0·924–0·967), which was inferior to AI (AUC 0·997 [0·991–1·000], p=0·0040). The medical algorithmic audit demonstrated the system's robustness across demographic factors, tracers, scanners, and centres. The AI's predictions were independently prognostic for overall mortality (adjusted hazard ratio 1·44 [95% CI 1·19–1·74], p<0·0001).

Interpretation AI-based screening of cardiac amyloidosis-suggestive uptake in patients undergoing scintigraphy was reliable, eliminated inter-rater variability, and portended prognostic value, with potential implications for identification, referral, and management pathways.

Funding Pfizer.

Lancet Digit Health 2024; 6: e251–60

See **Comment** page e231

*First authors contributed equally

Department of Biomedical Imaging and Image-Guided Therapy, Division of Nuclear Medicine, Medical University of Vienna, Vienna, Austria (C P Spielvogel PhD, D Haberl MSc, K Kluge MD, T Traub-Weidinger MD, T Nakuz MD, A Göllner MD, D Amereller MD, M Weber PhD, X Li PhD, A R Haug MD, R Calabretta MD, Prof M Hacker MD); **Christian Doppler Laboratory for Applied Metabolomics, Medical University of Vienna, Vienna, Austria** (J Ning MD, A R Haug); **Institute of Cardiovascular Science, University College London, London, UK** (R H Davies MD, I Pierce PhD, T A Treibel MD, C Nitsche MD PhD); **Department of Medicine II, Division of Cardiology, Medical University of Vienna, Vienna, Austria** (K Mascherbauer MD, Prof J Mascherbauer MD, A Kammerlander MD, Prof C Hengstenberg MD, C Nitsche); **Bart's Heart Centre, St Bartholomew's Hospital, West Smithfield, London, London, UK** (R H Davies, I Pierce, K Patel MD, L J Menezes MD, T A Treibel, C Nitsche); **Department of Nuclear Medicine, Third Xiangya Hospital, Central South University, Changsha, China** (M Zhao MD); **Department of Nuclear Medicine, Second Xiangya Hospital, Central South University, Changsha, China** (X Ma MD); **Karl Landsteiner University of Health Sciences, Department of**

Introduction

Cardiac amyloidosis is caused by the deposition of misfolded proteins in the heart, eventually leading to heart failure and death.[1] The two main proteins deposited in the heart are transthyretin (ATTR) and immunoglobulin light chain (AL). Until 2016, myocardial biopsy represented the

Internal Medicine 3, University Hospital St Pölten, Krems, Austria (Prof J Mascherbauer); **Department of Experimental and Clinical Biomedical Sciences, Nuclear Medicine Unit, University of Florence, Florence, Italy** (M Starace MD, A Monaci MD, Roberto Sciagra MD)

Correspondence to: Dr Christian Nitsche, Department of Internal Medicine II, Division of Cardiology, Medical University of Vienna, A-1090 Vienna, Austria
christian.nitsche@meduniwien.ac.at

Research in context

Evidence before this study

We searched PubMed on Sept 26, 2023, from database inception using the terms (("cardiac") AND ("amyloidosis") AND ("scintigraphy") AND ("artificial intelligence" OR "deep learning" OR "machine learning" OR "predictive modelling" OR "predictive modeling" OR "modelling" OR "modeling")). Publications concerned with the application of artificial intelligence (AI) or other modelling approaches for the diagnosis of cardiac amyloidosis based on scintigraphy imaging data were considered. All publications identified were in English. Two publications were concerned with deep learning-based models for the detection of cardiac amyloidosis. Both studies showed the feasibility of detecting high-grade cardiac uptake on whole-body scintigraphy scans with bone-avid tracers with varying extents of validation. The studies investigated a small number of tracers and validated their findings on non-consecutively sampled patients originating from one or multiple centres within one country. The potential risks and clinical relevance of the studies' approaches were only marginally elucidated, and the connection between the obtained results and clinical outcomes has not been clarified. To ensure the clinical value of these systems for the screening of cardiac amyloidosis, validation on consecutively sampled real-world data is required.

Added value of this study

To our knowledge, this international study of 16 241 patients is the largest study yet for the development and validation of an automated system for the detection of cardiac amyloidosis. This is the first study that associates the AI-based detection of cardiac amyloidosis with clinical outcomes, guaranteeing clinical relevance of the findings. Further, this study employs a comprehensive all-comer cohort (ie, all consecutive patients who underwent whole body ^{99m}-Tc scintigraphy with bone-avid tracers) acquired as part of clinical routine to ensure the usefulness for screening. It is the first study validating all available scintigraphy tracers currently used in clinical practice and capable of indicating presence of cardiac amyloidosis. Additionally, this study highlighted the discrepancy between physicians' assessments using the current clinical standard and investigated the comparative performance of the AI system with physicians.

Implications of all the available evidence

We expect an added benefit for clinicians and patients through the implementation of the developed system in parallel with the current clinical assessment. This system will reduce the number of undetected cases of cardiac amyloidosis, particularly among non-cardiology referrals and in smaller centres with less experience in cardiac amyloidosis. Lastly, this study extends the knowledge on inter-rater variability and prevalence in cardiac amyloidosis for patients undergoing scintigraphy with bone-avid tracers.

only modality to reliably diagnose cardiac amyloidosis; however, this method has associated inherent risks. Scintigraphy with amyloid-avid tracers has revolutionised the diagnostic pathway, because intense cardiac uptake in the absence of pathological light chains now enables a non-biopsy diagnosis of ATTR-cardiac amyloidosis.[2] Other causes of intense cardiac uptake are much less common and include AL-cardiac amyloidosis as well as very rare subtypes of cardiac amyloidosis,[3] whereas non-cardiac amyloidosis related conditions are extremely scarce. This scenario makes high-grade cardiac uptake a reliable feature for the presence of cardiac amyloidosis.

The spectrum of referral indications for ^{99m}Tc scintigraphy is broad, whereby the specific query for the presence of abnormal cardiac uptake only makes out a minority of referrals.[4] Hence, cardiac uptake might occur as an incidental finding and confront physicians from non-cardiac specialties. Therefore, the diagnosis of cardiac amyloidosis can be lost by nuclear medicine specialists not informing the treating physician, by the unawareness of non-cardiology physicians for the disease, and, rarely, due to the failure to identify abnormal cardiac uptake.

The correct identification of affected patients and streamlined referral to dedicated specialists are of particular importance, because disease-modifying, life-prolonging therapies are now available.[5,6] Automated detection of abnormal cardiac uptake using artificial intelligence (AI) can assist this process by enabling the efficient analysis of large datasets with consecutive activation of dedicated health-care pathways. One smaller sized study published in 2023 showed that deep learning might detect cardiac amyloidosis-suggestive uptake on planar scintigraphy scans with high sensitivity and specificity.[7] However, this research did not explore the association between predictions and clinical outcomes, the performance across different radiopharmaceutical tracers, or the performance in a comprehensive, all-comer cohort (ie, all consecutive patients who underwent whole body ^{99m}-Tc scintigraphy with bone-avid tracers).

We aimed to develop and validate an AI system for automated and objective detection of high-grade cardiac uptake and to assess its prognostic value in an international, multicentre, cross-tracer study.

Methods

Study design and patients

In this international multicentre retrospective study, we developed an AI system and trained it on a dataset of consecutive patients undergoing ^{99m}Tc-scintigraphy between Jan 4, 2010, and Aug 19, 2020, at the Vienna

General Hospital, Vienna, Austria, a university-affiliated tertiary care centre. External validation was performed on a dataset of consecutively recruited patients from the Barts Health NHS Trust consisting of five hospital sites in London, UK (Barts Heart Centre, Whipps Cross University Hospital, Royal London Hospital, Mile End Hospital, and Newham University Hospital; recruitment from Oct 1, 2014, to Sept 29, 2022), as well as selected scans from two centres in China (Frist Xiangya Hospital, Changsha and Second Xiangya Hospital, Changsha; recruitment from Jan 1, 2021, to Oct 31, 2022) and one centre in Italy (Careggi University Hospital, Florence; Jan 1, 2011, to May 23, 2023). Chinese scans included consecutive patients evaluated for suspected cardiac amyloidosis. For the Italian cohort, equal numbers of cases with and without cardiac amyloidosis-suggestive uptake were selected. The purposes of imaging for the cohorts from Austria and the UK are in the appendix (pp 12–13). Indications for the Italian cohort were not collected. The primary aim of the validation was to assess robustness of the AI system's predictions across multiple centres, scanners, and tracers.

Inclusion criteria were set up to be as inclusive as possible for the cohorts from Austria and the UK with the aim of acquiring all patients referred to whole-body ^{99m}Tc-scintigraphy with an anterior view. Patients scanned with a tracer known to be less sensitive but rather specific for cardiac amyloidosis (eg, ^{99m}Tc-methylene diphosphonate [^{99m}Tc-MDP]) were also eligible, because it was not the primary aim of this study to assess disease prevalence, which was done by previous work.[8] Exclusion criteria were image acquisition at less than 2 h (^{99m}Tc-3,3-diphosphono-1,2-propanodicarboxylic acid [DPD], ^{99m}Tc-hydroxymethylene diphosphonate [HMDP], and ^{99m}Tc-MDP) or less than 1 h (^{99m}Tc-pyrophosphate [PYP]) after tracer injection (as is conventionally performed and recommended by the current guidelines[9,10]) and if patients' imaging and clinical data could not be linked (eg, due to incompleteness or inconsistencies in the recorded study dates, birth dates, patient names, or identifiers of either clinical reports or DICOM files).

This study was granted ethical approval by the institutional review boards of the Medical University of Vienna (1557/2020) and the Barts Health NHS Trust (160139) and was performed in accordance with the 1964 Declaration of Helsinki. Informed consent was waived by both institutional review boards due to the retrospective nature of the study. The collection of patient data in the Chinese cohorts was approved by the respective institutional review boards (Frist Xiangya Hospital, Changsha and Second Xiangya Hospital, Changsha), who also waived informed consent. For the Italian cohort, patient signature on the examination authorisation form included the consent to the use of anonymised data for research purposes.

Imaging and grading

Patients were scanned using either Discovery 670 hybrid gamma camera (GE HealthCare Technologies, Chicago, IL, USA), Varicam (GE HealthCare Technologies, Chicago, IL, USA), IRIX (Koninklijke Philips, Amsterdam, Netherlands), Encore2 (Siemens Healthineers, Erlangen, Germany), or Tandem 870 CZT (GE HealthCare Technologies, Chicago, IL, USA) in Austria; Encore2 (Siemens Healthineers, Erlangen, Germany), BrightView (Koninklijke Philips, Amsterdam, Netherlands), or Skylight (Koninklijke Philips, Amsterdam, Netherlands) in the UK; Precedence 16 (Koninklijke Philips, Amsterdam, Netherlands) or Skylight (Koninklijke Philips, Amsterdam, Netherlands) in China; and Discovery 360 (GE HealthCare Technologies, Chicago, IL, USA) in Italy. ^{99m}Tc-DPD, ^{99m}Tc-HMDP, and ^{99m}Tc-MDP (Austria, UK, and Italy) were used with acquisition 2 h or more after injection. ^{99m}Tc-PYP (China) was used with 1 h or more after tracer injection. ^{99m}Tc-scintigraphy images were acquired according to the current guidelines.[9] The presence of cardiac tracer uptake was annotated according to the Perugini classification, whereby grade 0 represents no cardiac uptake, grade 1 represents minor cardiac uptake (heart less than bones), and grades 2 or 3 represent intense cardiac uptake (heart equal to or greater than bones). Perugini grade 2 or more is considered highly sensitive for ATTR-cardiac amyloidosis and highly specific for cardiac amyloidosis in general.[2,11] For the present analysis, grading was therefore dichotomised into cardiac amyloidosis-suggestive uptake (positive) with Perugini grade of 2 or more, and non-cardiac amyloidosis-suggestive uptake (negative) with Perugini grade of 1 or less. Grading was performed for the purpose of this study by a centralised core-laboratory (Medical University of Vienna) consisting of three independent imaging specialists (CN [cardiologist], and TT-W and JN [nuclear medicine physicians]) who were masked to each other and had extensive experience in the analysis of scintigraphy scans (≥5 years). In cases of discrepancy, which mostly occurred in borderline grade 1 versus grade 2, a board of two additional imaging specialists (RC and ARH; ≥10 years experience) performed a consensus reading to determine the grade.

See Online for appendix

AI system development and validation

Development of the AI system was based on multiple components (appendix p 32). First, the system loaded the scintigraphy scans from DICOM images and performed histogram standardisation to adjust pixel intensity distributions to a normalised range. Next, a deep learning-based thorax detection was performed on the standardised image. This object detection gave the output of a bounding box used for subsequent cropping of the image. The cropped thorax was fed to a convolutional neural network that gave the final classification of a positive versus a negative scan. In more detail, the model output a quantitative cardiac amyloid marker (CAM)

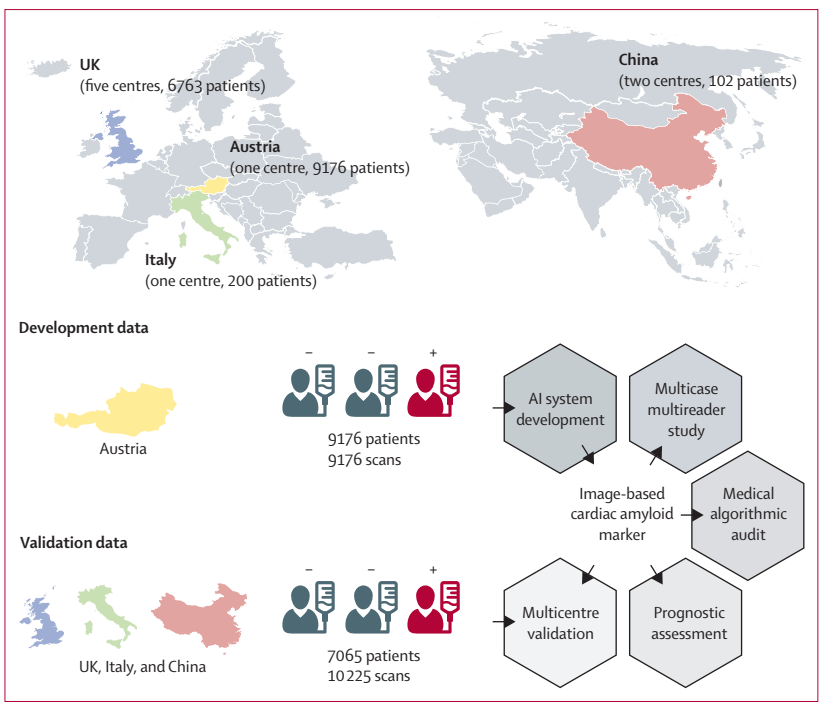

Figure 1: **Cohort origin overview**
Patients from nine centres in four countries and two continents were enrolled. The largest cohort in terms of patient number, the Austrian cohort, was employed for the development and initial validation of the AI system. The AI-based imaging biomarker output by the AI system was validated using the remaining patients from eight centres. AI=artificial intelligence.

ranging between 0 and 1, which represents the confidence score for detecting a positive scan. A CAM of at least 0·5 was considered as abnormal uptake suggestive of cardiac amyloidosis. Additionally, the system output a thorax detection quality score, indicating the confidence for a successful detection, and an attention map that visually represents which regions in the input image were most important for the given prediction. Both measures were primarily generated to enable continuous risk monitoring of the system.[12] Initial training and performance assessment were performed using ten-fold stratified cross-validation based on the Austrian cohort. The AI system was subsequently evaluated on independent scans of the UK, Chinese, and Italian cohorts, masked to the clinical annotation and diagnosis. Details and model specifications are explained in the medical algorithmic audit documents in the appendix (pp 4–10).

Multicase multireader study

In addition to the multicentre validation, a multicase multireader study was performed to assess interobserver variability and clinical value by comparing the AI system's performance with clinical expert readers. The diagnostic performance of the AI system was compared with a group of five board-certified nuclear medicine physicians (TN, AG, TT-W, DA, and MH). The ground truth for the comparison was based on the consensus assessment of a cardiologist specialising in cardiac amyloidosis and two nuclear medicine specialists (CN, RC, and ARH), all with extensive expertise in the analysis of scintigraphy using bone-avid tracers. Overall, 100 positive and 100 negative scans (based on the ground truth assessment) randomly sampled from the UK cohort were included in the multicase multireader study. The five readers were masked to the ground truth as well as to the AI system's prediction. The inter-rater variability of the clinical readers was assessed using Fleiss' kappa.

Clinical data

Baseline patient characteristics of sex, age, and comorbidities were retrieved from the respective hospital medical records systems or from the patient information included in the DICOM files. All-cause mortality was captured using national data via the Austrian Death Registry, and the UK National Health Service (Spine) and was 100% complete for the Austrian and UK cohorts. Information on hospitalisations due to heart failure was available for the Austrian cohort only, and determined from three sources, covering hospitalisations in all Austrian hospitals: patient records of the Medical University of Vienna, Vienna-Health-Association database, and the nation-wide electronic health records.

Statistical analysis

Continuous data are expressed as mean with SD or as median with IQR; categorical variables are presented as numbers and percentages. For the patient characteristics, inter-group differences were calculated using either Mann–Whitney *U* test or *t* test for numerical variables and either Fisher's exact test or χ^2 test for categorical values. Tests for normality to select the appropriate test were conducted using the Shapiro–Wilk test. Comparison of area under the curves (AUCs) was performed using the DeLong test. Machine learning performance metrics and associated formulas are shown in the appendix (p 11). For the Austrian cohort, mean performance and CIs for the AI system's prediction were calculated over the ten-fold cross-validation. For the calculation of CIs in the validation cohorts and for the statistical comparison of the AI system's performance with nuclear medicine physicians, bootstrapped samples with 1000 iterations were used. Kaplan–Meier estimates and Cox regression analyses were used to evaluate the prognostic significance of a positive CAM, starting with the date of scintigraphy. In patients with multiple scans, only the initial scan was included for outcome analysis, and all scans were used for the machine learning analysis to ensure applicability in real-world scenarios. Multivariate adjustment was performed for the following demographic factors and comorbidities: age, sex, presence of arterial hypertension, diabetes, coronary artery disease, atrial fibrillation, liver disease, and cancer. $p \leq 0{\cdot}05$ was considered statistically significant. The software used in the analysis is listed in the appendix (p 14).

We adhered to the TRIPOD guidelines,[13] and to the Checklist for Artificial Intelligence in Medical Imaging.[14]

Further, a medical algorithmic audit[15] was performed to identify failure modes, risks, and to establish principles for the improvement, continuous monitoring, and the safe use of the AI system.

Role of the funding source

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

Results

24 935 patients were enrolled, of whom 8694 patients were excluded (appendix p 33), the remaining 16 241 patients were included in the study. The final study population was grouped into four cohorts according to the country of the recruiting centre: Austria (n=9176), UK (n=6763), Italy (n=200), and China (n=102; figure 1). Clinical parameters stratified according to the presence of a cardiac amyloidosis-suggestive cardiac uptake for the entire cohort are shown in the table and for the individual cohorts in the appendix (pp 15–18). The mean age of patients was 62·6 years (SD 16·0), 9493 (58%) of 16 241 patients were female, and 6748 (42%) were male. Of the 15 939 patients with available information, 10 730 (66%) had cancer and 14 398 (89%) patients were referred for non-cardiac indications (appendix pp 15–16). Overall, 433 (3%) of 16 241 patients had a positive initial scan as per visual assessment, with a prevalence of 142 (2%) of 9176 patients in the Austrian cohort, 125 (2%) of 6763 patients in the UK cohort, 63 (62%) of 102 patients in the Chinese cohort, and 103 (52%) of 200 patients in the Italian cohort. Data from endomyocardial tissue biopsies (for the Austrian cohort) are in the appendix (p 19).

In the ten-fold stratified cross-validation scheme for the Austrian cohort, the prediction system achieved an AUC of 1·000 (95% CI 1·000–1·000), accuracy of 0·999 (0·998–0·999), sensitivity of 0·958 (0·920–0·997), specificity of 0·999 (0·999–1·000), positive predictive value of 0·951 (0·923– 0·986), and negative predictive value of 0·999 (0·999–1·000). Performances per fold are in the appendix (p 20). External validation yielded comparable performance between the UK and Italian cohorts, with AUCs of 0·997 (0·993–0·999) for the UK cohort and 1·000 (0·999–1·000) for the Italian cohort, and the performance in the Chinese cohort was slightly lower with an AUC of 0·925 (0·871–0·971; figure 2). A comprehensive comparison of diagnostic performance metrics for each cohort is in the appendix (p 21). Among the five centres included in the UK cohort, the performance per centre was consistently high with AUCs ranging from 0·988 to 1·000; AUCs for the two Chinese centres were 0·915 and 0·933.

Performance of subgroups (demographic factors, scanner, tracer, and centre) was consistent with the overall performance (appendix pp 22, 34). 64 (0·3%) of 19 401 predictions were incorrect. The mean AUC across all subgroups was 0·981 (95% CI 0·972–0·990).

	Entire cohort (n=16 241)	Non-cardiac amyloidosis-suggestive (Perugini grade 0 or 1; n=15 808)	Cardiac amyloidosis-suggestive (Perugini grade 2 or 3; n=433)	p value
Age, years	62·6 (16·0)	62·2 (15·9)	77·4 (10·6)	<0·0001
Sex				
Male	6748 (42%)	6409 (41%)	339 (78%)	<0·0001
Female	9493 (58%)	9399 (59%)	94 (22%)	..
Perugini grade				
0	15 498 (95%)	15 498 (98%)	0	<0·0001
1	310 (2%)	310 (2%)	0	..
2	243 (2%)	0	243 (56%)	..
3	190 (1%)	0	190 (44%)	..
Tracer				
DPD	9945 (61%)	9688 (61%)	257 (59%)	<0·0001
HMDP	200 (1%)	97 (1%)	103 (24%)	..
MDP	5994 (37%)	5984 (38%)	10 (2%)	..
PYP	102 (1%)	39 (<1%)	63 (15%)	..
Cohort				
Austria	9176 (56%)	9034 (57%)	142 (33%)	<0·0001
UK	6763 (42%)	6638 (42%)	125 (29%)	..
China	102 (1%)	39 (<1%)	63 (15%)	..
Italy	200 (1%)	97 (1%)	103 (24%)	..
Hypertension				
No	12 160 (75%)	11 961 (76%)	199 (46%)	0·51
Yes	3779 (23%)	3711 (23%)	68 (16%)	..
NA	302 (2%)	136 (1%)	166 (38%)	..
Diabetes				
No	13 537 (83%)	13 310 (84%)	227 (52%)	1·00
Yes	2402 (15%)	2362 (15%)	40 (9%)	..
NA	302 (2%)	136 (1%)	166 (38%)	..
Atrial fibrillation				
No	14 256 (88%)	14 117 (89%)	139 (32%)	<0·0001
Yes	1683 (10%)	1555 (10%)	128 (30%)	..
NA	302 (2%)	136 (1%)	166 (38%)	..
Liver disease				
No	14 955 (92%)	14 696 (93%)	259 (60%)	0·028
Yes	984 (6%)	976 (6%)	8 (2%)	..
NA	302 (2%)	136 (1%)	166 (38%)	..
Cancer				
No	5209 (32%)	5018 (32%)	191 (44%)	<0·0001
Yes	10 730 (66%)	10 654 (67%)	76 (18%)	..
NA	302 (2%)	136 (1%)	166 (38%)	..
Coronary artery disease				
No	13 822 (85%)	13 636 (86%)	186 (43%)	<0·0001
Yes	2117 (13%)	2036 (13%)	81 (19%)	..
NA	302 (2%)	136 (1%)	166 (38%)	..
Cardiology referral				
No	14 398 (89%)	14 316 (91%)	82 (19%)	<0·0001
Yes	1643 (10%)	1395 (9%)	248 (57%)	..
NA	200 (1%)	97 (1%)	103 (24%)	..

Data are mean (SD) or n (%). Perugini grade annotation as per fist scan. DPD=3,3-diphosphono-1,2-propanodicarboxylic acid. HMDP=hydroxymethylene diphosphonate. MDP=methylene diphosphonate. PYP=pyrophosphate. NA=not available.

Table: **Patient characteristics for the overall cohort**

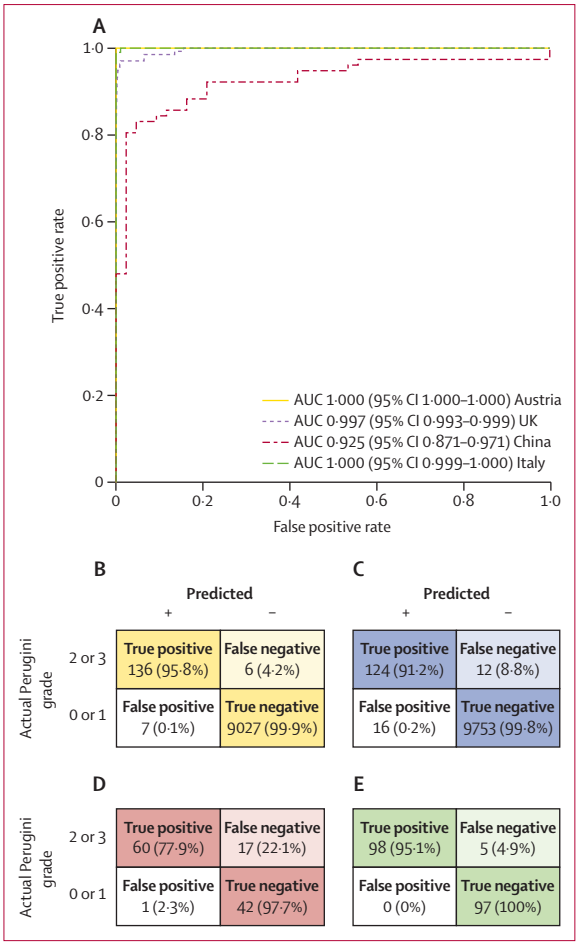

Figure 2: **Performance of the AI system in the four cohorts**
(A) Receiver operating characteristic curve representing the AI system's performance for the four cohorts. Confusion matrices are shown for the Austrian (B), UK (C), Chinese (D), and Italian (E) cohorts. Plus signs indicate a positive cardiac amyloid marker (>0·5) as predicted by the AI system, negative signs indicate a negative cardiac amyloid marker (≤0·5). AI=artificial intelligence.

Subgroups with lower AUC (<0·960) were consistently associated with the Chinese cohort. The subgroup with the lowest performance was related to Xiangya Second Hospital, China (AUC 0·915). Each sample was assigned to one failure mode. The most common failure mode resulting in false positives was determined to be extracardiac uptake (eg, unilateral left-sided breast uptake; figure 3C), accounting for 16 (67%) of 24 false positives across all cohorts. The most common failure mode among false negative predictions was low image quality or injection artifacts, which occurred in 17 (43%) of 40 false negative predictions. The number of instances for each failure mode and examples for each failure mode are in the appendix (pp 23, 35). Class activation maps for positive predictions were correctly focusing on the heart region (figure 3B).

As a consequence of the insights gained via the medical algorithmic audit, an AI-based algorithm detecting potential false positives was developed, implemented, and trained on the Austrian cohort data (appendix p 36). This safety mechanism was then tested on the UK cohort where it reduced the number of false positives by seven (44%) of 16 and led to an improvement of the positive predictive value from 0·886 to 0·932. No additional false negatives were introduced. Because these safety measures were developed and applied post-hoc, the resulting improvements were not included in any performance measurements.

The five observers disagreed in 22 (11%) of 200 of cases. Considering the task, the inter-observer variability was mediocre with a Fleiss' kappa of 0·89. The clinical expert raters achieved a mean AUC of 0·946 (95% CI 0·924–0·967), a mean sensitivity of 0·893 (0·851–0·935), and a mean specificity of 0·998 (0·994–1·000; appendix p 24). The AI system's performance was superior compared with the clinical rater's performance, with an AUC of 0·997 (p=0·0040), a sensitivity of 0·931, and a specificity of 1·000 (appendix p 37).

After a median follow-up of 4·6 years (IQR 1·4–5·6) after scintigraphy, 4101 (26%) of the 15 939 patients of the Austrian and UK cohort had died. A positive CAM, as predicted by the AI system, was significantly associated with mortality in the overall cohort (crude hazard ratio [HR] 1·98 [95% CI 1·67–2·34]; log-rank p<0·0001; figure 4). This prognostic effect was similar for the Austrian (1·77 [1·32–2·37]; log-rank p<0·0001) and UK cohorts (2·34 [1·84–2·97]; log-rank p<0·0001). The AI system's prediction remained significantly associated with mortality after multivariate adjustment for important demographic confounders and comorbidities in the overall (adjusted HR [adjHR] 1·44 [95% CI 1·19–1·74]; appendix p 25), Austrian (1·38 [1·02–1·87]; appendix p 26), and UK cohorts (1·52 [1·19–1·94]; appendix p 27). CAM positivity was prognostic across all patients with scans acquired using ^{99m}Tc-DPD (crude HR 1·63 [1·30–2·03]) and ^{99m}Tc-MDP (2·11 [1·05–4·23]; appendix p 38). Data on heart failure hospitalisations were only available for the Austrian cohort, whereby 150 (2%) of 9176 patients had an event. A positive CAM was a significant predictor both in univariate analysis (crude HR 17·52 [11·05–27·76]; log-rank p<0·0001) and after multivariate adjustment (adjHR 3·16 [1·90–5·25]; p<0·0001; appendix pp 28, 39).

Discussion

In this multinational, multicentre, multiscanner, cross-tracer study, we developed and validated a novel AI-based prediction system for the automated and standardised identification of patients with cardiac amyloidosis-suggestive uptake on scintigraphy. We show that intense cardiac uptake can be automatically and reliably detected across all tracers used in the identification of cardiac amyloidosis. The prediction system also conveyed significant prognostic information. We conclude that AI-based screening for cardiac amyloidosis-suggestive uptake

among patients referred for whole-body scintigraphy with bone-avid tracers might represent a valuable tool in identifying cardiac amyloidosis and streamlining management pathways with the ultimate goal of reducing mortality by early initiation of disease-modifying amyloid-specific treatment.

Scintigraphy using ^{99m}Tc-labelled tracers accounts for more than 1·4 million examinations in the USA every year and is performed predominantly for oncological, rheumatological, and lately increasingly for cardiological indications.[16] The prevalence of ATTR-related cardiac amyloidosis in the general population is estimated to be approximately 1·5%, with approximately 3% of patients older than 80 years affected who are referred to scintigraphy for non-cardiac reasons.[4] Assuming correct recognition, this would yield a total of approximately 21 000 newly diagnosed cases of ATTR-associated cardiac amyloidosis annually for the USA only. Given the availability of new disease-modifying therapies, there is an urgent need for reliable and early disease recognition, and for activation of efficient management pathways to meet the high socioeconomic disease burden.

Accurate, objective, and standardised detection of cardiac amyloidosis-associated uptake might be achieved using AI, which has been demonstrated by a 2023 study by Delbarre and colleagues.[7] Comparable to our study, Delbarre and colleagues developed a detection model for high-grade uptake on ^{99m}Tc-labelled planar bone scintigraphy (^{99m}Tc-HMDP and ^{99m}Tc-DPD). Scans were retrospectively selected, whereas our approach aimed to imitate real-life conditions by including consecutive all-comer cohorts referred for scintigraphy. Hence, the prevalence of positives was much lower in our study than in Delbarre and colleagues' study (2·7% *vs* 8·1%). These differences in study design are essential because they have a direct impact on the respective system performance. For instance, lower than tested disease prevalence will yield lower positive predictive values when implementing the AI system in a real-life setting. Also, grade 1 cardiac uptake is a frequent cause of false positive predictions and was more prevalent with a consecutive inclusion approach in our study (2% *vs* <1%). However, despite so-called false positivity as per definition of high-grade uptake, patients with subtle cardiac tracer accumulation require careful consideration, because this finding might represent either benign residual blood pool activity, ATTR-associated cardiac amyloidosis, or AL-associated cardiac amyloidosis—conditions which require entirely different management strategies.[4,8] Hence, a positive prediction in case of subtle cardiac tracer uptake should always instigate confirmation of myocardial tracer origin by single-photon-emission computed tomography (SPECT)-CT.

With the aim to maximise the clinical applicability of our system, we tested its performance across different subgroups, countries, scanners, and tracers which can identify patients with cardiac amyloidosis. The excellent performance results were consistent across studied subgroups, countries, and tracers, with only PYP scans performed in China showing slightly lower AUC values compared with other cohorts. This finding might be explained by differences in imaging protocols between PYP and other tracers. Most notably, the shorter post-injection times employed for PYP scans can lead to increased background uptake, reducing image quality.[17,18] Constant improvement and maintenance of a medical AI system is pivotal to ensure safety and optimise its performance and clinical use. With the use of a medical

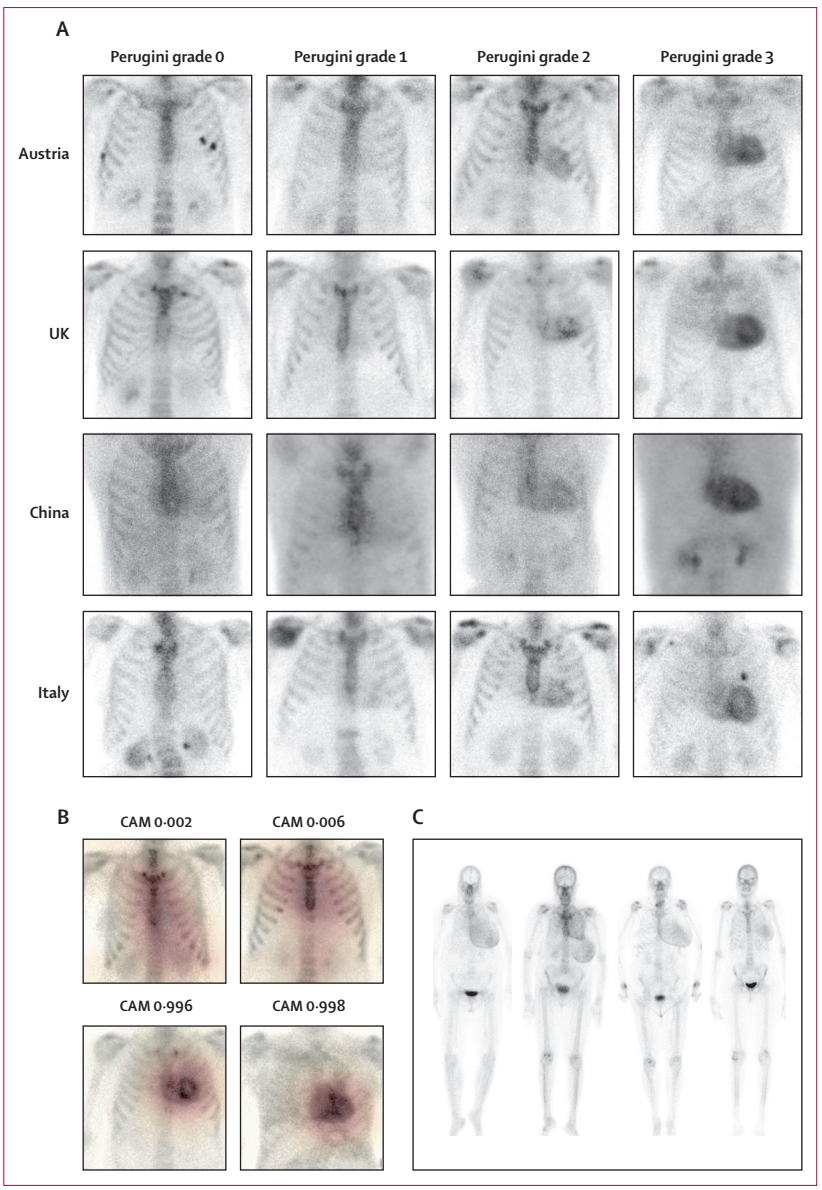

Figure 3: **Summary of audit results**
(A) Examples of thorax regions for each cohort. (B) Class activation maps indicating the spatial importance of image regions for a positive CAM prediction. The artificial intelligence system correctly focuses on the heart region. (C) Example failure modes resulting from increased tracer uptake in the left breast, leading to false positive predictions. CAM=cardiac amyloid marker.

algorithmic audit, we were able to reduce the number of false positives by seven (44%) of 16, further improving the system's positive predictive value. Our results further demonstrate the importance of reporting a wide range of performance metrics to allow for a thorough assessment of the system's predictive performance. For example,

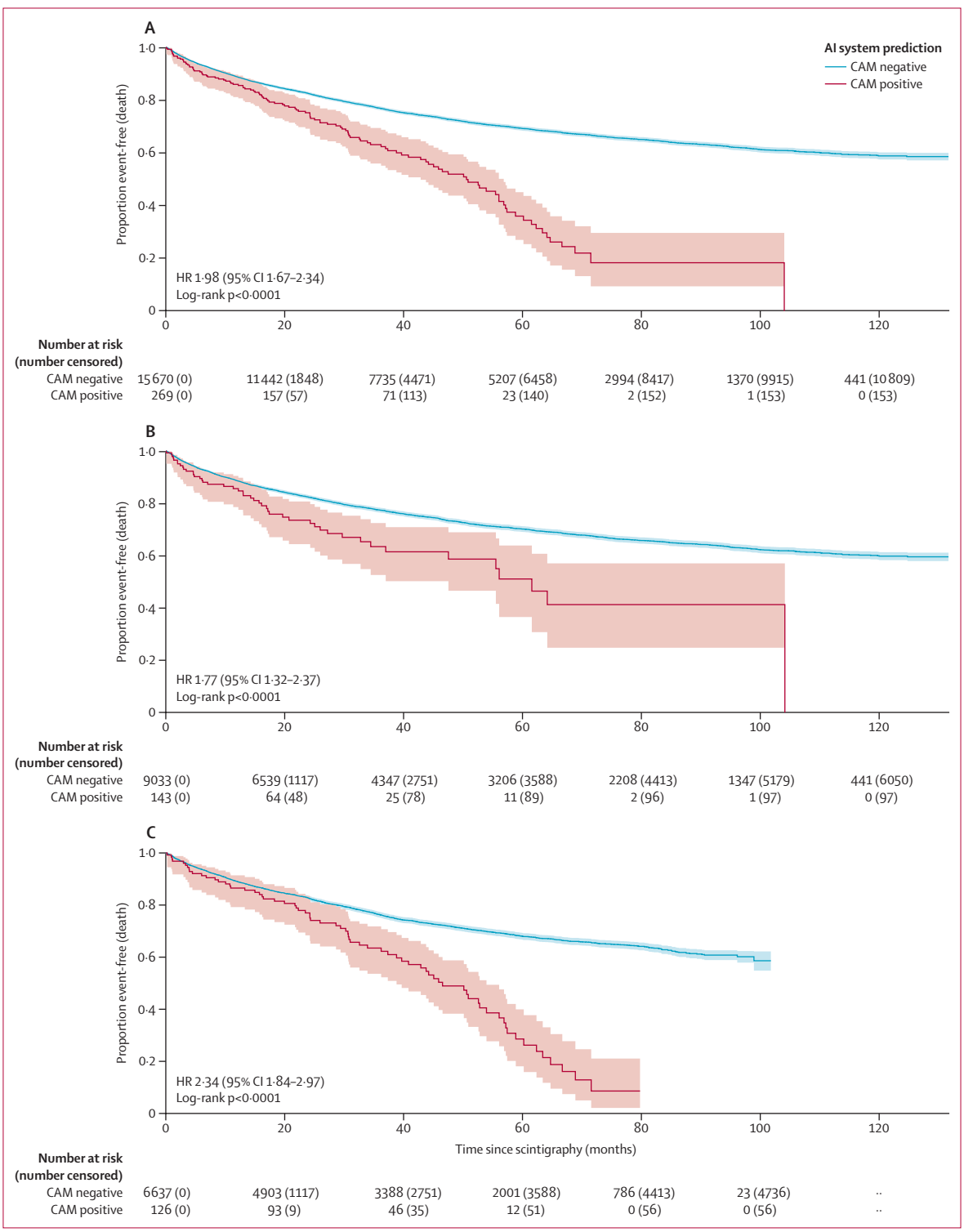

Figure 4: **Kaplan–Meier estimator for overall mortality in patients stratified based on the predicted cardiac amyloid marker**
(A) Overall cohort. (B) Austrian cohort. (C) UK cohort. AI=artificial intelligence. CAM=cardiac amyloid marker. HR=hazard ratio.

many studies on the predictive performance of diagnostic medical AI systems fail to report the positive predictive value or confusion matrices.[19,20] This selective choice of reported performance metrics might prevent the assessment of failure modes and implementation of appropriate risk mitigation strategies (as performed in the present study) and therefore reduce the quality of prediction systems in general.[15]

Importantly, our system also conveyed strong prognostic information. A positive cardiac amyloid marker, as predicted by the AI system, was independently associated with an increased mortality hazard in the overall cohort, and in the Austrian and UK cohorts separately. These results are remarkable considering the high prevalence of cancer in this all-comer cohort with associated high mortality rates. These findings highlight the progressive disease nature of cardiac amyloidosis, which can be tackled nowadays with already available and emerging treatment options.[5,21–23] In light of these new therapies, early disease recognition and efficient management pathways will become increasingly important patient care components, which might be effectively streamlined with the help of AI-based systems.

Embedding the developed system into the clinical workflow might be useful on different levels. First, incorporating the AI system as an in-line tool, which sits directly on the scanning software and has the ability for immediate image analysis during image acquisition, would enable automated real-time detection of cardiac amyloidosis-suggestive uptake and trigger further diagnostic steps, such as the necessity of performing additional SPECT-CT. Automated activation of management pathways such as notification of or referral to dedicated specialist centres could help reduce time to diagnosis and to treatment initiation. This time reduction could be of particular importance for outpatient imaging centres where cardiology expertise might be limited and referral pathways less well established. Second, the system can be applied for the retrospective screening of cohorts to identify patients requiring dedicated cardiac amyloidosis care or to recruit subjects for clinical studies. Due to the low prevalence of ATTR-associated cardiac amyloidosis, large amounts of imaging data must be screened in a time-consuming manual procedure to derive sufficiently large cohorts for expressive clinical studies. Using the automated prediction system, hundreds of patients can be assessed for ATTR-cardiac amyloidosis in a few seconds. Objectifying and standardising detection of patients with cardiac amyloidosis-suggestive uptake would also allow the screening of larger populations at high-risk, as recently recommended.[24–26]

Several limitations of the present study merit comment, including the relatively small and non-consecutively acquired cohort of ^{99m}Tc-PYP and ^{99m}Tc-HMDP scans. Still, the results derived from this data provide a proof-of-concept allowing for a first estimation of the AI system's applicability to screen for cardiac amyloidosis-suggestive uptake in ^{99m}Tc-PYP and ^{99m}Tc-HMDP scans. Further, the ground truth for this study was based on the clinical standard of visual image assessment. Despite the thorough annotation by multiple readers, occasional misjudgement cannot be ruled out. Ground truth based on histology from heart biopsies and SPECT-CT would be ideal, especially for the comparison of AI system and clinical reader performance. However, (1) retrieving biopsy results for a cohort of the present size is unrealistic, and (2) additional SPECT-CT of the chest is not routinely performed (especially in non-cardiac referrals), and the results of the present study therefore reflect real-life conditions, which makes it applicable to clinical routine. Additionally, grade-1 uptake was defined as negative in the present study. As detailed earlier, grade 1 might also represent cardiac amyloidosis, and future refinements of AI-based cardiac amyloidosis detection will therefore need to focus on differentiating benign causes (eg, residual blood pool activity) from cardiac amyloidosis-associated uptake. This differentiation will require planar scans coupled with SPECT-CT and cardiac biopsy to define true cardiac amyloidosis. Lastly, we did not fine-tune the threshold for the presented cardiac amyloid marker. However, it is feasible to adjust and optimise the threshold according to specific clinical requirements and individual situations, allowing for customisation of sensitivity and specificity.

In conclusion, this study presents the development and validation of a high-performance AI-based prediction system for the detection of cardiac amyloidosis-associated uptake on scintigraphy. Consistent performance across countries, scanners, tracers, and patient subgroups suggests broad clinical applicability. The AI system also predicted worse clinical outcomes, which highlights its potential to be implemented in the clinical workflow, where it could help reduce time to diagnosis and to initiation of disease-modifying therapies. Inclusion of consecutive all-comer referrals suggests strong performance when implementing the developed system in a real-life setting.

Contributors

CPS, DH, CN, and MH conceived the study and planned the experiments. CN, CPS, DH, RHD, IP, MZ, XM, TAT, XL, KP, LJM, RS, KM, AK, CH, JM, MS, and AM participated in gathering and cleaning the data. CN, MZ, XM, RC, TT-W, KK, ARH, RS, MS, AM, and JN performed data annotation. TN, AG, DA, TT-W, CPS, DH, and MH participated in the multicase multireader study. CPS and DH performed experiments and analysis, developed the associated software, and performed the validation. MW, CN, and CPS performed the statistical analysis. KK, CPS, and DH performed the medical algorithmic audit. CPS, DH, and CN designed the figures and wrote the manuscript. All authors reviewed the manuscript and provided critical feedback. CPS and CN have accessed and verified the data. CPS and CN are independent of the funding source. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

CN reports speaker fees or institutional research grants from Pfizer and advisory board honoraria from Prothena. TAT is co-founder and

shareholder of Myocardium AI. RHD has received payment for consultancy work and owns shares in Myocardium AI. All other authors declare no competing interests.

Data sharing

The image data used to train and test the artificial intelligence system are not shareable under the current agreements established for the means of this study. The model associated with this publication can be obtained by contacting the Technology Transfer Office of the Medical University of Vienna at https://www.meduniwien.ac.at/technologietransfer. Requests for access to de-identified intermediate data generated within this study can be made to the corresponding author.

Acknowledgments

The present study was funded by Pfizer (research grant number 68467373 awarded to CN).

Editorial note: The Lancet Group takes a neutral position with respect to territorial claims in published maps and institutional affiliations.

References

1 Quarta CC, Kruger JL, Falk RH. Cardiac amyloidosis. *Circulation* 2012; **126:** e178–82.
2 Gillmore JD, Maurer MS, Falk RH, et al. Nonbiopsy diagnosis of cardiac transthyretin amyloidosis. *Circulation* 2016; **133:** 2404–12.
3 Quarta CC, Obici L, Guidalotti PL, et al. High 99mTc-DPD myocardial uptake in a patient with apolipoprotein AI-related amyloidotic cardiomyopathy. *Amyloid* 2013; **20:** 48–51.
4 Nitsche C, Mascherbauer K, Calabretta R, et al. Prevalence and outcomes of cardiac amyloidosis in all-comer referrals for bone scintigraphy. *J Nucl Med* 2022; published online June 16. https://doi.org/10.2967/jnumed.122.264041.
5 Maurer MS, Schwartz JH, Gundapaneni B, et al. Tafamidis treatment for patients with transthyretin amyloid cardiomyopathy. *N Engl J Med* 2018; **379:** 1007–16.
6 Kastritis E, Palladini G, Minnema MC, et al. Daratumumab-based treatment for immunoglobulin light-chain amyloidosis. *N Engl J Med* 2021; **385:** 46–58.
7 Delbarre MA, Girardon F, Roquette L, et al. Deep learning on bone scintigraphy to detect abnormal cardiac uptake at risk of cardiac amyloidosis. *JACC Cardiovasc Imaging* 2023; **16:** 1085–95.
8 Nitsche C, Mascherbauer K, Wollenweber T, et al. The complexity of subtle cardiac tracer uptake on bone scintigraphy. *JACC Cardiovasc Imaging* 2022; **15:** 1516–18.
9 Dorbala S, Ando Y, Bokhari S, et al. ASNC/AHA/ASE/EANM/HFSA/ISA/SCMR/SNMMI expert consensus recommendations for multimodality imaging in cardiac amyloidosis: part 1 of 2-evidence base and standardized methods of imaging. *J Nucl Cardiol* 2019; **26:** 2065–123.
10 Van den Wyngaert T, Strobel K, Kampen WU, et al. The EANM practice guidelines for bone scintigraphy. *Eur J Nucl Med Mol Imaging* 2016; **43:** 1723–38.
11 Dorbala S, Ando Y, Bokhari S, et al. Correction to: ASNC/AHA/ASE/EANM/HFSA/ISA/SCMR/SNMMI expert consensus recommendations for multimodality imaging in cardiac amyloidosis: part 1 of 2-evidence base and standardized methods of imaging. *J Nucl Cardiol* 2021; **28:** 1761–62.
12 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3:** e745–50.
13 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Ann Intern Med* 2015; **162:** 735–36.
14 Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2020; **2:** e200029.
15 Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden-Rayner L. The medical algorithmic audit. *Lancet Digit Health* 2022; **4:** e384–97.
16 OECD, Nuclear Energy Agency. The supply of medical isotopes. Paris: OECD publishing, 2019.
17 Weiler-Sagie M, Ben-Haim S. Variability in bone-seeking tracers and imaging protocols for the diagnosis of cardiac amyloidosis: the more the merrier? *J Nucl Cardiol* 2022; **29:** 319–22.
18 Sutherland N, Sivaratnam D, Pearson M, Better N. Use of SPECT imaging to confirm ATTR cardiac amyloid in a Tc99m pyrophosphate study artefactually compromised by a recent V/Q scan. *J Nucl Cardiol* 2023; **30:** 2242–44.
19 Vasey B, Ursprung S, Beddoe B, et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. *JAMA Netw Open* 2021; **4:** e211276.
20 Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* 2021; **4:** 65.
21 Benson MD, Waddington-Cruz M, Berk JL, et al. Inotersen treatment for patients with hereditary transthyretin amyloidosis. *N Engl J Med* 2018; **379:** 22–31.
22 Adams D, Gonzalez-Duarte A, O'Riordan WD, et al. Patisiran, an RNAi therapeutic, for hereditary transthyretin amyloidosis. *N Engl J Med* 2018; **379:** 11–21.
23 Garcia-Pavia P, Aus dem Siepen F, Donal E, et al. Phase 1 trial of antibody NI006 for depletion of cardiac transthyretin amyloid. *N Engl J Med* 2023; **389:** 239–50.
24 Witteles RM, Bokhari S, Damy T, et al. Screening for transthyretin amyloid cardiomyopathy in everyday practice. *JACC Heart Fail* 2019; **7:** 709–16.
25 Nitsche C, Scully PR, Patel KP, et al. Prevalence and outcomes of concomitant aortic stenosis and cardiac amyloidosis. *J Am Coll Cardiol* 2021; **77:** 128–39.
26 Nitsche C. How prevalent is cardiac amyloidosis and which patients should be screened? *Eur J Heart Fail* 2022; **24:** 2352–54.