Original Article

Fully automated guideline-compliant diameter measurements of the thoracic aorta on ECG-gated CT angiography using deep learning

Maurice Pradella[1]^, Thomas Weikert[1]^, Jonathan I. Sperl[2]^, Rainer Kärgel[2]^, Joshy Cyriac[1]^, Rita Achermann[1], Alexander W. Sauter[1]^, Jens Bremerich[1]^, Bram Stieltjes[1]^, Philipp Brantner[1,3]^, Gregor Sommer[1]^

[1]Department of Radiology, Clinic of Radiology & Nuclear Medicine, University Hospital Basel, University of Basel, Petersgraben 4, 4031 Basel, Switzerland; [2]Siemens Healthineers, Siemensstraße 3, 91301 Forchheim, Germany; [3]Regional Hospitals Rheinfelden and Laufenburg, Riburgerstrasse 12, 4310 Rheinfelden, Switzerland

Contributions: (I) Conception and design: M Pradella, T Weikert, JI Sperl, AW Sauter, J Bremerich, B Stieltjes, P Brantner, G Sommer; (II) Administrative support: M Pradella, J Bremerich, B Stieltjes, P Brantner, G Sommer; (III) Provision of study material or patients: M Pradella, J Cyriac, JI Sperl, R Kärgel, J Bremerich, P Brantner, G Sommer; (IV) Collection and assembly of data: M Pradella, T Weikert, AW Sauter, J Bremerich, B Stieltjes, P Brantner, G Sommer; (V) Data analysis and interpretation: M Pradella, T Weikert, J Cyriac, R Achermann, AW Sauter, J Bremerich, B Stieltjes, P Brantner, G Sommer; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Maurice Pradella, MD. Department of Radiology, Clinic of Radiology & Nuclear Medicine, University Hospital Basel, University of Basel, Petersgraben 4, 4031 Basel, Switzerland. Email: maurice.pradella@usb.ch.

Background: Manually performed diameter measurements on ECG-gated CT-angiography (CTA) represent the gold standard for diagnosis of thoracic aortic dilatation. However, they are time-consuming and show high inter-reader variability. Therefore, we aimed to evaluate the accuracy of measurements of a deep learning-(DL)-algorithm in comparison to those of radiologists and evaluated measurement times (MT).
Methods: We retrospectively analyzed 405 ECG-gated CTA exams of 371 consecutive patients with suspected aortic dilatation between May 2010 and June 2019. The DL-algorithm prototype detected aortic landmarks (deep reinforcement learning) and segmented the lumen of the thoracic aorta (multi-layer convolutional neural network). It performed measurements according to AHA-guidelines and created visual outputs. Manual measurements were performed by radiologists using centerline technique. Human performance variability (HPV), MT and DL-performance were analyzed in a research setting using a linear mixed model based on 21 randomly selected, repeatedly measured cases. DL-algorithm results were then evaluated in a clinical setting using matched differences. If the differences were within 5 mm for all locations, the cases was regarded as coherent; if there was a discrepancy >5 mm at least at one location (incl. missing values), the case was completely reviewed.
Results: HPV ranged up to ±3.4 mm in repeated measurements under research conditions. In the clinical setting, 2,778/3,192 (87.0%) of DL-algorithm's measurements were coherent. Mean differences of paired measurements between DL-algorithm and radiologists at aortic sinus and ascending aorta were −0.45±5.52 and −0.02±3.36 mm. Detailed analysis revealed that measurements at the aortic root were over-/underestimated due to a tilted measurement plane. In total, calculated time saved by DL-algorithm was 3:10 minutes/case.

^ ORCID: Maurice Pradella, 0000-0003-2449-7835; Thomas Weikert, 0000-0001-9274-053X; Jonathan I. Sperl, 0000-0003-4528-1507; Rainer Kärgel, 0000-0003-2024-5390; Joshy Cyriac, 0000-0002-4584-0623; Alexander W. Sauter, 0000-0002-6707-2258; Jens Bremerich, 0000-0002-1002-8483; Bram Stieltjes, 0000-0002-5961-802X; Philipp Brantner, 0000-0003-3996-3966; Gregor Sommer, 0000-0002-8952-0808.

Conclusions: The DL-algorithm provided coherent results to radiologists at almost 90% of measurement locations, while the majority of discrepent cases were located at the aortic root. In summary, the DL-algorithm assisted radiologists in performing AHA-compliant measurements by saving 50% of time per case.

Keywords: Deep learning; aortic aneurysm; computed tomography angiography; dimensional measurement accuracy; observer variation; time management

Submitted Feb 09, 2021. Accepted for publication May 27, 2021.
doi: 10.21037/qims-21-142
View this article at: https://dx.doi.org/10.21037/qims-21-142

Introduction

Thoracic aortic dilatation occurs with an incidence of approximately 6–16 cases per 100,000 people/year and there is an increasing prevalence and incidence of dilatation of the thoracic aorta (1-4). Regardless of the cause of dilatation, the risk of aortic dissection or rupture rises with increasing diameters (5). This leads to high mortality rates. For example, in the USA, aneurysms of the thoracic and abdominal aorta are the 14th leading cause of death in people older than 55 years (6). Main factors that cause an increase in aortic diameter are patient age, genetic disorders (such as Marfan syndrome), as well as valve pathologies such as a bicuspid aortic valve (7,8).

Imaging is the sole option to detect aortic dilatation, being typically an asymptomatic disease, and only cross-sectional imaging can depict the entire aortic arch; opposite to echocardiography which can only be used to visualize the aortic root. Current guidelines recommend ECG-gated CT angiography (CTA) which is considered superior to other imaging modalities (9). However, measurements differ frequently. It is well known that transverse diameter measurements are inaccurate and considered obsolete (10). Centerline-based measurements have become best practice and were established about 15 years ago (11). However, the process of evaluating the dimensions of the thoracic aorta by measurements perpendicular to the vessel centerline is still time-consuming with 5–6 minutes per case (12,13). Currently, centerline fitting is performed automatically, but measurement locations have to be chosen manually. Due to incorrectly placed centerlines or failed automatic fitting, there is often the necessity for manual adjustments/interaction (13). This increases measuring times further and is a source of variability which ranges up to 5 mm even among expert readers in a research setting (14,15).

There are limited studies describing tools for automatic aortic segmentation/measurements that, for example, detect abdominal aortic aneurysms, measure the descending aortic diameter prior to stent graft planning, segment and measure aortic diameters in native scans of the thoracic aorta in CT scans, or help to improve reading follow-up CT scans according to guidelines (16-19).

In this work, we analyzed the performance of a novel DL-algorithm that automatically detects the thoracic aorta, places the centerline, identifies measurement locations and performs measurements according to current American Heart Association (AHA) guidelines.

The accuracy of the DL-algorithm was analyzed in a patient cohort with suspected aortic dilatation. In a research setting, we first compared its measurements to radiologists' measurements who used the established semi-automatic procedure in order to evaluate inter- and intra-reader variability and the expected savings in terms of measurement time (*MT*). This was followed by an evaluation of the whole cohort in a clinical setting.

Methods

Ethics

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). All data was encoded prior to any analysis to preserve patient anonymity. The Ethics Commitee for Northwest and Central Switzerland approved this study (ID: 2019-01053) and individual consent for this retrospective analysis was waived.

Study population

A total of 371 consecutive patients who underwent ECG-gated CTA at our institution between May 2010 and June 2019 and whose radiologic reports included standardized diameter measurements were identified and included in this study (*Figure 1*). Those patients either were suspected to

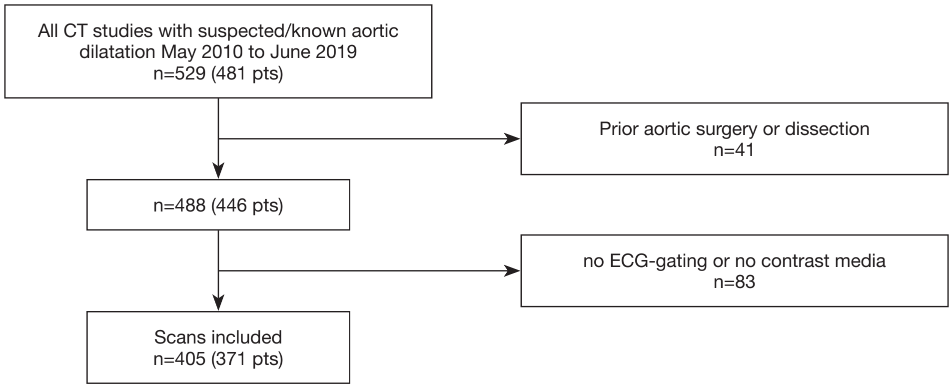

Figure 1 Flow chart of the dataset A. All scans with suspected or known dilatation and reported, standardized measurements were included. Pts, patients; DL, deep learning.

Table 1 Baseline characteristics

Baseline data	Dataset A	Dataset B (inter-rater subset)	Dataset C (follow-up subset)
Number of patients	371	21	32
Age (years)	65.2±11.7	67.4±14.0	61.1±11.1
Female sex	98 (24.6%)	5 (23.8%)	5 (15.6%)
Number of CT scans	405	21	34

Baseline characteristics and study populations for both datasets (Dataset A: main cohort, Dataset B: randomly selected cases for inter-/intra-reader analysis, Dataset C: patients with more than one exam). Dataset C, age at first scan. CT, computed tomography

have dilatation (for example an aortic root diameter based on echocardiography of more than 40 mm) or underwent CT exams in the context of known dilatation. Exclusion criteria were aortic pathologies other than dilatation (acute or chronic dissection, rupture or intramural hematoma) or prior surgery of the thoracic aorta; the DL-algorithm was not built to evaluate those conditions. Baseline data of patients and overview of processed cases are shown in *Table 1*. While the full dataset (dataset A) was used to evaluate the overall diagnostic performance of the DL-algorithm, a subset of 21 CT studies (dataset B, inter-rater subset) was randomly selected to perform an analysis under research conditions. Thereby, the inter- and intra-observer viariability associated with the common established, semi-automatic workflow was analyzed. Another subset (dataset C, follow-up subset) was created to evaluate the subcohort of patients who underwent more than one exam.

CT scan

All scans were performed on one of four CT scanners (Somatom Sensation 64, AS+, Edge or Definition Flash, Siemens Healthineers, Forchheim, Germany). Each exam was prospectively ECG-gated to minimize motion artifacts by cardiac movement. Image acquisition was performed during diastole, either as low pitch spiral acquisition with dose modulation over multiple heart beats (Sensation 64, AS+ or Edge scanners) or during one heart beat (Definition Flash scanner).

Bolus tracking was performed in the ascending aorta; trigger was ≥100 HU with a 10 s delay. 70–100 mL of contrast agent for thoracic scans were administered with a flow rate of 3–4 mL/s. No pharmacologic agent was used for heart rate control in any scan. We generally used the thinnest soft tissue kernel available (1.0 mm slice thickness, increment 0.6 mm, resolution 512×512 pixels).

Measurement tools

Established semi-automatic workflow

Measurements were performed perpendicular to blood flow axis using the centerline technique in the postprocessing software Syngo.via (Siemens Healthineers, Forchheim, Germany) (10,14). The aortic centerline was automatically detected; radiologists could adjust the centerline in case it was not fitted well. If the automatic centerline wasn't available, radiologists placed it manually. Measuring points were used according to current AHA-guidelines (9): aortic sinus (AS), sinotubular junction (STJ), ascending aorta (AA), proximal aortic arch (PA), mid aortic arch (MA), and distal aortic arch (DA) (9).

Fully-automatic DL-based workflow

DL-algorithm measurements were performed by an in-house deployed prototype software (Chest AI, version 0.2.9.2, Siemens Healthineers, Forchheim, Germany). Its development was completely independent from this study. The thinnest soft tissue kernel series per case was sent to the dedicated workstation which processed the cases one by one. No further human input was necessary.

The DL-algorithm fully and automatically performed three consecutive steps: detection of aortic landmarks, segmentation of the lumen, and diameter measurements (incl. detection of measurement locations).

First, landmark detection based on Deep Reinforcement Learning was performed to detect six landmarks along the thoracic aorta: aortic root, aortic arch center, brachiocephalic artery bifurcation, left common carotid artery, left subclavian artery, celiac trunk. The principles of the underlying algorithm have been described by Ghesu *et al.* (20). The algorithm has been trained on more than 10,000 data sets (CT data plus manual labelling of the six landmarks).

The aortic root landmark was used to define a Region of Interest (ROI) for the segmentation algorithm. The segmentation was performed using an adversarial deep image-to-image network (DI2IN), which is a multi-layer convolutional neural network (CNN) taking the CT data (cropped to the ROI) as an input and providing a segmentation mask as an output. The technical approach of network topology and training strategy has first been developed in the context of liver segmentation and is easily adapted to other organs like the aorta by providing corresponding CT data and annotations (manually segmented aorta masks) (21). Training was performed on more than 1,000 CT data sets covering both native and contrast-enhanced data with and without ECG-gating; these data sets were completely independent from this study.

Given the segmented aorta mask, a centerline model was used to generate the aortic centerline. The centerline was used in combination with the pre-computed aortic landmarks to identify the measurement planes at multiple locations according to the AHA guidelines (*Figure 2*) (9).

In each of the planes, multiple diameters were measured by computing intersections of rays starting from the centerline with the aortic mask. Based on these diameters, the maximum in-plane diameter was reported. Visual output series were created in axial and sagittal orientation as well as a 3D volume rendering.

Image reading and data evaluation

Research setting, inter-rater variability of semi-automatic workflow compared to DL-algorithm and measurement times

Three readers, R1, R2, and R3 with 2, 4.5 and 8 years of experience, respectively, performed measurements in dataset B (inter-rater subset) twice with a blanking period of at least 7 days. Readers were blinded to reports. Reader R2 and R3 were both fellowship-trained in cardiovascular radiology. The reading was performed in a calm environment, no telephone or clinical duties were present to establish optimal conditions for measurements. Each reader noted MT of each case after it was loaded in Syngo.via software up until all locations were measured.

Intraclass correlations (ICC) were calculated for both intra- and inter-reader agreement evaluation for each measurement location (22). To compare the the results of the DL-algorithm with human performance and variability, we set up a linear mixed model based on the human measurements with reader and case as random effects and location as a fixed effect (23). Then, a predicted value (gold standard) and its 95% prediction interval for each location and case was estimated in order to evaluate human performance variability (HPV). To obtain robust prediction intervals we applied bootstrap methods taking into account the hierarchical structure of the data (R:library fabricatr). Finally, the proportion of DL-measurements outside the prediction interval was used to test (chi-squared) whether the proportion of outliers is compatible with the expected number of 5% (gold standard).

Clinical setting, performance evaluation of the DL-algorithm

All 405 scans included in dataset A underwent fully automatic processing by the DL-algorithm. Each case was processed twice to evaluate technical feasibility; to verify reproducibility, those measurements were compared with

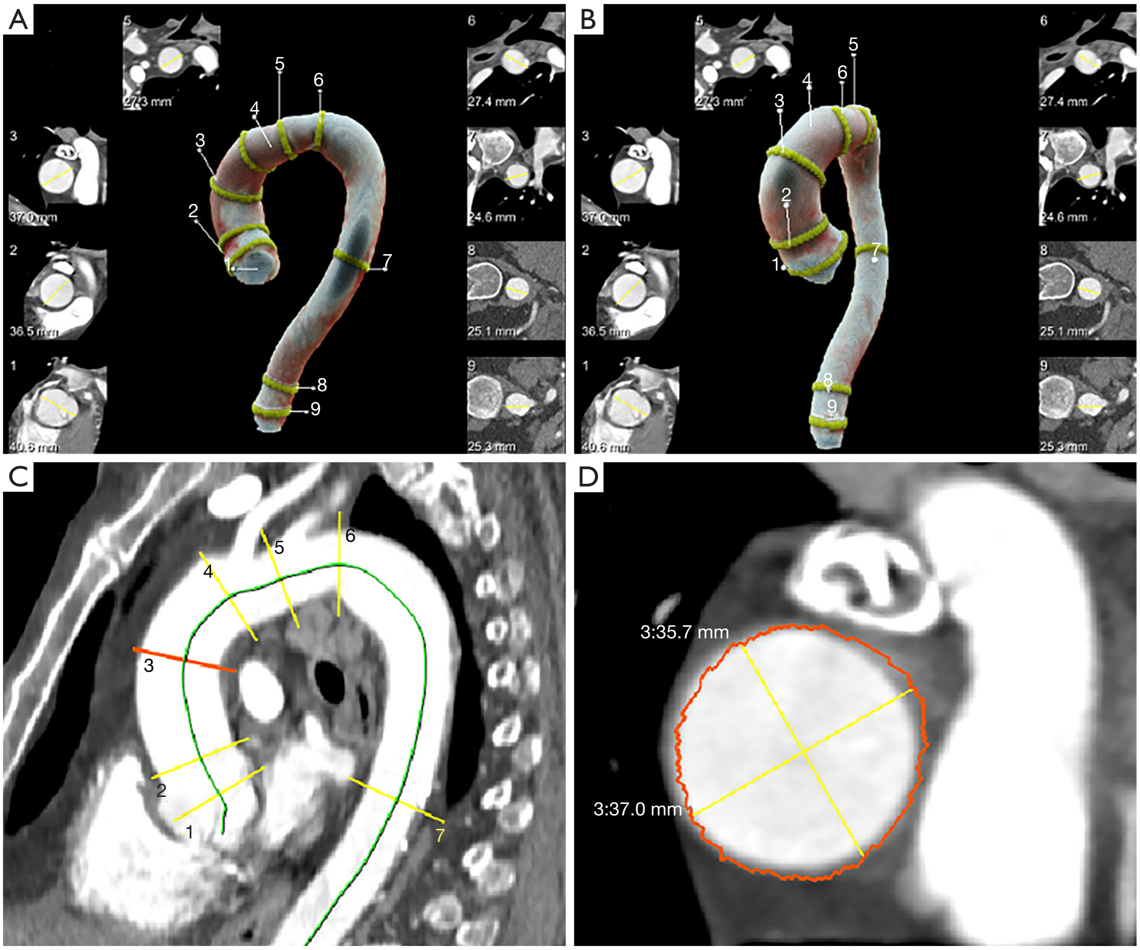

Figure 2 Visual output of DL-measurements in a case with normal diameters and within human variance. (A) VRT with lateral view on thoracic aorta with thumbnails of measurements at each location; (B) VRT with anterior view on thoracic aorta; (C) Non-linear projection of the aortic centerline into a 2D plane on thoracic aorta with measurement plane at each location, measurement plane of ascending aorta highlighted in orange; (D) Measurement of AA on cross-sectional images orthogonal to the aortic centerline based on (C). AA, ascending aorta; AHA, American Heart Association; DL, deep learning; VRT, volume rendering technique.

each other. MT required for processing were automatically noted by the DL-algorithm.

The results provided by the DL-algorithm were then compared to the original diameter measurements that were retrospectively extracted from the written reports using a Python-based script. These original measurements were initially performed by residents, afterwards they were discussed with a senior, board-certified radiologist who was free to overrule measurements and who finalized the report. Mid descending aorta (MDA) and distal descending aorta (DDA) measurements were not included in the original reports as a trade-off to optimize clinical efficiency since most dilatations are found at the aortic root or ascending aorta.

Cases with a difference of >5 mm for at least one measurement location between the two methods were regarded as discrepant (cutoff-value based on Quint *et al.* (15), this includes missing measurements). In these cases, visual outputs were used for a full review by a fellowship-trained, cardiovascular radiologist with 4.5 years of experience (R2). In addition, an analysis of classification change (dilatation versus no dilatation) between DL-measurements and original reports was performed. We defined relevant dilatation of the aorta as ≥45 mm at AS, STJ, and AA and ≥40 mm at all other locations in order to evaluate misclassification (based on current literature which

Table 2 Technical success rates

Technical success rates	Dataset A (371 pts/405 cases)
Number of cases processed by the DL-algorithm	399/405 (98.5%)
Number of processed cases by the DL-algorithm with all measurements available	341/399 (85.5%)
Aortic sinus	399/399 (100%)
Sinotubular junction	399/399 (100%)
Ascending aorta	396/399 (99.2%)
Proximal arch	396/399 (99.2%)
Mid arch	394/399 (98.7%)
Distal arch	341/399 (85.5%)
Mid descending aorta	341/399 (85.5%)
Distal descending aorta	341/399 (85.5%)
Total technical success rate of DL-algorithm for all locations	3,007/3,192 (94.2%)

Technical success rates of the DL-algorithm divided by location. DL, deep learning.

also represents the standard at our institution) (2,24). For review analysis, AS and STJ were grouped as aortic root, PA, MA, and DA as aortic arch, and MDA and DDA as descending aorta.

Dataset C which included all patients with more than one exam was analyzed in regards if there was a difference of >5 mm of diameters between two scans for the DL-algorithm or the reports. The results can be found in the supplements.

Data organization was performed with Excel (Microsoft Corporation, Redmont, USA) and Python (Python Software Foundation, Wilmington, USA). R (R Foundation for Statistical Computing, Vienna, Austria) and SPSS (IBM, Armonk, USA) were used for statistical analysis. We plotted the data in scatterplots and calculated Pearson correlation coefficients (PCC) for each location. In addition, Bland-Altman plots were created to compare reported with DL-measurements. To compare absolute diameters, Mann-Whitney-U-Test was used. A P value <0.05 was defined to indicate statistical significance. We also calculated mean diameter measurements and standard deviations by the DL-algorithm for each location, sorted by sex and age group (Table S1).

Results

Success rates of automatic processing

Fully-automated diameter measurements by the DL-algorithm were technically successful in 399 of 405 cases (98.5%) and at 3007/3192 locations (94.2%, *Table 2*). Measurements from AS until mid arch were available in 98.7%, complete measurements at all locations in 85.5% of all cases. The algorithm's technical failure rate was highest in the descending aorta and distal aortic arch (14.5%).

Inter- & intra-reader comparison in research setting

The randomly selected dataset B (inter-rater subset) used for the inter- and intra-reader analyses consisted of 12 cases with normal diameters, 8 with aortic dilatation and one with aortic coarctation. *Figure 3* shows the manual measurement results for all 21 patients from the subset along with the respective DL-measurements.

Overall, inter-rater agreement between radiologists was excellent: Average ICC for intra-reader agreement over all locations was 0.94 [range: 0.81 (STJ) – 0.98 (AA)]. The largest variation was observed for STJ measurements of R1 and R3 (ICC: 0.76 and 0.70, respectively) indicating only moderate/good agreement. Average ICC for inter-reader agreement over all locations was 0.94 [range: 0.85 (STJ) – 0.98 (AA)]. An overview of all ICC can be found in the supplements (Table S2).

The prediction interval for HPV that was calculated based on the repeated measurements by the three readers varied for each location with a median width of ±2.6 mm, a maximum width of ±3.4 mm, and a minimum width of ±2.5 mm.

The DL-algorithm measurements were statistically more

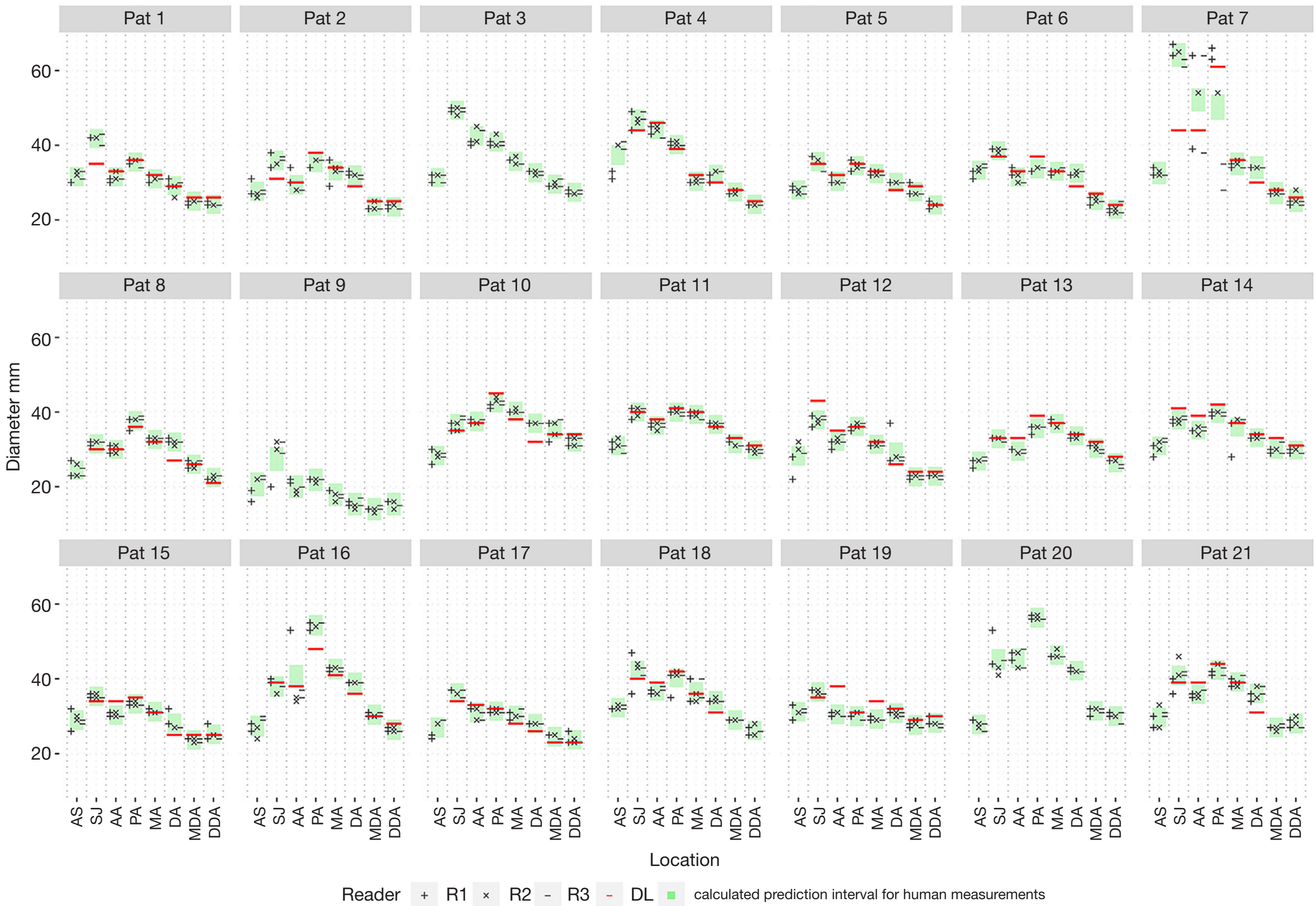

Figure 3 Inter-reader comparison for all 21 patients. Each plot represents all measurements for one patient for the locations from AS to DDA on the x-axis. For each location, the measurements by the three readers (symbols "+", "x" and "–"), the DL-algorithm (red line) and the predicted gold standard interval (green box) whose calculation was based on measurements by readers 1–3 (R1–3) are shown. Note: Since there was no variance in DL-measurements (which represents 100% preciseness), the red line for DL-measurements represents the two measurements performed per location. AS, aortic sinus; SJ, sinotubular junction; AA, ascending aorta; PA, proximal arch; MA, Mid arch; DA, distal arch; MDA, mid descending aorta; DDA, distal descending aorta; DL, deep learning; Pat, patient number.

often outside the 95% prediction interval compared to the expected percentage of 5% as shown by the chi-squared test (22.5%, P<0.0001, 95% CI: 15.99–30.51).

Evaluation of DL-algorithm measurement and classification accuracy in clinical setting

The accuracy analysis of dataset A showed that the automated measurements in 2,540/3,192 locations (79.6%, *Table 3*) differed from the human measurements by less than a 5 mm interval and therefore, were counted as coherent. 145/399 cases (36.3%) showed a difference of >5 mm for at least one location (this also includes cases with missing measurements). After a detailed review of all measurements for these 145 cases, 2,778/3,192 (87.0%) measurements were identified as coherent. Aside from the aortic root (537/798, 67.3%), the ascending aorta (364/399, 91.2%), aortic arch (1,123/1,197, 94.2%), and descending aorta (754/798, 94.5%) showed high rates of coherent measurements. In the majority of reviewed cases, the estimation error found at the aortic root was due to a tilted measurement plane (*Figure 4*). Of all reviewed cases, classification of dilatation remained unchanged in 76/145 cases (52.4%) while a change of classification occurred in 69/145 cases (47.6%, *Table 3*).

Table 3 Measurement and classification accuracy by location

Dataset A (399 cases/3,192 locations)	All locations	Root	Ascending aorta	Arch	Descending aorta
Cases with initially correct estimation (within 5 mm interval for all measurements)	254/399 (63.7%)				NA§
Reviewed cases†	145/399 (36.3%)				
Reviewed locations†	1,160				
Correct measurement	748/1,160 (64.5%)	24/290 (8.3%)	110/145 (75.9%)	361/435 (83.0%)	246/290 (84.8%)
Wrong measurement	377/1,160 (32.5%)	266/290 (90.3%)	32/145 (22.1%)	69/435 (15.9%)	14/290 (4.8%)
Missing measurement	35/1,160 (3.0%)	0/290	3/145 (2.0%)	5/435 (1.1%)	30/290 (10.3%)
No change of classification	76/145 (52.4%)	96/145 (66.2%)	134/145 (92.4%)	130/145 (89.7%)	NA
Change of classification	69/145 (47.6%)	49/145 (33.8%)	11/145 (7.6%)	15/145 (10.3%)	NA
Aneurysms misclassified by DL-algorithm in reviewed cases	34/145‡ (23.4%)	18/145 (12.4%)	11/145 (7.6%)	11/145 (7.6%)	NA
Finally correct measurements (incl. reviewed locations)	2,778/3,192 (87.0%)	537/798 (67.3%)	364/399 (91.2%)	1,123/1,197 (93.8%)	754/798 (94.5%)
Aneurysms misclassified by DL-algorithm (all cases)	34/399‡ (8.5%)	18/399 (4.5%)	11/399 (2.8%)	11/399 (2.8%)	NA

Measurement and classification accuracy by location. †, cases with a difference >5 mm between DL and original measurements underwent detailed review of all measurements. ‡, in 6 cases the misclassified aneurysm extended to multiple locations, those cases were only counted once. §, not available in original reports. DL, deep learning; NA, not available.

An aneurysm was misclassified in 34/399 cases (8.5%). An overview of measured diameters sorted by sex and age groups can be found in the supplement (Table S1).

Mean differences of matched measurements by the algorithm and the report were −0.45 mm at AS and −0.02 mm at AA. Bland-Altman analysis revealed wider limits of agreement (±1.96 SD) at AS than at AA (AS: +10.37 and −11.28 mm, AA: +6.56 and −6.61 mm). PCC were 0.676 for AS (moderate correlation) and 0.906 for AA (high correlation) (*Figure 5*). Mean differences of matched measurements at STJ, PA, MA and DA were +3.25 mm, +0.32 mm, −1.21 mm, and +1.53 mm, respectively. The Bland-Altman and scatterplots for these locations can be found in the supplement (Figures S1-S4).

DL-algorithm measurements were performed twice per case, the measured diameters were the exact same in every case and at every location, meaning perfect reproducibility (exact to eight decimal places).

Measurement times

In our research setting, the mean MT for the three human readers was 4:48±1:55 min per case (range 2:00–13:00 min, Table S2). Significant differences were seen when comparing the less experienced reader (R1) with the two more experienced, fellowship-trained readers (R1–R2: P<0.001, R1–R3: P=0.001).

The DL-algorithm performed measurements autonomously in 2:19±0:22 min (incl. generation of visual outputs), which was significantly faster compared to human reader (P<0.001).

The calculated average time to analyze a case with support of the DL-algorithm would be 1:38 min. We accounted for a failure rate of 13% (success rate: 87.0%, *Table 3*), a human MT for one location would be 36 seconds (4:48 min total for 8 locations) plus one minute to check visual outputs. This would result in an average of 3:10 min saved for measurements per case.

Discussion

In this study, we evaluated the accuracy of a DL-algorithm to perform thoracic aorta diameter measurements according to AHA-guidelines in more than 350 patients and further

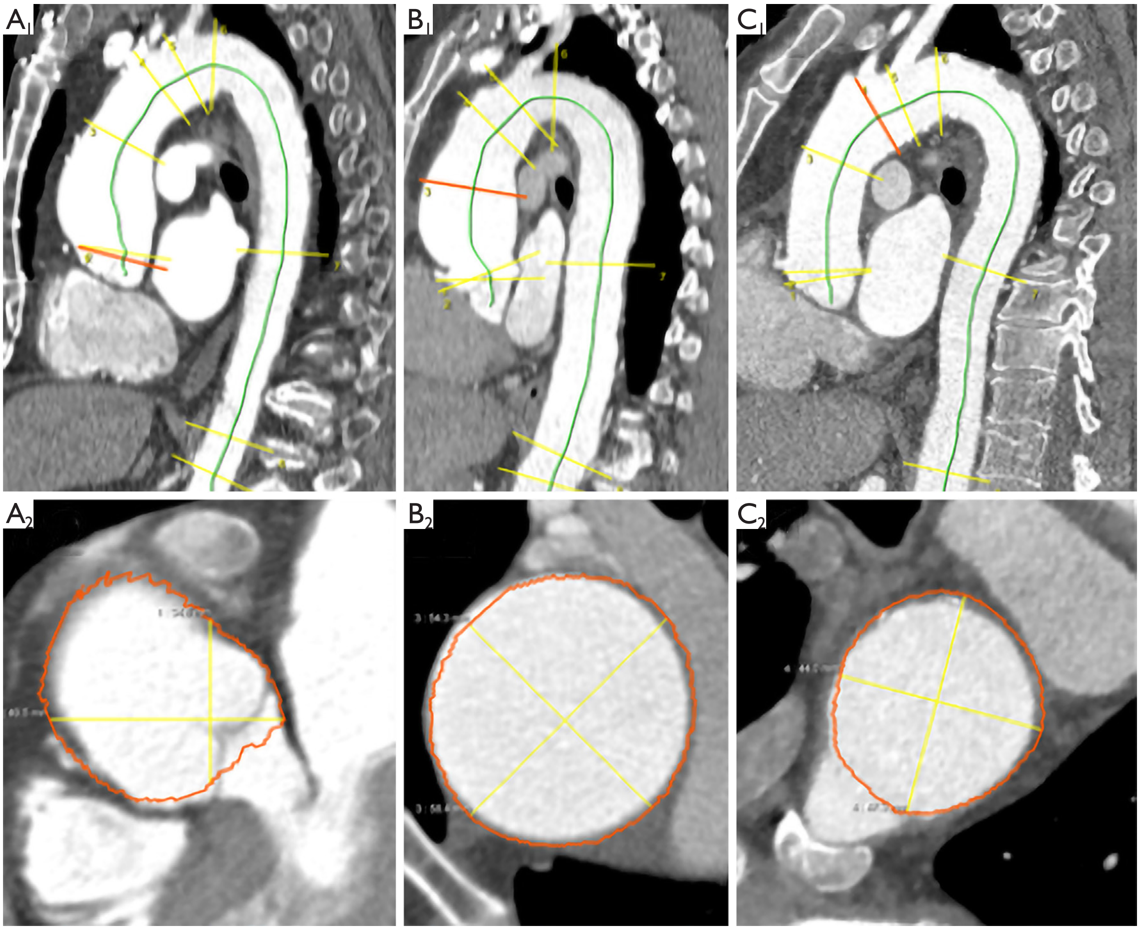

Figure 4 Example cases. (A1) tilted measurement plane causing overestimation. AS plane (orange) showed tilt. (A2) the actual measured plane was oblique coronal, left ventricular outflow tract, aortic valve with two of three leaflets and aortic sinus were visible. (B1) correct angle of AA measurement plane. Note that AS and STJ planes are tilted. (B2) acceptable measurement of an AA aneurysm with a diameter of 5.8cm. (C1) correct angle of PA measurement plane. (C2) correct measurement of PA aneurysm with a diameter of 4.7 cm. AS, aortic sinus; AA, ascending aorta; STJ, sinotubular junction; PA, proximal arch.

compared them to radiologists' measurements. We showed four major results: (I) there was a mean variance of up to ±3.4 mm for radiologists in a research setting (this constitutes perfect measurement conditions) using the established semi-automatic workflow. This observed variance is in agreement or slightly lower than previously described variability of up to 5 mm difference between expert readers (15). Opposite to that, the DL-algorithm was highly precise but less accurate in repeated measurements. (II) We showed that time required by human readers for guideline-compliant measurements under these perfect conditions was about five minutes per case. (III) At 87.0% of all measurement locations in our clinical cohort, the DL-algorithm provided measurement results within the expected margin of variance and therefore, were coherent to results of human readers. This finding resulted in an expected time-saving of more than 3 min per case for the radiologist. (IV) In review of discrepant cases, errors by the DL-algorithm were found predominantly at the aortic root (in 139/145 cases); these cases could be easily identified by DL-algorithms' visual outputs and therefore reduce interaction/re-measurements to a minimum.

In general, there are obstacles in how to measure diameters of tubular structures on CT data. Centerline based measurements are today's gold standard and have superiority over the double oblique technique based on multiplanar reformations as previously demonstrated (14,15). Nevertheless, the accuracy of measurements is debatable.

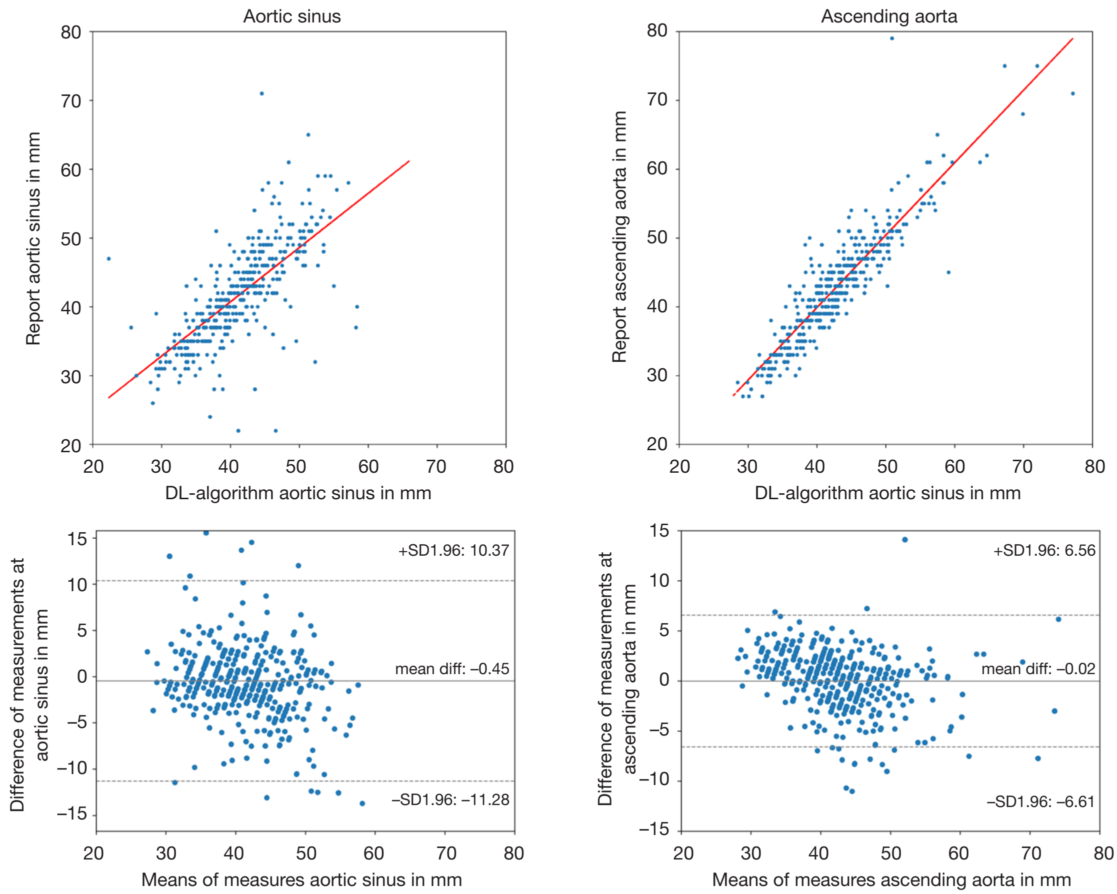

Figure 5 Scatter and Bland-Altman plots of measurements (in mm) at AS and AA. AS, aortic sinus; AA, ascending aorta; STJ, sinotubular junction; DL, deep learning; Diff, difference; SD, standard deviation.

Elefteriades *et al.* commented that "1–2 mm is not enough to detect change" and "you cannot have confidence in measured change of 3 or 4 mm" (25). These statements match our findings: as we investigated inter-observer variability between radiologists in the research setting, ICC analysis showed excellent agreement overall, but the agreement at STJ was only moderate to good between observers. Our prediction model interval of mean values which gives an estimate of expected variability ranged between ±2.5–3.4 mm depending on case and location. The range is likely to be even wider in a clinical setting as the current assessments were performed under perfect conditions (quiet environment, no telephone calls). In total, DL-measurements were outside the prediction intervall but this was mainly caused by a few outliers. In daily clinical practice, multiple factors influence aortic diameter measurements, for example: CT scan technique, reader experience, measurement technique used, stress, knowledge of previously reported diameter. Hence, we agree that one must consider an impreciseness of up to 3–4 mm (25). Compared to a study from McComb which used native CT datasets from a US lung cancer screening trial to investigate normal aortic diameters, our cohort of known and suspect dilatation had higher diameters for the aortic root and ascending aorta and similar diameters for arch and descending aorta (26). While the average patient age is similar, our inclusion criteria consisted of known or suspected dilatation which can explain this observation. In general, absolute diameter of more than 55 mm is of highest relevance since these patients usually require surgical therapy (9,26). Therefore, in our cohort, we considered a measurement difference of >5 mm as relevant which was also justified by Quint *et al.* who found 5mm differences within the 95% confidence interval between expert readers (15).

Opposite to human measurements, there was no variance in repeated DL-measurements of the same case, representing perfect preciseness. This was also found in another study which showed that AI-support reduces variability of aortic measurement (19). Statistically, over all test cases, there was a difference between human and DL-measurements but this was likely caused by rather large differences in a few cases. These errors could easily be detected by inspection of DL-algorithms' visual outputs. In most cases and locations, the measurements by the DL-algorithm were quite similar to human measurements. The crucial question is how to weigh human variability or neutralization of variability in human measurements by a DL-algorithm versus a few inaccurate measurements, respectively. We believe that the variability in human measurements is bound to remain and appreciate DL's neutral measurements to achieve more objectivity.

In-depth analysis of undetermined cases (145/399) revealed that most differences were located at the aortic root (n=131, 90.8%). The aortic root might have to be re-measured in a third of cases (133/399, 33.3%) but the majority of all measurements (2,778/3,192, 87.0%) were approved via easily assessable DL-measurement outputs (14). Our sub-analysis of follow-up cases did not show a case with a true diameter increase of >5 mm at any location between two scans. DL produced two false-positive cases which were caused by a tilted measurement plane at the aortic root. In the rest of cases, DL-measurements were coherent; however, the full effect of the DL-algorithm on follow-up exams requires an analysis of a larger follow-up cohort.

In 34/399 cases, an aneurysm was misclassified by the algorithm which meant that the DL-measurement was below our cutoff. These situations could be easily identified by the visual outputs of the DL-algoritm. Opposite to that, in 35 cases radiologists overestimated diameters resulting in diagnosis of aneurysm. Deficits in understanding software applications or stress could cause those errors. In addition, centerline based measurements require a more complex understanding of both technique and anatomy.

Potential time savings are a major advantage of the DL-algorithm: our readers needed about 5 minutes per case for centerline measurements which is similar to in the literature reported centerline MT (13,19). The DL-algorithm would save more than 3 minutes per case, which could easily add up to multiple hours per week (27). It is important to mention that the algorithms' measurements do not involve human input so the radiologist can continue to assess the exam until the measurements are completed. In the majority of incoherent cases in our study, only the aortic root had to be re-measured. In a few cases, measurements of aortic arch and descending aorta were incoherent or missing which could be explained by the ostia of the supraaortal arteries and lack of landmark identification by the DL-algorithm.

The DL-algorithm provided a high processing rate of cases (>97%), still 6 cases were not calculated which could either be a general software error (reproducible and non-reproducible) or an error in landmark detection. Furthermore, it provided additional information: DL-measurements of mid and distal descending aorta were available, and review showed that they were correct in 84.8% of cases. Those measurements were not available in our radiologic reports because they were omitted in the standard workflow since most aneurysms are located at the aortic root and AA.

There are several limitations including some with special regard to the use of DL-software (28): first, this was a single center, retrospective analysis. Second, since all scans were ECG-gated, motion artifacts were minimized. Additionally, aortic replacement surgery or post-stenting were excluded. Third, imaging data from only one manufacturer was analyzed; performance on exams acquired on scanners from other vendors might vary. Fourth, the reference standard for dataset A were measurements extracted from the radiology reports and not re-measured in a reseach setting. Fifth, the algorithm is a measuring tool, it was not built to make diagnoses or detect pathologies like intramural hematoma or aortic dissection, which have to be evaluated by the radiologist.

In summary, the evaluated DL-algorithm performed fully automatic, guideline-compliant aortic measurements reliably in 87% of all measurements and performed repeated measurements of the same CT scan with zero variance. In about one third of cases, the aortic root had to be re-measured, however time savings in the order of 3 minutes per case were still observed. Thereby, it is a foundation for a tool supporting radiologists in guideline-compliant aortic measurements.

Acknowledgments

We thank Mitchell A. Collins for proofreading this manuscript.
Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/qims-21-142). JS is an employee of Siemens

Healthineers and received personal fees, RK is a consultant for Siemens Healthineers. JS and RK both helped in installation and maintance of the software but were not involved in study design, data analysis or interpretation. They report that they have a patent US2020/0160527Al pending to Siemens Healthineers. The other authors have no conflict of interest to declare.

Disclaimers: Siemens Healthineers provided the prototype DL-algorithm. Two co-authors are affiliated with Siemens Healthineers (Jonathan. I. Sperl, employee and R. Kärgel, consultant). Siemens Healthineers had no influence on study design and data analysis.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). All data was encoded prior to any analysis to preserve patient anonymity. The Ethics Commitee for Northwest and Central Switzerland approved this study (ID: 2019-01053) and individual consent for this retrospective analysis was waived.

References

1. Booher AM, Eagle KA. Diagnosis and management issues in thoracic aortic aneurysm. Am Heart J 2011;162:38-46.e1.
2. Goldfinger JZ, Halperin JL, Marin ML, Stewart AS, Eagle KA, Fuster V. Thoracic aortic aneurysm and dissection. J Am Coll Cardiol 2014;64:1725-39.
3. Mokashi SA, Svensson LG. Guidelines for the management of thoracic aortic disease in 2017. Gen Thorac Cardiovasc Surg 2019;67:59-65.
4. Olsson C, Thelin S, Stahle E, Ekbom A, Granath F. Thoracic aortic aneurysm and dissection: increasing prevalence and improved outcomes reported in a nationwide population-based study of more than 14,000 cases from 1987 to 2002. Circulation 2006;114:2611-8.
5. Davies RR, Goldstein LJ, Coady MA, Tittle SL, Rizzo JA, Kopf GS, Elefteriades JA. Yearly rupture or dissection rates for thoracic aortic aneurysms: simple prediction based on size. Ann Thorac Surg 2002;73:17-27; discussion -8.
6. Clouse WD, Hallett JW Jr, Schaff HV, Gayari MM, Ilstrup DM, Melton LJ 3rd. Improved prognosis of thoracic aortic aneurysms: a population-based study. JAMA 1998;280:1926-9.
7. Keane MG, Wiegers SE, Plappert T, Pochettino A, Bavaria JE, Sutton MG. Bicuspid aortic valves are associated with aortic dilatation out of proportion to coexistent valvular lesions. Circulation 2000;102:III35-9.
8. Ikonomidis JS, Ivey CR, Wheeler JB, Akerman AW, Rice A, Patel RK, Stroud RE, Shah AA, Hughes CG, Ferrari G, Mukherjee R, Jones JA. Plasma biomarkers for distinguishing etiologic subtypes of thoracic aortic aneurysm disease. J Thorac Cardiovasc Surg 2013;145:1326-33.
9. Hiratzka LF, Bakris GL, Beckman JA, Bersin RM, Carr VF, Casey DE Jr, et al. 2010 ACCF/AHA/AATS/ACR/ASA/SCA/SCAI/SIR/STS/SVM guidelines for the diagnosis and management of patients with Thoracic Aortic Disease: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, American Association for Thoracic Surgery, American College of Radiology, American Stroke Association, Society of Cardiovascular Anesthesiologists, Society for Cardiovascular Angiography and Interventions, Society of Interventional Radiology, Society of Thoracic Surgeons, and Society for Vascular Medicine. Circulation 2010;121:e266-369.
10. Mendoza DD, Kochar M, Devereux RB, Basson CT, Min JK, Holmes K, Dietz HC, Milewicz DM, LeMaire SA, Pyeritz RE, Bavaria JE, Maslen CL, Song H, Kroner BL, Eagle KA, Weinsaft JW; GenTAC (National Registry of Genetically Triggered Thoracic Aortic Aneurysms and Cardiovascular Conditions) Study Investigators. Impact of image analysis methodology on diagnostic and surgical classification of patients with thoracic aortic aneurysms. Ann Thorac Surg 2011;92:904-12.
11. Boskamp T, Rinck D, Link F, Kummerlen B, Stamm G, Mildenberger P. New vessel analysis tool for morphometric quantification and visualization of vessels in CT and MR imaging data sets. Radiographics 2004;24:287-97.
12. Green DB, Palumbo MC, Lau C. Imaging of

Thoracoabdominal Aortic Aneurysms. J Thorac Imaging 2018;33:358-65.

13. Müller-Eschner M, Rengier F, Partovi S, Weber TF, Kopp-Schneider A, Geisbüsch P, Kauczor HU, von Tengg-Kobligk H. Accuracy and variability of semiautomatic centerline analysis versus manual aortic measurement techniques for TEVAR. Eur J Vasc Endovasc Surg 2013;45:241-7.
14. Rengier F, Weber TF, Giesel FL, Bockler D, Kauczor HU, von Tengg-Kobligk H. Centerline analysis of aortic CT angiographic examinations: benefits and limitations. AJR Am J Roentgenol 2009;192:W255-63.
15. Quint LE, Liu PS, Booher AM, Watcharotone K, Myles JD. Proximal thoracic aortic diameter measurements at CT: repeatability and reproducibility according to measurement method. Int J Cardiovasc Imaging 2013;29:479-88.
16. Lu JT, Brooks R, Hahn S, Chen J, Buch V, Kotecha G, et al. editors. DeepAAA: Clinically Applicable and Generalizable Detection of Abdominal Aortic Aneurysm Using Deep Learning. Cham: Springer International Publishing, 2019.
17. Biesdorf A, Rohr K, Feng D, von Tengg-Kobligk H, Rengier F, Böckler D, Kauczor HU, Wörz S. Segmentation and quantification of the aortic arch using joint 3D model-based segmentation and elastic image registration. Med Image Anal 2012;16:1187-201.
18. Sedghi Gamechi Z, Bons LR, Giordano M, Bos D, Budde RPJ, Kofoed KF, Pedersen JH, Roos-Hesselink JW, de Bruijne M. Automated 3D segmentation and diameter measurement of the thoracic aorta on non-contrast enhanced CT. Eur Radiol 2019;29:4613-23.
19. Rueckel J, Reidler P, Fink N, Sperl J, Geyer T, Fabritius MP, Ricke J, Ingrisch M, Sabel BO. Artificial intelligence assistance improves reporting efficiency of thoracic aortic aneurysm CT follow-up. Eur J Radiol 2021;134:109424.
20. Ghesu FC, Georgescu B, Zheng Y, Grbic S, Maier A, Hornegger J, Comaniciu D. Multi-Scale Deep Reinforcement Learning for Real-Time 3D-Landmark Detection in CT Scans. IEEE Trans Pattern Anal Mach Intell 2019;41:176-89.
21. Yang D, Xu D, Zhou SK, Georgescu B, Chen M, Grbic S, Metaxas D, Comaniciu D. editors. Automatic Liver Segmentation Using an Adversarial Image-to-Image Network. Cham: Springer International Publishing, 2017.
22. Popović ZB, Thomas JD. Assessing observer variability: a user's guide. Cardiovasc Diagn Ther 2017;7:317-24.
23. Koller M. robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. 2016 2016;75:24.
24. Michelena HI, Khanna AD, Mahoney D, Margaryan E, Topilsky Y, Suri RM, et al. Incidence of aortic complications in patients with bicuspid aortic valves. JAMA 2011;306:1104-12.
25. Elefteriades JA, Farkas EA. Thoracic aortic aneurysm clinically pertinent controversies and uncertainties. J Am Coll Cardiol 2010;55:841-57.
26. McComb BL, Munden RF, Duan F, Jain AA, Tuite C, Chiles C. Normative reference values of thoracic aortic diameter in American College of Radiology Imaging Network (ACRIN 6654) arm of National Lung Screening Trial. Clin Imaging 2016;40:936-43.
27. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, Erickson BJ, Kallmes DF. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 2015;22:1191-8.
28. Bluemke DA, Moy L, Bredella MA, Ertl-Wagner BB, Fowler KJ, Goh VJ, Halpern EF, Hess CP, Schiebler ML, Weiss CR. Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers-From the Radiology Editorial Board. Radiology 2020;294:487-9.

Cite this article as: Pradella M, Weikert T, Sperl JI, Kärgel R, Cyriac J, Achermann R, Sauter AW, Bremerich J, Stieltjes B, Brantner P, Sommer G. Fully automated guideline-compliant diameter measurements of the thoracic aorta on ECG-gated CT angiography using deep learning. Quant Imaging Med Surg 2021;11(10):4245-4257. doi: 10.21037/qims-21-142