Published in final edited form as:
Eur Radiol. 2023 February ; 33(2): 1102–1111. doi:10.1007/s00330-022-09056-z.

Inter-observer variability of expert-derived morphologic risk predictors in aortic dissection

Martin J. Willemink[1], **Domenico Mastrodicasa**[1,2], **Mohammad H. Madani**[1], **Marina Codari**[1], **Leonid L. Chepelev**[1], **Gabriel Mistelbauer**[1], **Kate Hanneman**[3], **Maral Ouzounian**[4], **Daniel Ocazionez**[5], **Rana O. Afifi**[6], **Joan M. Lacomis**[7], **Luigi Lovato**[8], **Davide Pacini**[9], **Gianluca Folesani**[9], **Ricarda Hinzpeter**[10], **Hatem Alkadhi**[10], **Arthur E. Stillman**[11], **Anna M. Sailer**[1], **Valery L. Turner**[1], **Virginia Hinostroza**[1], **Kathrin Bäumler**[1], **Anne S. Chin**[12], **Nicholas S. Burris**[13], **D. Craig Miller**[14], **Michael P. Fischbein**[14], **Dominik Fleischmann**[1,2]

[1]Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA

[2]Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA, USA

[3]Department of Medical Imaging, Peter Munk Cardiac Centre, Toronto General Hospital, University of Toronto, Toronto, Canada

[4]Department of Surgery, University of Toronto, Toronto, Canada

[5]Department of Radiology, McGovern Medical School at The University of Texas Health Science Center at Houston (UTHealth), Houston, TX, USA

[6]Department of Cardiothoracic and Vascular Surgery, McGovern Medical School at The University of Texas Health Science Center at Houston (UTHealth), Houston, TX, USA

[7]Department of Radiology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

✉Dominik Fleischmann, d.fleischmann@stanford.edu.
Martin J. Willemink and Domenico Mastrodicasa contributed equally to the manuscript.
Extended author information available on the last page of the article

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-09056-z.

Declarations
Guarantor The scientific guarantor of this publication is Dominik Fleischmann, MD.

Conflict of interest Martin. J. Willemink is a Junior Deputy Editor of European Radiology. They have not taken part in the review or selection process of this article.
The other authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Study subjects or cohorts overlap Some of the subjects ($n = 47$, out of a total of 72) included in this study were part of the population ($n = 83$) included in reference #14 (Sailer et al doi:10.1161/CIRCIMAGING.116.005709)

Methodology

- retrospective
- experimental
- multi-center study

[8]Department of Radiology, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Policlinico di S. Orsola, Bologna, Italy

[9]Department of Cardiac Surgery, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Policlinico di S. Orsola, Bologna, Italy

[10]Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

[11]Department of Radiology, Emory University, Atlanta, GA, USA

[12]Département de Radiologie, Centre Hospitalier de l'Université de Montréal, Montreal, Canada

[13]Department of Radiology, University of Michigan, Ann Arbor, MI, USA

[14]Department of Cardiothoracic Surgery, Stanford University School of Medicine, Stanford, CA, USA

Abstract

Objectives—Establishing the reproducibility of expert-derived measurements on CTA exams of aortic dissection is clinically important and paramount for ground-truth determination for machine learning.

Methods—Four independent observers retrospectively evaluated CTA exams of 72 patients with uncomplicated Stanford type B aortic dissection and assessed the reproducibility of a recently proposed combination of four morphologic risk predictors (maximum aortic diameter, false lumen circumferential angle, false lumen outflow, and intercostal arteries). For the first interobserver variability assessment, 47 CTA scans from one aortic center were evaluated by expert-observer 1 in an unconstrained clinical assessment without a standardized workflow and compared to a composite of three expert-observers (observers 2–4) using a standardized workflow. A second inter-observer variability assessment on 30 out of the 47 CTA scans compared observers 3 and 4 with a constrained, standardized workflow. A third inter-observer variability assessment was done after specialized training and tested between observers 3 and 4 in an external population of 25 CTA scans. Inter-observer agreement was assessed with intraclass correlation coefficients (ICCs) and Bland-Altman plots.

Results—Pre-training ICCs of the four morphologic features ranged from 0.04 (−0.05 to 0.13) to 0.68 (0.49–0.81) between observer 1 and observers 2–4 and from 0.50 (0.32–0.69) to 0.89 (0.78–0.95) between observers 3 and 4. ICCs improved after training ranging from 0.69 (0.52–0.87) to 0.97 (0.94–0.99), and Bland-Altman analysis showed decreased bias and limits of agreement.

Conclusions—Manual morphologic feature measurements on CTA images can be optimized resulting in improved inter-observer reliability. This is essential for robust ground-truth determination for machine learning models.

Keywords

Aortic dissection; Computed tomography angiography; Variability, inter-observer

Introduction

Aortic dissection is a devastating disease with an incidence of 4 to 7 per 100,000 people per year [1–5]. The optimal treatment of patients with initially uncomplicated Stanford type B aortic dissection (uTBAD) is currently unknown and much debated. Advocates for thoracic endovascular aortic repair (TEVAR) emphasize the high rate of late complications in medically managed patients with uTBAD. However, the level of evidence to support preventive TEVAR is poor [6]. There is broad consensus among aortic specialists that better risk stratification is needed to accurately identify patients at high risk for late adverse events. Only high-risk patients are likely to benefit from preventive TEVAR [6], but reliable criteria for clinical risk stratification are yet to be identified. Several studies have attempted to identify high-risk morphologic features on computed tomography angiography (CTA) imaging features portending an increased risk of late complications in uTBAD [7]. Proposed morphologic imaging markers include aortic size [8–19], false lumen size [8–10, 14, 20, 21], primary entry tear size and location [10, 14, 21–23], false lumen patency and saccular configuration [9, 13, 14, 18, 21, 22], and false lumen longitudinal extent and configuration [13–15, 20, 22, 24, 25]. Only the reproducibility of relatively simple aortic diameter measurements has been evaluated in several studies noting substantial differences among expert readers [26–29].

Machine learning models have the potential to automatically identify morphologic features on CTA exams of patients with aortic dissection [30]. Training such models typically requires robust ground truth measurements, conducted manually by experts. We sought (i) to determine the inter-observer variability of expert-derived measurements of morphologic risk features on CTA images of uTBAD patients and (ii) to optimize the reproducibility of measurements. Highly reproducible measurements are not only essential for initial assessment and follow-up of clinical CTA exams but are also a prerequisite for the development of new machine learning models.

Materials and methods

Subjects

The study population was derived from the Registry of Aortic Diseases to Model Adverse Events and Progression (ROADMAP). In summary, patients with uTBAD who survived the initial hospitalization without complications and who underwent CTA within the first 2 weeks after the initial event were included. Clinical and imaging data were collected from a total of 10 tertiary aortic centers in North America and Europe. A total of 72 patients with uTBAD were included in the current study. The cohort was divided into two groups: a group of 47 patients from a single aortic center (Stanford University Hospital, Stanford, CA) and a group of 25 randomly selected patients from various aortic centers (Emory University Hospital, Atlanta, GA; University of Pittsburgh Hospital, Pittsburgh, PA; University Hospital of Texas Houston, Houston, TX; University Hospital of Toronto, Toronto, Canada; University of Bologna, Bologna, Italy; University Hospital Zurich, Zurich, Switzerland).

Morphologic imaging features

Expert observers evaluated the following four morphologic imaging markers using the same post-processing software on a dedicated workstation (iNtuition, TeraRecon, Inc.): (i) maximum aortic diameter, (ii) relative false lumen circumference, (iii) false lumen drainage pattern, and (iv) the number of identifiable intercostal arteries along the dissected aorta (Fig. 1). A detailed explanation of each morphologic imaging marker is described elsewhere [14]. In short, the process to extract the morphologic features begins with the semiautomated identification of the aortic centerline from the cardiac apex to the aortic bifurcation. The aortic centerline represents the common longitudinal axis of the entire aorta (true and false lumen combined) if dissected. The aortic centerline was subsequently used to obtain cross-sectional views orthogonal to the long axis of the aorta and to localize the maximum aortic diameter of the dissected aorta. The maximum aortic diameter was measured in millimeters. In the same orthogonal plane, the relative false lumen circumference was measured in angular degrees (°) (Fig. 1). The false lumen drainage pattern was assessed for each branch vessel (left subclavian artery, celiac artery, superior mesenteric artery, left/right renal arteries, inferior mesenteric artery, and left/right common iliac arteries) by determining the lumen of origin (true lumen, false lumen, or both). False lumen outflow was then quantified as the weighted sum of all branches draining the false lumen in milliliters per minute [14] (Fig. 2). Lastly, the number (*n*) of identifiable intercostal arteries branching off the aorta in the dissected portion was recorded. Data were automatically extracted and saved for further processing. The time required for the complete assessment of morphologic features from a single CT dataset was approximately 45 min.

Interobserver variability assessment

Three experiments were conducted to establish the reproducibility of morphologic CT features using the following scenarios: (i) unconstrained clinical assessment, where imaging features were evaluated in a clinical-type scenario without a standardized workflow so the reader could vary the sequence of imaging features on subsequent reads; (ii) constrained predefined standardized workflow, where imaging features were assessed in a standard order programmed into the post-processing workstation; and (iii) and constrained predefined standardized workflow *after* standardized training (Fig. 2).

Unconstrained clinical assessment (Experiment 1)

The purpose of the first evaluation was to simulate a clinical scenario where CT scans are evaluated by different cardiovascular radiologists with potentially different workflows and data extraction strategies. CTA scans of the 47 patients from a single tertiary aortic center (Stanford University Medical Center) were independently evaluated twice. The first evaluation was done by a pairwise comparison of a single observer (observer 1) who evaluated all 47 patients with a mix of three observers (observers 2, 3, and 4) who independently evaluated 16, 16, and 15 of these 47 patients, respectively (Fig. 2). Observer 1 evaluated the CT scans without a pre-set standardized workflow, and data were manually recorded for further processing. Observers 2–4 used a pre-defined, standardized workflow that allowed them to systematically evaluate each morphologic feature in the same order. Data were extracted and saved automatically. All four observers were physicians with at least 8 years of experience in cardiovascular imaging with considerable expertise specifically with CTA assessment of aortic dissection.

Constrained predefined standardized workflow (Experiment 2)

The purpose of the second evaluation was to systematically evaluate the inter-observer variability between two expert readers using an identical, standardized workflow, but before feature-specific training. For this experiment, 30 out of 47 CT scans were independently evaluated by two observers, respectively (observers 3 and 4, Fig. 2).

Optimization guidelines through feature-specific observer training

After the first two independent evaluations in the single aortic center cohort (Fig. 2), discrepancies between observers 3 and 4 were reviewed in consensus and together with a senior cardiovascular radiologist with more than 20 years of experience in order to identify sources of discrepancies and refine criteria for feature classification and measurements. An optimized assessment guideline was developed, which included clear rules and examples for typical and less typical cases of uTBAD (Fig. 3).

For *maximum aortic diameter*, the measurement was specified to include the aortic wall. We observed that the maximum aortic diameter often comprises a finite zone within the proximal descending thoracic aorta, rather than a unique point. The absence of a clearly defined plane impedes the measurement of the *false lumen circumferential angle*, as this feature can change substantially within a short range. In order to standardize these measurements, the instruction was as follows: (i) In the presence of a clear single location with maximum diameter, both maximum aortic diameter and false lumen circumferential angle were measured in that single plane; (ii) In all other cases, the observers identified a range in which the aorta maintained its maximum diameter. If the false lumen circumferential angle changed substantially within this range, it was recommended to record three measurements: the center of the range, 1 cm proximal, and 1 cm distal. The average of these three measurements was then used as a final result.

For the *false lumen outflow* pattern, if the dissection flap was not overriding the ostium of the aortic branch, it was defined as either true or false lumen. If the dissection flap was visible in the middle of the ostium, the *false lumen outflow* pattern was defined as both false lumen and true lumen (Fig. 3A). In case the evaluation of the dissection flap position was still not clear, then a previous or follow-up scan was assessed.

Constrained predefined standardized workflow after standardized training (Experiment 3)

The purpose of the third experiment was to address the previously unforeseen sources of discrepancies even amongst expert readers, by having readers undergo specific image feature training and establishing instructions for instances where measurements may be less certain.

To assess the effect of this expert-geared image analysis instruction, a third evaluation was performed (Fig. 1) in which two observers (observers 3 and 4) independently evaluated 25 previously unseen external CTA scans following the preset standardized workflow as before, but applying guidelines laid out during specific image feature training and applying the newly established rules and measurement strategy.

Statistical analysis

Continuous variables are listed as median (interquartile range) and categorical variables are listed as numbers (percentage). Differences between observers were evaluated. Individual probabilities of developing late complications were calculated based on the morphologic imaging markers and presence of connective tissue disease as described elsewhere [14]. Patients were divided into three risk categories: low (risk score < 6.05), intermediate (6.05 ≤ risk score ≤ 7.00), or high (risk score > 7.00) and subsequent reclassification rates between observers were calculated [14].

Pre-training and post-training inter-observer reliability were assessed with the single measures intraclass correlation coefficient (ICC) for continuous variables and linearly weighted kappa for categorical variables. Kappa-values were interpreted as: no agreement 0, poor agreement < 0.20, fair agreement 0.21–0.40, moderate agreement 0.41–0.60, good agreement 0.61–0.80, and excellent agreement > 0.80 [31, 32]. ICC values between 0.70 and 0.80 were considered good and ICC values greater than 0.80 were considered excellent [32]. The limits of inter-observer agreement were defined as the mean difference ± 1.96 × the standard deviation and displayed with Bland-Altman plots [33]. The proportions of agreement between observers were calculated for determining the outflow pattern (true lumen, false lumen, or both lumens) of each branch starting at the left subclavian artery (left subclavian artery, celiac artery, superior mesenteric artery, left and right renal arteries, inferior mesenteric artery, and left and right common iliac arteries). Analyses were performed with RStudio version 1.2 (RStudio) and SPSS Version 26 (IBM). *p* values below 0.05 were considered statistically significant. Results are reported according to the GRRAS Guidelines for Reporting Reliability and Agreement Studies [34].

Results

Differences between observers

Pre-training and post-training differences between observers are listed in Tables 1, 2 and 3, respectively, and displayed in Fig. 4. Mean pre-training differences between observer 1 and mixed observers 2–4 in the unconstrained clinical assessment of $N = 47$ CT scans were 0.8 (−1.7 to 3.8) mm for maximum aortic diameter, −17.0 (−30.5 to 3.8) degrees for false lumen circumferential angle, 141.3 (0.0 to 418.8) mL/min for false lumen outflow, and 2 (1 to 5) for the number of identifiable intercostal arteries.

The pre-training differences with the constrained predefined standardized workflow between observers 3 and 4 in $N = 30$ CT scans were slightly smaller: 0.2 (−1.5 to 2.7) mm for maximum aortic diameter, 0.4 (−14.0 to 16.1) degrees for false lumen circumferential angle, 0.0 (0.0 to 268.0) mL/min for false lumen outflow, and 0 (−2 to 1) for the number of identifiable intercostal arteries.

After training, the differences between observers 3 and 4 in $N = 25$ CT scans decreased further: 0.1 (−0.4 to 0.5) mm for maximum aortic diameter, −2.2 (−5.5 to 2.6) degrees for false lumen circumferential angle, 0.0 (0.0 to 0.0) mL/min for false lumen outflow, and 0 (−1 to 0) for the number of identifiable intercostal arteries.

Risk classification

The risk category reclassification rate was highest before training with 44.7% reclassifications between observer 1 and the mix of observers 2–4 in the unconstrained clinical assessment of $N = 47$ CT scans. The pre-training reclassification rate between observers 3 and 4 in $N = 30$ CT scans assessed according to the constrained predefined standardized workflow was slightly lower (30.3%). After training, the reclassification rate decreased to 16.0% between observers 3 and 4 in $N = 25$ CT scans. Specific reclassification results are listed in Table 4.

Inter-observer agreement and reliability

Pre-training and post-training ICCs and linearly weighted kappas are listed in Table 5. Bland-Altman plots (Supplementary Figure 2) indicated that limits of agreements were widest pre-training between observer 1 and the mix of observers 2–4 in the unconstrained clinical assessment of $N = 47$ CT scans, with a slight decrease for the pre-training assessment of observers 3 and 4 in $N = 30$ CT scans with the constrained predefined standardized workflow. The limits of agreements were the narrowest for the post-training assessment between observers 3 and 4 in $N = 25$ CT scans.

The pre-training and post-training proportions of the agreement for the false lumen outflow assessment are displayed in Supplementary Figure 3. Agreement was lowest between observer 1 and the mix of observers 2–4 pre-training using unconstrained clinical assessment, and highest after training between observers 3 and 4.

Discussion

Our study shows that manual expert-driven measurements of morphologic risk features in patients with uTBAD patients can be obtained with excellent inter-observer reliability and agreement provided that both, a standardized workflow and feature-specific rules are well-defined. Inter-observer variability of four morphologic imaging features before workflow standardization with automated data extraction and before workflow optimization through feature-definition and training was suboptimal, even among subspecialty cardiovascular radiologists. After implementing specific feature training and detailing methods of adjudication for complex features, the inter-observer variability improved substantially for all features. These results are important in illustrating how we can currently optimize the reproducibility of aortic measurements in the clinical setting, in the setting of imaging research, but also because developing reliable machine learning models to automate such tasks in the future requires robust ground truth information for training, validation, and testing purposes [35, 36]. Implementation of specific image feature training and establishing clear and exhaustive reporting guidelines are particularly important and necessary for the assessment of complex imaging variables encountered in the setting of aortic dissection. We will apply such standards in an ongoing effort to externally validate an imaging feature–

based risk prediction model in the retrospective registry of aortic diseases to model adverse events and progression (ROADMAP) cohort.

We evaluated four morphologic imaging markers that have been described previously [7, 14]. Our systematic approach mimicked a clinical scenario with a mix of observers as well as a scenario where the same observers evaluated all scans. The most frequently used and clinically most relevant imaging marker for patients with aortic dissection is the maximum aortic diameter. Our results indicate that this is the most reproducible imaging feature out of the four assessed markers. Although maximum aortic diameter was relatively reproducible, there was still a significant improvement after standardized training. The other morphologic imaging features included false lumen circumferential angle, false lumen outflow, and the number of identifiable intercostal arteries. The reproducibility of these features was initially limited but improved substantially after applying the optimization guideline. Similar to reproducibility, reclassification between observers also improved, which is essential both in clinical practice and in defining the ground truth for machine learning models.

Previous studies have evaluated the inter-observer variability of maximum aortic diameter measurements in patients with abdominal aortic aneurysms. Similar to our results, the majority of studies found a relatively poor inter-observer variability [37–39]. The maximum aortic diameter, however, was the most reproducible imaging feature out of the four assessed features in our study. Other studies evaluated the variability of clinical measurements of the thoracic aorta, including the aortic root as measured on echocardiography, CT, and MR exams [40, 41], and the maximum aortic diameter as measured on CT [28]. To the best of our knowledge, the reproducibility of the other three morphologic imaging features in this study (false lumen circumferential angle, false lumen outflow, and the number of identifiable intercostal arteries), and the reproducibility of many other candidate morphological predictors in the literature has not been evaluated before. Elefteriades and colleagues reviewed discrepancies in measurements of the thoracic aorta and concluded that discrepancies arise from image formatting, diameter definition (including or excluding aortic wall), irregular aortic contour, and inconsistent reporting [42]. As the current study suggests, most of these discrepancies can be prevented by providing clear guidelines resulting in improved interobserver variability.

Our study has some limitations. First, this was a retrospective study based on previously gathered CT imaging data. The quality of more recent CT images may be better, resulting in improved inter-observer variability. Second, due to feasibility reasons, not all scans from the ROADMAP registry have been evaluated for this interobserver variability study. However, the sample size of the presented work was large enough to assess our objectives. Last, variability across sites is different than within a site. We cannot be entirely certain that the improved reproducibility is due to training since the pre- and post-training populations were different. But in general, we would expect that variability increases when external data are used. Even experienced readers could have problems analyzing images from another site (different protocol, resolution, noise, contrast, etc.). The fact that the interobserver agreement improved after training and testing in external datasets is therefore encouraging. Further research should evaluate the interobserver variability across sites.

In conclusion, our systematic inter-observer variability study showed that manual measurements of morphologic risk features on CT images of uTBAD patients can be optimized, resulting in excellent inter-observer reliability and agreement, and a decrease in reclassifications. This is important because apart from optimizing the reproducibility of aortic measurements in the clinical setting, establishing robust ground truth information is essential for training, validation, and testing of machine learning models. It is therefore recommended to apply standardization and optimization guidelines before determining ground truth labels for machine learning models in complex organs such as the aorta.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank Shannon G. Walters, RT, MS, from the Stanford 3D and Quantitative Imaging Laboratory for his support with the image processing workflow.

Funding

This study has received funding from the American Heart Association grant numbers 18POST34030192 (MJW) and 826389 (MC), and a research grant (5T32EB009035) from the National Institute of Biomedical Imaging and Bioengineering (DM).

Abbreviations

CT	Computed tomography
CTA	Computed tomography angiography
GRRAS	Guidelines for reporting reliability and agreement studies
ICC	Intraclass correlation coefficient
ROADMAP	Registry of aortic dissections to model adverse events and progression
TEVAR	Thoracic endovascular aortic repair
uTBAD	Uncomplicated acute Stanford type B aortic dissection

References

1. Hagan PG, Nienaber CA, Isselbacher EM et al. (2000) The International Registry of Acute Aortic Dissection (IRAD): new insights into an old disease. JAMA 283:897–903 [PubMed: 10685714]
2. Howard DP, Banerjee A, Fairhead JF et al. (2013) Population-based study of incidence and outcome of acute aortic dissection and premorbid risk factor control: 10-year results from the Oxford Vascular Study. Circulation 127:2031–2037 [PubMed: 23599348]
3. Landenhed M, Engstrom G, Gottsater A et al. (2015) Risk profiles for aortic dissection and ruptured or surgically treated aneurysms: a prospective cohort study. J Am Heart Assoc 4:e001513 [PubMed: 25609416]
4. Nienaber CA, Clough RE (2015) Management of acute aortic dissection. Lancet 385:800–811 [PubMed: 25662791]

5. Fleischmann D, Afifi RO, Casanegra AI et al. (2022) Imaging and Surveillance of Chronic Aortic Dissection: A Scientific Statement From the American Heart Association. Circ Cardiovasc Imaging 10.1161/HCI.0000000000000075:HCI0000000000000075

6. MacGillivray TE, Gleason TG, Patel HJ et al. (2022) The Society of Thoracic Surgeons/American Association for Thoracic Surgery clinical practice guidelines on the management of type B aortic dissection. J Thorac Cardiovasc Surg 10.1016/j.jtcvs.2021.11.091

7. Spinelli D, Benedetto F, Donato R et al. (2018) Current evidence in predictors of aortic growth and events in acute type B aortic dissection. J Vasc Surg 68:1925–1935 e1928 [PubMed: 30115384]

8. Chang CP, Liu JC, Liou YM, Chang SS, Chen JY (2008) The role of false lumen size in prediction of in-hospital complications after acute type B aortic dissection. J Am Coll Cardiol 52:1170–1176 [PubMed: 18804746]

9. Delsart P, Beregi JP, Devos P, Haulon S, Midulla M, Mounier-Vehier C (2014) Thrombocytopenia: an early marker of late mortality in type B aortic dissection. Heart Vessels 29:220–230 [PubMed: 23604315]

10. Evangelista A, Salas A, Ribera A et al. (2012) Long-term outcome of aortic dissection with patent false lumen: predictive role of entry tear size and location. Circulation 125:3133–3141 [PubMed: 22615344]

11. Grommes J, Greiner A, Bendermacher B et al. (2014) Risk factors for mortality and failure of conservative treatment after aortic type B dissection. J Thorac Cardiovasc Surg 148(2155–2160):e2151

12. Jonker FH, Trimarchi S, Rampoldi V et al. (2012) Aortic expansion after acute type B aortic dissection. Ann Thorac Surg 94:1223– 1229 [PubMed: 22776085]

13. Kudo T, Mikamo A, Kurazumi H, Suzuki R, Morikage N, Hamano K (2014) Predictors of late aortic events after Stanford type B acute aortic dissection. J Thorac Cardiovasc Surg 148:98–104 [PubMed: 24029294]

14. Sailer AM, van Kuijk SM, Nelemans PJ et al. (2017) Computed tomography imaging features in acute uncomplicated Stanford type-B aortic dissection predict late adverse events. Circ Cardiovasc Imaging:10

15. Sueyoshi E, Nagayama H, Hayashida T, Sakamoto I, Uetani M (2013) Comparison of outcome in aortic dissection with single false lumen versus multiple false lumens: CT assessment. Radiology 267:368–375 [PubMed: 23297333]

16. Sueyoshi E, Sakamoto I, Hayashi K, Yamaguchi T, Imada T (2004) Growth rate of aortic diameter in patients with type B aortic dissection during the chronic phase. Circulation 110:II256–II261 [PubMed: 15364872]

17. Tanaka A, Sakakibara M, Ishii H et al. (2014) Influence of the false lumen status on clinical outcomes in patients with acute type B aortic dissection. J Vasc Surg 59:321–326 [PubMed: 24139979]

18. Tolenaar JL, Froehlich W, Jonker FH et al. (2014) Predicting in-hospital mortality in acute type B aortic dissection: evidence from International Registry of Acute Aortic Dissection. Circulation 130: S45–S50 [PubMed: 25200055]

19. Trimarchi S, Tolenaar JL, Jonker FH et al. (2013) Importance of false lumen thrombosis in type B aortic dissection prognosis. J Thorac Cardiovasc Surg 145:S208–S212 [PubMed: 23260434]

20. Tolenaar JL, van Keulen JW, Jonker FH et al. (2013) Morphologic predictors of aortic dilatation in type B aortic dissection. J Vasc Surg 58:1220–1225 [PubMed: 23827338]

21. Ueki C, Sakaguchi G, Shimamoto T, Komiya T (2014) Prognostic factors in patients with uncomplicated acute type B aortic dissection. Ann Thorac Surg 97:767–77 discussion 773 [PubMed: 24360090]

22. Kamman AV, Brunkwall J, Verhoeven EL, Heijmen RH, Trimarchi S, Trialists A (2017) Predictors of aortic growth in uncomplicated type B aortic dissection from the Acute Dissection Stent Grafting or Best Medical Treatment (ADSORB) database. J Vasc Surg 65:964–971 e963 [PubMed: 27876516]

23. Loewe C, Czerny M, Sodeck GH et al. (2012) A new mechanism by which an acute type B aortic dissection is primarily complicated, becomes complicated, or remains uncomplicated. Ann Thorac Surg 93:1215–1222 [PubMed: 22402280]

24. Kamman AV, Jonker FHW, Sechtem U et al. (2017) Predictors of stable aortic dimensions in medically managed acute aortic syndromes. Ann Vasc Surg 42:143–149 [PubMed: 28390915]

25. Kitamura T, Torii S, Oka N et al. (2015) Impact of the entry site on late outcome in acute Stanford type B aortic dissectiondagger. Eur J Cardiothorac Surg 48:655–66 discussion 661–652 [PubMed: 25602060]

26. Quint LE, Liu PS, Booher AM, Watcharotone K, Myles JD (2013) Proximal thoracic aortic diameter measurements at CT: repeatability and reproducibility according to measurement method. Int J Cardiovasc Imaging 29:479–488 [PubMed: 22864960]

27. Regeer MV, van Rosendael PJ, Kamperidis V et al. (2015) Effect of statins on aortic root growth rate in patients with bicuspid aortic valve anatomy. Int J Cardiovasc Imaging 31:1583–1590 [PubMed: 26319215]

28. Rudarakanchana N, Bicknell CD, Cheshire NJ et al. (2014) Variation in maximum diameter measurements of descending thoracic aortic aneurysms using unformatted planes versus images corrected to aortic centerline. Eur J Vasc Endovasc Surg 47:19–26 [PubMed: 24183250]

29. Singh K, Jacobsen BK, Solberg S et al. (2003) Intra- and interobserver variability in the measurements of abdominal aortic and common iliac artery diameter with computed tomography. The Tromso study. Eur J Vasc Endovasc Surg 25:399–407 [PubMed: 12713777]

30. Hahn LD, Mistelbauer G, Higashigaito K et al. (2020) True and false lumen segmentation in uncomplicated type B aortic dissection using machine learning. Radiol Cardiothorac Imaging 2:e190179 [PubMed: 33778582]

31. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174 [PubMed: 843571]

32. van Hamersvelt RW, Willemink MJ, Takx RA et al. (2014) Cardiac valve calcifications on low-dose unenhanced ungated chest computed tomography: inter-observer and inter-examination reliability, agreement and variability. Eur Radiol 24:1557–1564 [PubMed: 24816936]

33. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310 [PubMed: 2868172]

34. Kottner J, Audige L, Brorson S et al. (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64:96–106 [PubMed: 21130355]

35. Wichmann JL, Willemink MJ, De Cecco CN (2020) Artificial intelligence and machine learning in radiology: current state and considerations for routine clinical implementation. Invest Radiol 55: 619–627 [PubMed: 32776769]

36. Willemink MJ, Koszek WA, Hardell C et al. (2020) Preparing medical imaging data for machine learning. Radiology 295:4–15 [PubMed: 32068507]

37. Cayne NS, Veith FJ, Lipsitz EC et al. (2004) Variability of maximal aortic aneurysm diameter measurements on CT scan: significance and methods to minimize. J Vasc Surg 39:811–815 [PubMed: 15071447]

38. Jaakkola P, Hippelainen M, Farin P, Rytkonen H, Kainulainen S, Partanen K (1996) Interobserver variability in measuring the dimensions of the abdominal aorta: comparison of ultrasound and computed tomography. Eur J Vasc Endovasc Surg 12:230–237 [PubMed: 8760988]

39. Lederle FA, Wilson SE, Johnson GR et al. (1995) Variability in measurement of abdominal aortic aneurysms. Abdominal Aortic Aneurysm Detection and Management Veterans Administration Cooperative Study Group. J Vasc Surg 21:945–952 [PubMed: 7776474]

40. Plonek T, Berezowski M, Bochenek M et al. (2019) A comparison of aortic root measurements by echocardiography and computed tomography. J Thorac Cardiovasc Surg 157:479–486 [PubMed: 30227996]

41. Freeman LA, Young PM, Foley TA, Williamson EE, Bruce CJ, Greason KL (2013) CT and MRI assessment of the aortic root and ascending aorta. AJR Am J Roentgenol 200:W581–W592 [PubMed: 23701088]

42. Elefteriades JA, Mukherjee SK, Mojibian H (2020) Discrepancies in measurement of the thoracic aorta: JACC review topic of the week. J Am Coll Cardiol 76:201–217 [PubMed: 32646571]

Key Points

- Clinical fashion manual measurements of aortic CTA imaging features showed poor inter-observer reproducibility.
- A standardized workflow with standardized training resulted in substantial improvements with excellent inter-observer reproducibility.
- Robust ground truth labels obtained manually with excellent inter-observer reproducibility are key to develop reliable machine learning models.

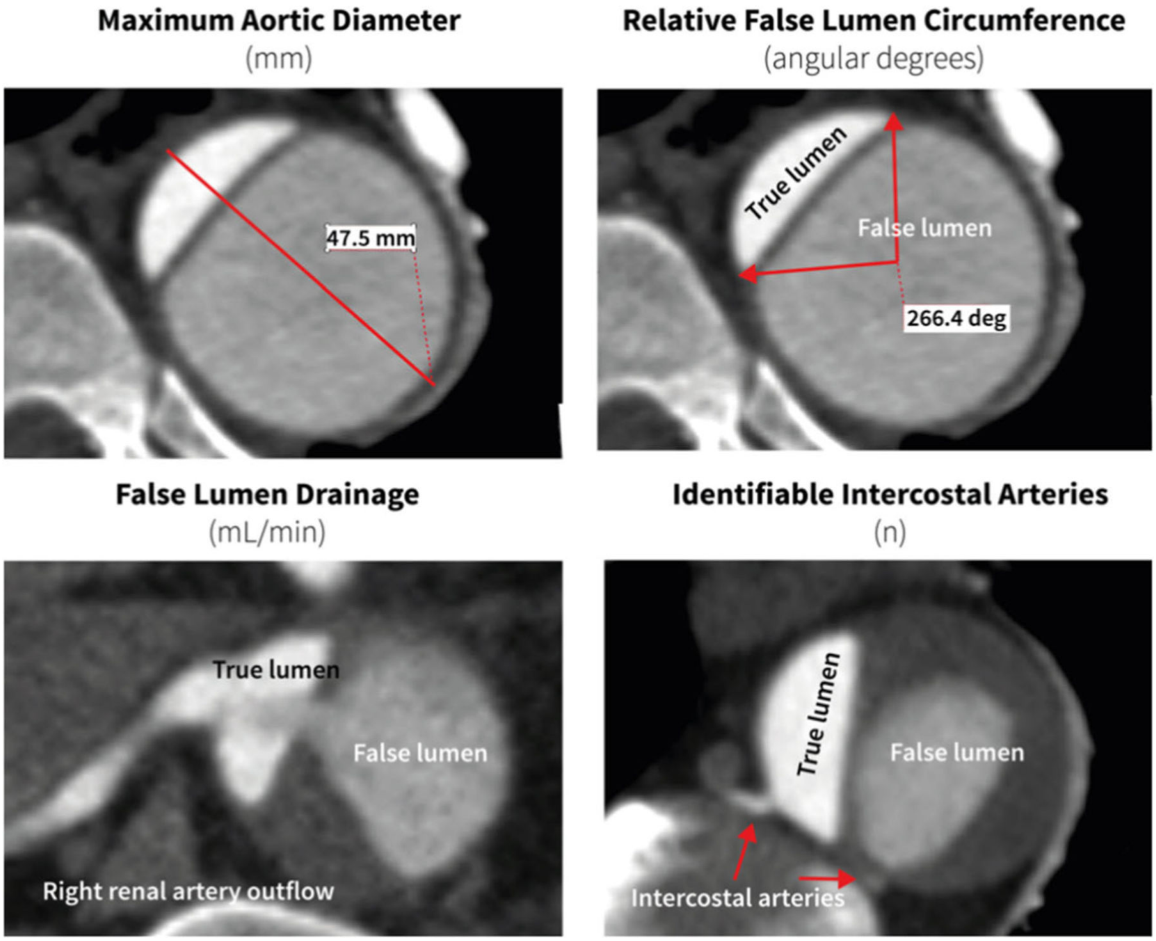

Fig. 1.

Assessed morphologic CT angiography imaging features

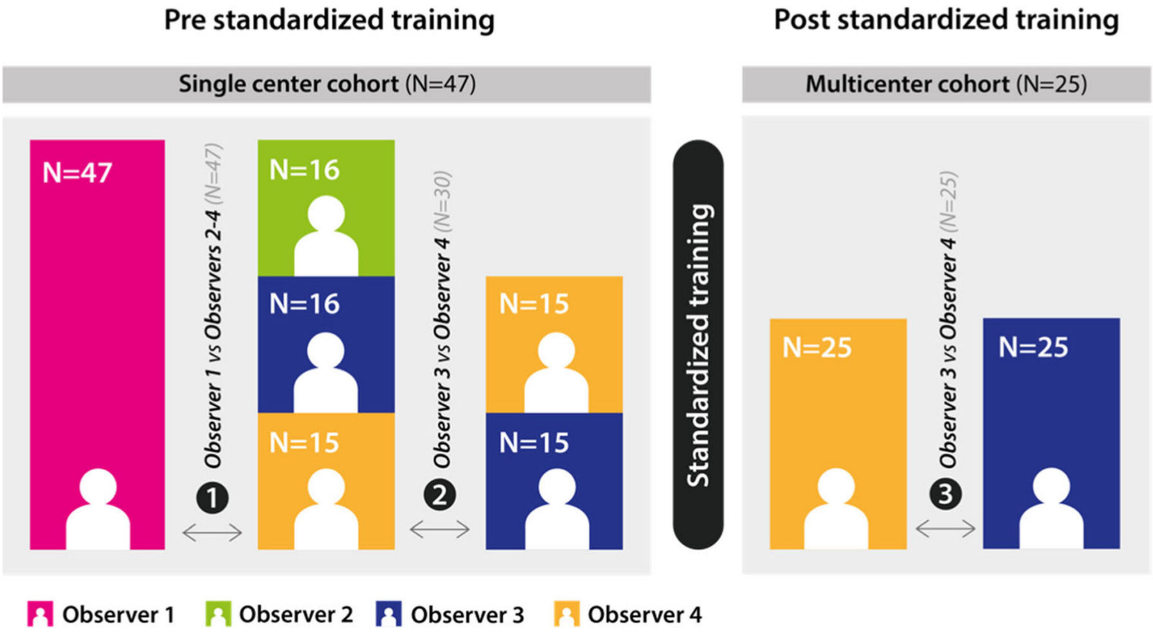
Pre standardized training
Post standardized training
Single center cohort (N=47)
Multicenter cohort (N=25)
N=47
Observer 1 vs Observers 2-4 (N=47)
1
N=16
N=16
N=15
Observer 3 vs Observer 4 (N=30)
2
N=15
N=15
Standardized training
N=25
Observer 3 vs Observer 4 (N=25)
3
N=25
Observer 1
Observer 2
Observer 3
Observer 4

Fig. 2.

Analysis flow diagram

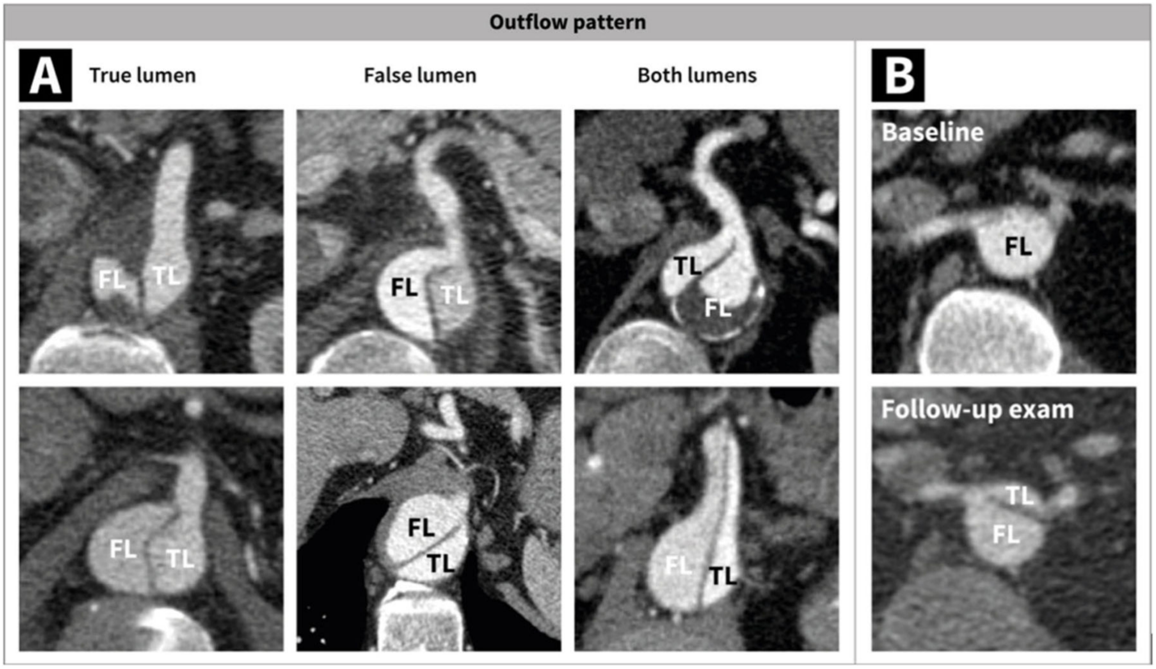

Fig. 3.

Examples of different outflow patterns as part of the standardized training guideline. Branch vessels can be supplied by the true lumen, the false lumen, or both lumens (**A**) In some cases, the outflow pattern is difficult to determine at baseline, while the follow-up exam is more clear (**B**). *FL*, false lumen; *TL*, true lumen

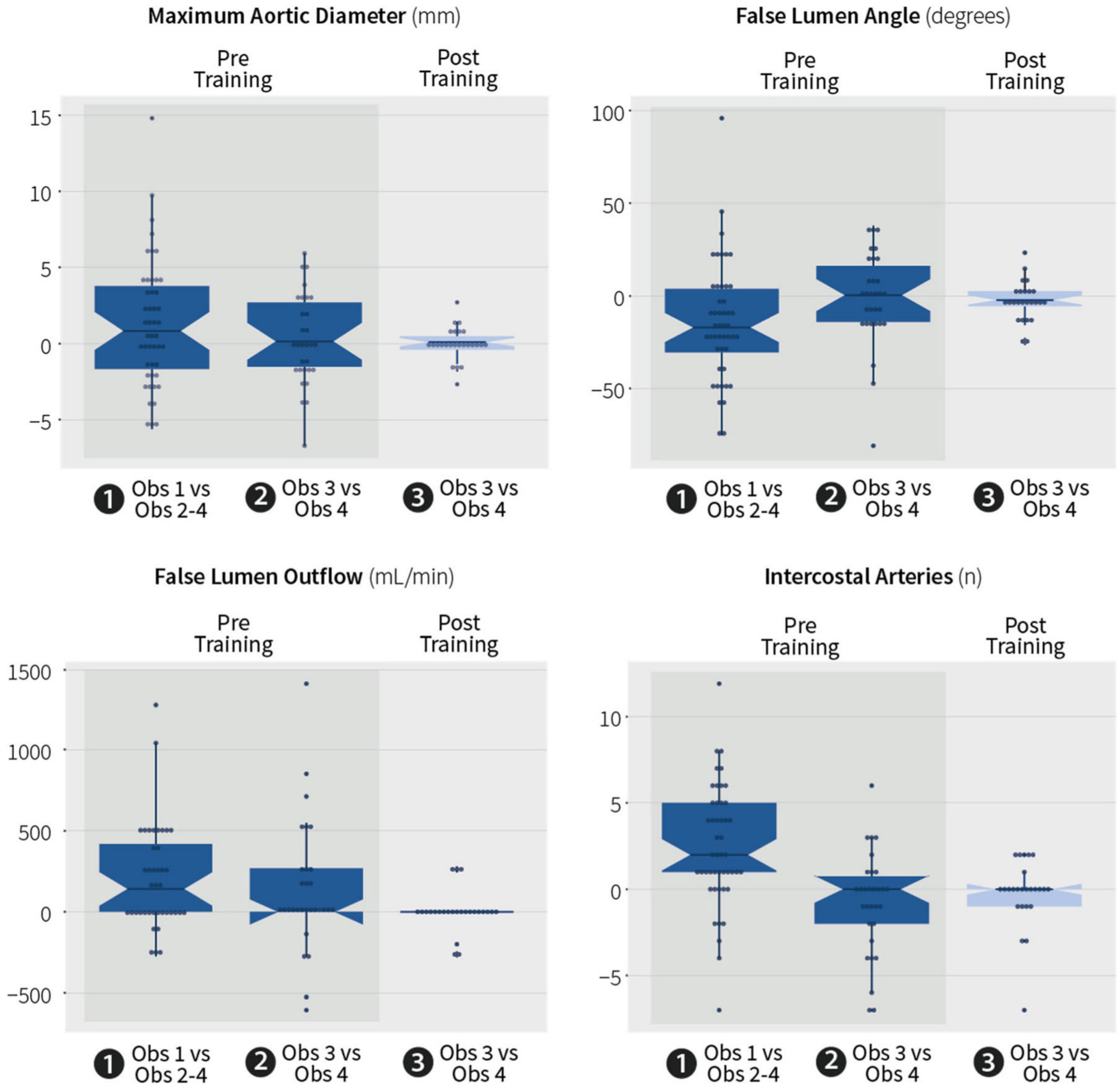

Fig. 4.

Pre- and post-training differences between observers for four morphologic imaging features.

Obs, observer

Table 1

Pre-training differences between observer 1 and observers 2–4 in an unconstrained clinical assessment

Pre-training (Experiment 1)		
	Observer 1 (47 CT exams)	**Mix of observers 2–4 (47 CT exams)**
Max aortic diameter (mm)	35.0 [33.0–39.5]	36.7 [33.1−39.9]
False lumen angle (degrees)	260.0 [250.0–270.0]	237.7 [221.6−261.5]
False lumen outflow (mL/min)	700 [488–938]	829 [633−1269]
Intercostal arteries (*n*)	13 [10–16]	16 [13−18]

Results listed as median (interquartiles)

Table 2

Pre-training differences between observers 3 and 4 in a constrained predefined standardized workflow

Pre-training (Experiment 2)		
	Observer 3 (30 CT exams)	**Observer 4 (30 CT exams)**
Max aortic diameter (mm)	36.6 [33.1−39.7]	35.8 [32.6−41.4]
False lumen angle (degrees)	249.0 [234.7−261.4]	250.3 [221.7−275.1]
False lumen outflow (mL/min)	810 [663−1086]	895 [650−1165]
Intercostal arteries (*n*)	16 [14−18]	17 [11−19]

Table 3

Post-training differences between observers 3 and 4 in a constrained predefined standardized workflow

Post-training (Experiment 3)		
	Observer 3 (25 CT exams)	**Observer 4 (25 CT exams)**
Max aortic diameter (mm)	41.6 [38.8−45.6]	40.9 [39.2−46.9]
False lumen angle (degrees)	244.2 [229.7−265.4]	249.0 [227.0−255.5]
False lumen outflow (mL/min)	400 [200−900]	500 [0−900]
Intercostal arteries (n)	15 [12−17]	16 [12−17]

Results listed as median (interquartiles)

Table 4

Risk reclassifications pre- and post-training

Pre-training												**Post-training**					
	21/47 (44.7%) reclassifications						***9/30 (30.0%) reclassifications***						***4/25 (16.0%) reclassifications***				
		Obs 1						Obs 3						Obs 3			
0		Low	Int	High				Low	Int	High				Low	Int	High	
Obs 2–4	Low	15	**7**	**6**	28	**Obs 4**	Low	17	**3**	**1**	21	**Obs 4**	Low	7	**2**	0	9
	Int	**4**	8	2	14		Int	**2**	2	0	4		Int	**1**	11	**1**	13
	High	0	**2**	3	5		High	**1**	**2**	2	5		High	0	0	3	3
		19	17	11	47			20	7	3	30			8	13	4	25

Int intermediate, *Obs* observer

Table 5

Agreement between observers

	Pre-training (Experiment 1) ***Unconstrained clinical assessment*** **Observer 1 vs. mix of observers 2–4 (47 CT exams)**	**Pre-training (Experiment 2)** ***Constrained predefined standardized workflow*** **Observer 3 vs. observer 4 (30 CT exams)**	**Post-training (Experiment 3)** ***Constrained predefined standardized workflow after training*** **Observer 3 vs. observer 4 (25 CT exams)**
Max aortic diameter (mm)	0.68 [0.49−0.81]	0.89 [0.78−0.95]	0.97 [0.94−0.99]
False lumen angle (degrees)	0.39 [0.12−0.61]	0.63 [0.35−0.80]	0.91 [0.82−0.96]
False lumen outflow (mL/min)	0.64 [0.30−0.82]	0.70 [0.46−0.85]	0.97 [0.92−0.98]
Intercostal arteries (*n*)	0.04 [−0.05−0.13]	0.50 [0.32−0.69]	0.69 [0.52−0.87]

Results listed as median (interquartiles). Unless otherwise specified, results are intraclass correlation coefficients (ICCs). Agreement for the number of intercostal arteries (categorical variable) was evaluated with linearly weighted kappa