Cardiothoracic Ratio Measurement Using Artificial Intelligence: Observer and Method Validation Studies

Pairash Saiviroonporn (pairash.sai@gmail.com)
Mahidol University
Kanchanaporn Rodbangyang
Mahidol University
Trongtum Tongdee
Mahidol University
Warasinee Chaisangmongkon
King Mongkut's University of Technology Thonburi
Pakorn Yodprom
Mahidol University
Thanogchai Siriapisith
Mahidol University
Suwimon Wonglaksanapimon
Mahidol University
Phakphoom Thiravit
Mahidol University

Research Article

Keywords: cardiothoracic ratio, deep learning, clinical validation, observer variation, AI

Posted Date: March 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-300251/v1

Abstract

Background

Artificial Intelligence (AI) technique for cardiothoracic ratio (CTR) measurement is a promising tool that has been technically validated but not clinically evaluated on a large dataset. This study observes and validates AI and manual methods for CTR measurement on a large dataset and investigates the clinical utility of the AI method.

Results

Five thousand normal chest x-rays and 2,517 images with cardiomegaly and CTR values, were analyzed using manual, AI-assisted, and AI only methods. AI methods obtained CTR values from a VGG-16 U-Net model. An in-house software was used to aid the study and to record measurement time. Intra and inter-observer experiments were performed on manual and AI-assisted methods and the average of each method was employed in a method variation study. AI outcomes were graded in the AI-assisted method as excellent (accepted by both users independently), good (required adjustment), and poor (failed outcome). Bland-Altman plot with coefficient of variation (CV), and coefficient of determination (R-squared) were employed to evaluate agreement and correlation between measurements. Finally, the performance of a cardiomegaly classification test was evaluated using a CTR cutoff at the standard (0.5), optimum, and maximum sensitivity. Manual CTR measurements on cardiomegaly data were comparable to the previous radiologist reports (CV of 2.13% *vs* 2.04%). The observer and method variations from the AI method were about three times higher than from the manual method (CV of 5.78% *vs* 2.13%). AI assistance resulted in 40% excellent, 56% good, and 4% poor grading. AI assistance significantly improved agreement on inter-observer measurement compared to manual methods (CV; bias: 1.72%; -0.61% *vs* 2.13%; -1.62%) and was faster to perform (2.2 ± 2.4 secs *vs* 10.6 ± 1.5 secs). R-squared and classification-test were not reliable indicators to verify that the AI method could replace manual operation.

Conclusion

AI alone is not suitable to replace manual operation due to its high variation, but it is useful to assist the radiologist because it can reduce observer variation and operation time. Agreement of measurement should be used to compare AI and manual methods, rather than R-square or classification performance tests.

Introduction

Chest radiography (CXR) is the most widely-used modality for screening of lung and heart diseases in clinical practice due to its easy accessibility and cost-effectiveness [1]. Cardiothoracic Ratio (CTR)

obtained from CXR is the preferred index to provide prognostic information on heart disease [1–4]. CTR is derived from a ratio of heart to internal thoracic diameters with a value of more than 0.5 considered as enlarged heart or cardiomegaly [2]. Calculation of CTR is usually performed manually which may introduce observer variation and is time consuming. Therefore, an automatic calculation of CTR could be a useful tool for clinicians and radiologists to improve accuracy and reduce workload.

Deep Learning (DL), subset of Artificial Intelligence (AI) methods, has demonstrated advance in medical imaging [5–8]. DL techniques have demonstrated excellent performance to classify CXR abnormalities [8], and to detect diabetic retinopathy in fundus images [6]. The technique has also been employed to automatically calculate CTR [9–12]. Presently, all DL techniques in CTR calculation are based on the U-Net model, the most successful convolutional network for biomedical image segmentation [13]. While DL techniques in CTR calculation have been technically validated, only two reports [9, 11] with small sample size (n = 100) were conducted in the clinical setting. Therefore, there is a need to validate this calculation technique on clinical perspective with big enough dataset before it can be implemented in real routine hospital setting.

We performed clinical validation using the DL technique from Chamveha *et al*. [10] on a large normal and cardiomegaly dataset. Observer and Method variations were evaluated to determine the measurement agreement between AI and manual methods. CTR values from our radiologist reports were employed as a reference to validate other methods in this study. We performed experiments on normal and cardiomegaly datasets to obtain AI performance as a whole, and separately only on cardiomegaly data to represent real clinical usage where CTR is only measured by a radiologist on suspicious of cardiomegaly. Finally, we explored the benefits of the AI method to assist the radiologist with CTR measurement.

Materials And Methods

Study population

This study was complied with the Declaration of Helsinki and approved by Siriraj Institutional Review Board (Si069/2020).
Informed consent was waived due to the retrospective nature of the study.
There were two data groups (normal and cardiomegaly) in the study. Data were acquired from chest x-ray radiologist reports between 2010–2019 from patients age greater than 17 years, and then their PA-upright CXR images were retrieved from the Picture Archiving Communication System (PACS) in our radiology department. Normal chest x-rays, and chest x-rays with cardiomegaly or enlarged heart size with CTR measurements were included. Five-thousand normal CXR images were randomly obtained and all 2,517 cardiomegaly images were acquired for a total of 7,517 images. The CTR values from radiologist reports were considered to be the reference method. Reference CTR values were only available for the cardiomegaly data because in our high patient volume clinical setting, radiologists measure CTR only on cases suspected of having cardiomegaly.

AI model

The Deep-Learning technique was based on U-Net with VGG-16 encoding [10, 14] modified from a vanilla U-Net [13], and trained on various public datasets [15–18] that had been manually segmented on lung and heart regions. The process automatically segmented heart and lung regions using the trained DL model and then find a spine line from the CXR image. These segmentations and line are then applied to find the heart and lung diameters and calculate the heart and lung ratio or CTR (Fig. 1B) on the normal and cardiomegaly data groups. These data were not part of a training or validation process of the DL model and so functioned to test the model [19]. A failed outcome was defined as the inability to segment lung or heart, or an unreasonable lung or heart size such as a heart size less than 3 mm.

Experimental setting

The study was designed to investigate observer and method variations of CTR measurements between manual and AI techniques, which were performed on both normal and cardiomegaly data groups (all data) and only on cardiomegaly group to emulate normal clinical practice at our hospital. Three CTR measurement methods were applied; manual, AI-assisted and AI only. A medical scientist with experience in medical image processing (PY) and a second-year radiology resident (KR) independently performed CTR measurements in the manual and AI-assisted methods with supervision from experienced chest radiologists (TT, TS, WS, and PT). The independent measurements were performed separately and two weeks apart on each dataset to reduce measurement bias. PY performed the measurement twice (intra-observer) and only once by KR, and the average of these three measurements on each method was used for method variation studies.

We developed a program using MATLAB software (R2019a, MathWorks, Inc., Natick, MA, USA) to assist the user operations as shown in Fig. 1A. The software provides graphical user interface for CTR measurement and records the user-interaction time of each measurement. In the manual method, users were presented with three lines of heart and chest borders in a default position (Fig. 1A) and asked to adjust these lines to the appropriate locations (Fig. 1B). In the AI-assisted method, the lines were positioned as suggested by the AI calculation and users could choose to accept them without further adjustment or disagree, which required adjustment of the lines. If there were any failure in the AI calculations, then the default line positions from the manual method were used. From user interaction in AI-assisted method, AI outcomes can be categorized into excellent, good, and poor categories. Any AI calculation failure was classified as poor and its data were excluded from the variation and correlation experiments. When AI outcomes were accepted by both users independently, it was classified as excellent. Finally, if any adjustment was required by the user then the outcome was considered to be good. The time of each case was measured from the start of line adjustment to acceptance (*i.e.*, hit the save button as in Fig. 1B).

Statistical analysis

Statistical analysis was performed on MATLAB and MedCalc (19.5.3, MedCalc software Ltd, Ostend, Belgium) software. The paired Student's t-test was used for parametric evaluation between measurement methods with the statistical significance level set at $p < 0.05$. Bland-Altman plot and linear correlation

were employed to evaluate agreement and correlation between measurement methods, respectively. Coefficient of variation (CV) signifying level of agreement was calculated from the standard deviation of the differences between two methods, then divided by their mean and expressed as a percentage. Thus, the lower the CV the better agreement was between two measurement methods. Coefficient of determination (R-Squared or R^2) was defined into four categories: poor (less than 0.5), moderate (0.5–0.75), good (0.75–0.9), and excellent (more than 0.9). Finally, the performance of the cardiomegaly classification test was evaluated using accuracy, sensitivity, specificity, area under receiver operating characteristics curve (AUC), and F1-score metrics on CTR cutoff values at 0.5 (the standard), the optimum (*i.e.*, maximize both sensitivity and specificity), and the maximum sensitivity (*i.e.*, to rule out cardiomegaly).

Results

Patient characteristics

There were 4,933 (1,431 males and 3,502 females; aged 42.5 ± 14.8 years) patients with normal CXRs, and 2,419 (675 males and 1,744 females; aged 64.2 ± 14.0 years) CXRs from patients with cardiomegaly. According to radiologist reports (reference method), the mean CTR value in the cardiomegaly group was 0.569 ± 0.047.

AI outcomes

With the AI-assisted method, 40% of outcomes were excellent, 56% were good, and 4% were poor (Fig. 2). Poor outcomes were mostly (97%; 290/299 cases) observed in the normal group, while only nine of 2,571 cardiomegaly cases had a poor outcome. Furthermore, most failures involved the heart segmentation calculation. Three examples of the poor outcome are displayed in Fig. 2J-2L, which also illustrates CTR measurements from the AI-assisted method.

Observer Variations

Intra- and inter-observer variations from manual and AI-assisted methods are presented in Fig. 3 and Table 1. Overall, the CV and bias of observer variations from both methods was lower than 2.2% and 1.7%, respectively, while the inter-observer variation of the manual method was 2.13%(CV) and − 1.62% (bias). The AI-assisted method significantly ($p < 0.001$) improved agreement on inter-observer measurements compared to the manual method (CV; bias: 1.72%; -0.61% *vs* 2.13%; -1.62%). Observer variations in the normal and the cardiomegaly group were comparable to that of both groups combined (data not shown). Therefore, the AI-assisted method increased observer agreement compared to the manual method.

Table 1
Bias, 95% CI, and coefficient of variation of intra- and inter-observer CTR measurements from Manual and AI-assisted methods on normal and cardiomegaly dataset.

Method	Intra-observer		Inter-observer	
	Bias (95% CI) (%)	CV(%)	Bias (95% CI) (%)	CV(%)
Manual	0.09 (-4.10 4.29)	1.51	-1.62 (-6.50 3.25)	2.13
AI-assisted	0.17 (-2.90 3.23)	1.10	-0.61 (-5.22 3.99)	1.72

Method Variations

CTR values from manual, AI, and AI-assisted methods were not significantly different on normal (0.455 ± 0.043, 0.447 ± 0.058, and 0.453 ± 0.044, respectively) and cardiomegaly (0.570 ± 0.045, 0.569 ± 0.049, and 0.570 ± 0.044, respectively) data. They also did not differ from the reference method (0.569 ± 0.047). Variations (CVs) from the reference method to manual and AI-assisted methods were in similar level to inter-observer variation of manual method (CVs of 2.04% and 2.23% *vs* 2.13, respectively). Our CTR measurements on cardiomegaly data, hence, were comparable to the previous reports by experience radiologists.

The CV of manual and AI methods (Fig. 4B and 4D, and Table 2), in contrast, were about three times higher, 5.78% and 5.61%, to from inter-observer variation in manual method on all, and cardiomegaly groups, respectively. Interestingly, even these two groups had similar high variation, their coefficient of determination were noticeable different: good (R^2 = 0.79) and poor (R^2 = 0.34) categories, as shown in Fig. 4A and 4C. CVs of manual and AI-assisted methods, on the other hands, were significantly lower than the inter-observer variation of manual method, 1.50% and 1.54% *vs* 2.13%, on all and cardiomegaly data groups, respectively, as displayed in Fig. 5B and 5D. Furthermore, unlike AI method, their coefficient of determinations was in a similar excellent category (R^2 > 0.9) as shown in Fig. 5A and 5C. AI-assisted method, therefore, had a similar variation as manual method while it was about three times higher in AI method. Furthermore, results from this study demonstrated that R-squared measurement was not a reliable indicator to verify AI method.

Table 2
Comparison of Bias, 95% CI, and coefficient of variation (CV) of CTR measurements.

Comparison	Normal and Cardiomegaly Data		Cardiomegaly Data	
	Bias (95% CI) (%)	CV(%)	Bias (95% CI) (%)	CV(%)
Manual vs **AI**	-0.93 (-15.0 13.14)	5.78	-0.09 (-13.30 13.12)	5.61
Manual vs **AI-assisted**	-0.08 (-4.22 4.07)	1.50	0.08 (-4.20 4.37)	1.54

The performances of cardiomegaly classification tests are presented in Table 3. All performance metrics from manual and AI-assisted methods were comparable on all cutoff points, and all were in the excellent

level (*e.g.*, AUC > 0.9 with all cutoff points around 0.5). The AI method gave similar outcomes only on the standard and optimum cutoffs, but provided a poor outcome on the cutoff point at the maximum sensitivity, (*e.g.*, accuracy of 34.8% with cutoff point around 0.2). The classification metrics were not reliable parameters to evaluate the performance of the AI method. For example, the AI only method had almost three times higher observer and method variations as compared to the manual method, but if standard or optimum cutoffs were used they would misleadingly suggest that the AI method also had the same excellent classification performance as the manual method. In contrast, if the cutoff for ruling out cardiomegaly (*i.e.*, cutoff at the maximum-sensitivity) was used then classification performance would be very poor. Thus, the cutoff criteria will dictate the outcome of the AI study in CTR measurement.

Table 3
Classification test on Manual, AI, and AI-assisted methods using cutoff points at the Standard (0.5), Optimum (maximum sensitivity and specificity), and maximum-sensitivity (Max-Sens).

Method	Test	CTR Cutoff	Sensitivity(%)	Specificity(%)	Accuracy(%)	F1	AUC
Manual	Standard	0.5	100	83.6	89.3	0.866	0.978
	Optimum	0.518	95.8	91.8	93.2	0.907	
	Max-Sens	0.505	100	86.0	90.9	0.884	
AI	Standard	0.5	97.5	82.8	87.9	0.849	0.962
	Optimum	0.521	90.3	91.0	90.8	0.872	
	Max-Sens	**0.219**	**100**	**0.10**	**34.8**	**0.516**	
AI-assisted	Standard	0.5	100	84.3	89.8	0.873	0.977
	Optimum	0.516	96.0	91.3	92.95	0.904	
	Max-Sens	0.501	100	85.0	90.2	0.876	

Measurement time

Average CTR measurement time for the manual method was about 10.6 ± 1.5 secs per case while it was almost five times faster (2.2 ± 2.4 secs) in the AI-assisted method on all data and on the cardiomegaly group alone (data not shown). From the AI-assisted method, the measurement times were 0 sec, 3.1 ± 1.8 secs, and 10.2 ± 1.3 for the excellent, good, and poor categories, respectively. The AI-assisted method then is almost five times faster to perform than manual method. Furthermore, if the data were selected from the AI and manual methods using the AI's excellent outcome criteria, then the CTR differences from such two methods will be in the range of ± 1.8% (calculated from 95% confidence interval of data). In other words, from data in this study, if the outcome from AI method differs from manual method by less than ±

1.8% then its outcome can be accepted without any further interaction from user (*i.e.*, an excellent category).

In summary, the observer and method variations from the AI method were about three times higher than from the manual method. CTR calculations from the AI method, however, are a very useful tool to assist the user, providing a better agreement and almost five times faster to perform. Furthermore, CV is a better parameter than R-squared or the classification performance test for the validation of AI in CTR measurement.

Discussions

CTR derived from CXR is a valuable index for evaluation of heart diseases, especially cardiomegaly [1–4]. To measure it, however, still requires manual operations that are user dependent and time consuming. Even with its usefulness, the measurement process is a burden in clinical practice. Recently, the AI method successfully provided automatic calculations of such an index and has been validated technically in various studies [9–12]. To use AI in the clinical setting, there is a need for clinical evaluation to assess the measurement agreement with manual method. However, there have been only two published pilot studies [9, 11] with small datasets that addressed this issue.

To our knowledge, this study was the first report of observer and method variations to validate CTR measurement using AI on a large dataset (n = 7,517). Using a modified U-Net deep-learning model (*i.e.*, 2D VGG-16 U-Net) for CTR calculation, AI was found to be not suitable to be used as an automated method for CTR measurement due to its high variations compared to the manual method. Its CTR calculations, on the other hand, can assist the user to obtain better results. Furthermore, the coefficient of determination (R^2) or classification performance test (*e.g.*, AUC) should not be employed because it may lead investigator to falsely conclude that the AI method can be employed as an automated method. Bland-Altman plot with Covariant of Variation (CV) parameters evaluated on a large data should be utilized instead to indicate agreement between these methods.

We found that the AI method can provide excellent outcomes in about 40% of the data, which is a desirable result for an automated method. However, there was about another 56% of good outcomes (*i.e.*, required adjustment by user) that needs improvement before the method can be used automatically. Most of the required adjustment was on heart diameter. Therefore, to aim for automated CTR measurement, the AI method needs to be improved on heart diameter calculation which is difficult to perform because its pixel value is low, and its edges are fused with the lung borders or thoracic spine [20]. In addition, the AI method also had about 4% failure rate (*i.e.*, poor outcome) most of which was in normal group (97%: 290/299). In routine clinical usage, which measures CTR only on suspicious cardiomegaly cases, thus, this failure is infrequent (9 failures in 2,517 cardiomegaly data). Nevertheless, most of the segmentation failure was on hearts with quite short diameters (*e.g.*, Fig. 2J). This may be due to an inadequate presentation of such heart data shape in the training dataset. Fine tuning the model

(retrain model on previous weighting data) with such heart shape dataset from local data should further reduce such failures.

We found that the AI-assisted method had lower inter-observer bias and variation than the manual method (CV and bias: 1.72% *vs* 2.13% and − 0.61 *vs* -1.62). This may be due to AI's excellent outcome in about 40% of data which can help to improve measurement agreement. Furthermore, it is almost five fold faster to perform than using the manual method, and increase F1 from 0.866 to 0.872 at the standard CTR cutoff point of 0.5. This is clearly demonstrated the usefulness of AI method to assist the CTR measurement. Our AI-assisted time performance was also in agreement with a recent study by Bercean *et al*. [9] which found a similar magnitude of time reduction (22.5 *vs* 5.1 secs, or 4.4 times). Even on a small dataset (n = 200), that study also found that the model-assisted method can improve individual radiologist's cardiomegaly F1 score (0.845 to 0.851) compared to the manual method.

We concluded that the classification performance test of AI method was not better than from the manual method, a finding at odds with a report by Li *et al*. [11] which found that the sensitivity and negative-predictive values of the AI method was significantly better than manual method. This may be due to two factors. First, the performance of deep learning algorithms in automated CTR measurement tasks depends on their ability to correctly locate heart and lung boundaries. In Li *et al*. [11], algorithms may have achieved more precise anatomical segmentations, although the authors did not provide precision metrics on an open dataset for comparison with the model we used [10]. Second, the algorithm in Li *et al*. [11] was trained and tested on the same dataset, while the model used in this paper was trained on an open dataset, and tested in an out-of-sample fashion. It would be useful to validate their finding by performing the classification test using their model on our dataset.

CTR measured from manual and AI-assisted methods were in substantial agreement with the reference method (CVs of 2.0 and 2.2%, respectively). The AI method, in contrast, had almost three times higher CVs on all comparisons. This strongly suggests that the AI method is not yet suitable to be employed as an automated method. However, its R^2 of all data (normal and cardiomegaly groups) and classification performance test at the standard or optimum cutoffs were in similar outcomes as other methods. This is because R^2 measures linear association rather than agreement of data [21, 22] and measurements with highly correlated data may have poor agreement [22], as in our case. Furthermore, the correlation typically depends on the range of measure. This is why the R^2 of the manual and AI methods was good in normal and cardiomegaly groups (R^2 = 0.79; CTR data range = 0.35–0.85), but poorly correlated in the cardiomegaly group alone (R^2 = 0.34; CTR data range = 0.52–0.85) (Fig. 4A and 4C). On the other hand, if the agreement measurement, like Bland-Altman plot and CV, presents with good agreement (Fig. 5B and 5D) then they will surely be highly correlated [22], as shown in Fig. 5A and 5C. Thus, the agreement measurement should be employed to evaluate the compatibility of AI to manual method in CTR measurement study.

Classification performance tests may also be misleading because they only provide information on the performance of normal and cardiomegaly groups, and not how the methods agree. For example, Fig. 2D

and 2F present cases where the AI method gave a false positive and negative result, respectively. These two data have an effect on classification test but most of the AI data did not have this effect (*i.e.*, AI's CTR data did not change the classification) as shown in Fig. 2E and Table 3. Still, we obtained excellent classification performance at the standard CTR cutoff (*e.g.*, AUC = 0.902). However, if this AI method were employed to rule out cardiomegaly patients (*i.e.*, using CTR cutoff at the maximum sensitivity), then the method would perform poorly (*e.g.*, accuracy of 34.8%) and should not replace the manual approach. Test agreement is the necessary for evaluation of the AI method if it were to be implemented as an automated method, and its agreement should be comparable to the manual method (CV = 2.1%).

We performed observer and method variation tests on a large dataset using only a modified U-Net Deep-Learning model because we wished to obtain baseline AI performance data. Our results, especially the manual measurement of 7,517 CXRs, will serve as a reference to evaluate other state-of-the-art AI models [23]. Our plan is to test these models on our dataset and accept the AI outcome only if it differs from our manual results by less than ± 1.8% (*i.e.*, an excellent category where the user can accept its outcome without adjustment). Any model with > 70% acceptance rate will be studied prospectively in a clinical setting and evaluated by our radiologists. Furthermore, at such acceptance rate, we will perform another retrospective study in our PACS data (around one million CXR images). Such a pioneering study would provide more insight into CTR values and useful information for clinicians.

There were some limitations in our dataset and methods. We used only normal and cardiomegaly data and there was no data from other pathologies, such as the fat pad of the pericardium or pleural effusion. These pathologic conditions may limit the DL model's ability to segment heart and lung, and may lower the performance of CTR measurement. Such data should be included in the future studies to better evaluate the performance of the model. Furthermore, we only investigated adult cases; evaluation of CTR measurement by AI in pediatric cases is needed. Next, we used only a publicly available dataset. Future studies using local datasets are needed to improve the model's performance. Finally, unlike most deep learning for CXR analysis studies, this study did not address the question of how AI can be trained to match human performance in CTR measurement, but focused on assessing the extent to which deep learning methods can benefit the radiologists' practice in a clinical setting. Future studies may focus more on the patterns of errors generated by the algorithms and suggest ways to improve its accuracy.

Conclusion

We conclude that AI should be employed to assist radiologists to perform CTR measurement because it can significantly reduce variations and is almost five-fold faster than the manual method. However, AI alone is not yet suitable for the measurement due to its high variations. Agreement of measurement, like the Bland-Altman plot and CV should be used to evaluate the comparability of AI to the manual method, while the coefficient of determination or classification performance test should be used with caution because it is not a reliable indicator.

Abbreviations

Abbreviation	Term
AI	Artificial Intelligence
CTR	CardioThoracic Ratio
CV	Coefficient of Variation
R-squared	Coefficient of Determination
CXR	Chest Radiography
DL	Deep Learning
PACS	Picture Archiving Communication System

Declarations

Ethics approval and consent to participate

This study was complied with the Declaration of Helsinki and approved by Siriraj Institutional Review Board (Si069/2020). Informed consent was waived due to the retrospective nature of the study according to the board policy.

Consent for publication

Not Applicable

Availability of data and materials

The datasets generated during and/or analyzed during the current study are not publicly available due to patients' data privacy policy by our hospital but are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no personal or professional conflicts of interest regarding any aspect of this study.

Funding disclosure

This study was supported by a Chalermphrakiat Grant (PS, TT, TS, and SW) from the Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.

Authors' contributions

PS was the principal investigator and participated in the design of the study, analyzed results, and drafted and revised the manuscript. KR and PY performed cardiothoracic ratio experiments and analyzed results.

KR also helped prepare data for manuscript. TT formulated the study design and supervised the experiment. WC provided technical and methodology support for the Deep Learning model. TS, SW, and PT supervised the experiment. All authors read and approved the final draft of the manuscript.

Acknowledgements

The authors gratefully acknowledge technical support from Perceptra Co, Ltd on U-Net with VGG-16 encoding used in this study

References

1. Hubbell FA, Greenfield S, Tyler JL, et al. The impact of routine admission chest x-ray films on patient care. N Engl J Med. 1985;312(4):209-13.
2. Danzer CS. The Cardiothoracic Ratio: An Index of Cardiac Enlargement. Am J Med Sci. 1919;157(4):157513- 521.
3. Dimopoulos K, Giannakoulas G, Bendayan I, et al. Cardiothoracic ratio from postero-anterior chest radiographs: a simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. Int J Cardiol. 2013;166(2):453-7.
4. Kearney MT, Fox KA, Lee AJ, et al. Predicting death due to progressive heart failure in patients with mild-to-moderate chronic heart failure. J Am Coll Cardiol. 2002;40(10):1801-8.
5. Biswas M, Kuppili V, Saba L, et al. State-of-the-art review on deep learning in medical imaging. Front Biosci (Landmark Ed). 2019;24:392-426.
6. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. Eye (Lond). 2020;34(3):451-60.
7. Lee JG, Jun S, Cho YW, et al. Deep Learning in Medical Imaging: General Overview. Korean J Radiol. 2017;18(4):570-84.
8. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med. 2018;15(11):e1002686.
9. Bercean B, Iarca S, Tenescu A, et al., editors. Assisting Radiologists Through Automatic Cardiothoracic Ratio Calculation. 2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI); 2020 21-23 May 2020.
10. Chamveha I, Promwiset T, Tongdee T, et al. Automated Cardiothoracic Ratio Calculation and Cardiomegaly Detection Using Deep Learning Approach. ArXiv. 2020:1-11.
11. Li Z, Hou Z, Chen C, et al. Automatic Cardiothoracic Ratio Calculation With Deep Learning. IEEE Access. 2019;7:37749-56.
12. Que Q, Tang Z, Wang R, et al. CardioXNet: Automated Detection for Cardiomegaly Based on Deep Learning. Annu Int Conf IEEE Eng Med Biol Soc. 2018;2018:612-5.
13. Ronneberger O, Fischer P, Brox T. U-net: Convolutional Networks for Biomedical Image Segmentation. MICCAI 2015; 18 May 2015; Munich, Germany: ArXiv; 2015.

14. Balakrishna C, Dadashzadeh S, Soltaninejad S. Automatic detection of lumen and media in the IVUS images using U-Net with VGG16 Encoder2018.
15. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33:590-7.
16. Shiraishi J, Katsuragawa S, Ikezoe J, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR Am J Roentgenol. 2000;174(1):71-4.
17. Wang X, Peng Y, Lu L, et al., editors. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 21-26 July 2017.
18. Xiaosong W, Yifan P, Le L, et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE Conference on Computer Vision and Pattern Recognition. 2017:2097-106.
19. Cicero M, Bilbily A, Colak E, et al. Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs. Investigative Radiology. 2016;52(5):281-7.
20. Arsalan M, Owais M, Mahmood T, et al. Artificial Intelligence-Based Diagnosis of Cardiac and Related Diseases. J Clin Med. 2020;9(3).
21. Bunce C. Correlation, Agreement, and Bland–Altman Analysis: Statistical Analysis of Method Comparison Studies. American Journal of Ophthalmology. 2009;148(1):4-6.
22. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Measures of agreement. Perspect Clin Res. 2017;8(4):187-91.
23. Lei T, Wang R, Wan Y, et al. Medical Image Segmentation Using Deep Learning: A Survey2020.

Figures

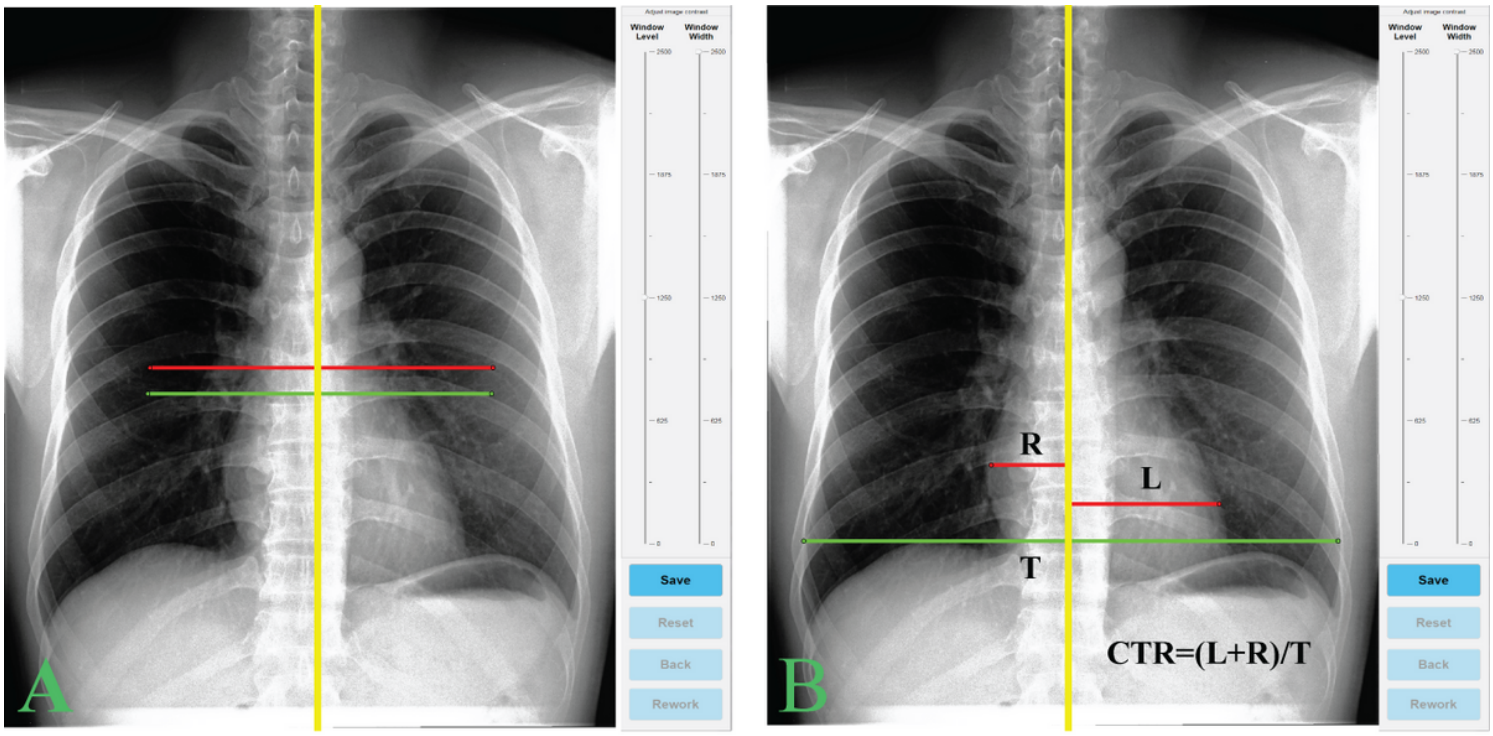

Figure 1

An in-house software for CTR measurement. The green and red lines are heart and chest border lines at default (A) and user-adjusted (B) positions, and the spine line is yellow. CTR was calculated from the ratio of these heart and chest lines (B).

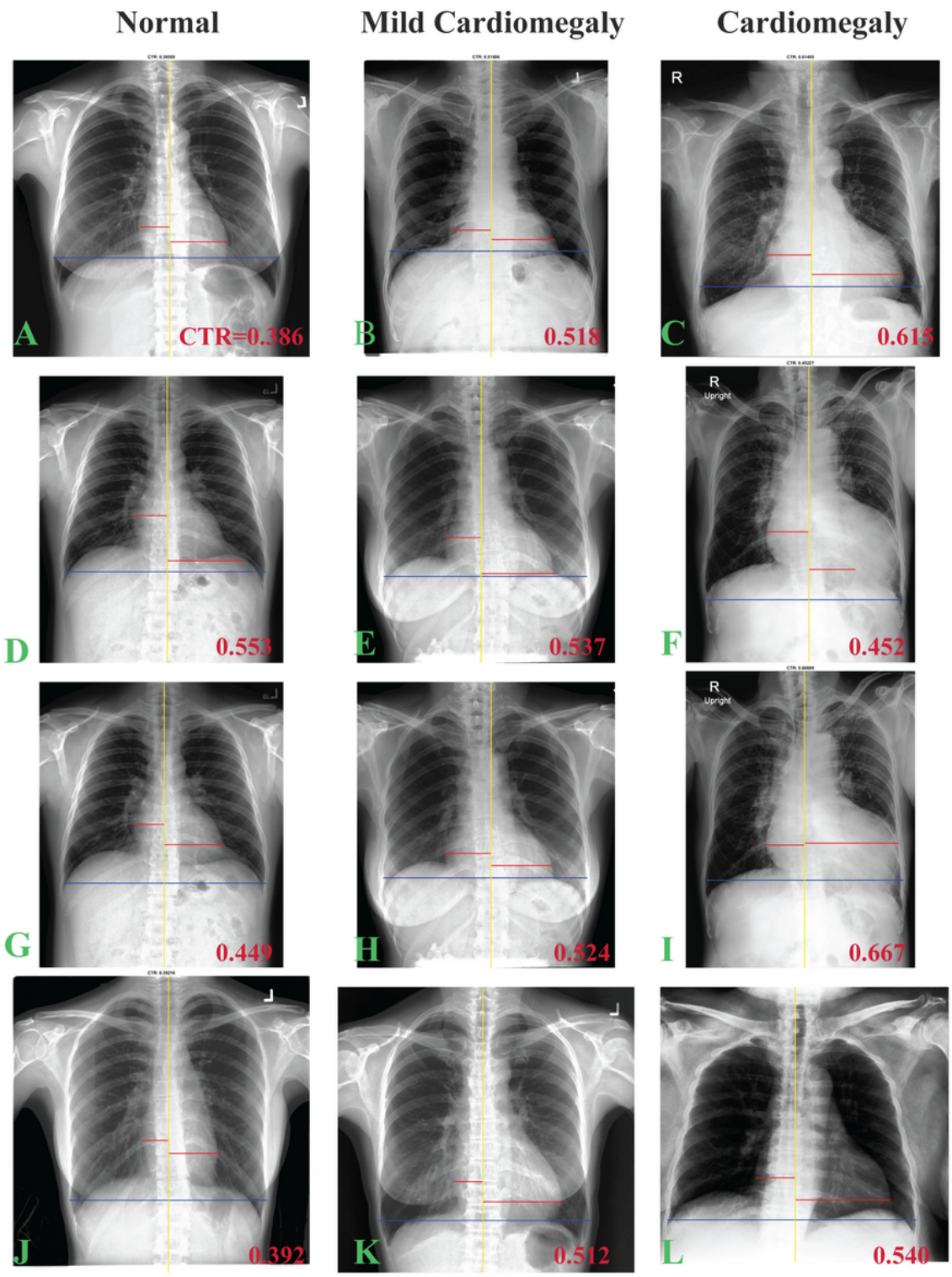

Figure 2

CTR measurements on normal, mild cardiomegaly and cardiomegaly cases using AI (the first and second rows) and manual (the third and fourth rows) methods. The first (A-C) and second (D-F) rows are CTR measurements by AI which accepted and rejected by user, respectively. The third row (G-I) is the user adjustment of the rejected AI operation (the second row) while the last row (J-L) demonstrates the failed

cases from AI operation required fully manual operation. CTR value is displayed at the lower right corner of each image.

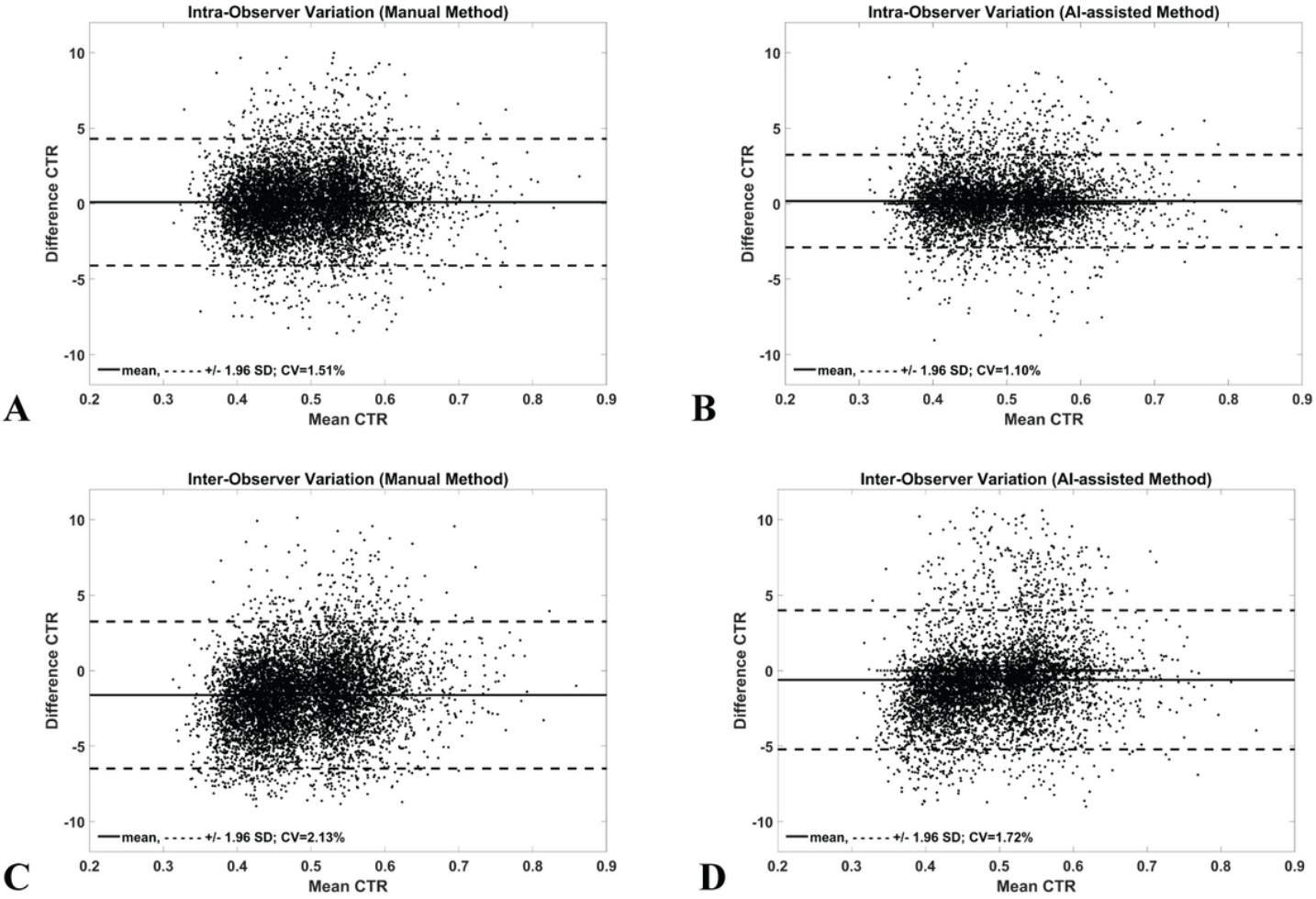

Figure 3

Bland-Altman plots of Manual (A and C) and AI-assisted (B and D) methods. Note: AI-assisted method had lower bias and CV as compared to Manual method on both intra- and inter-observer variation studies.

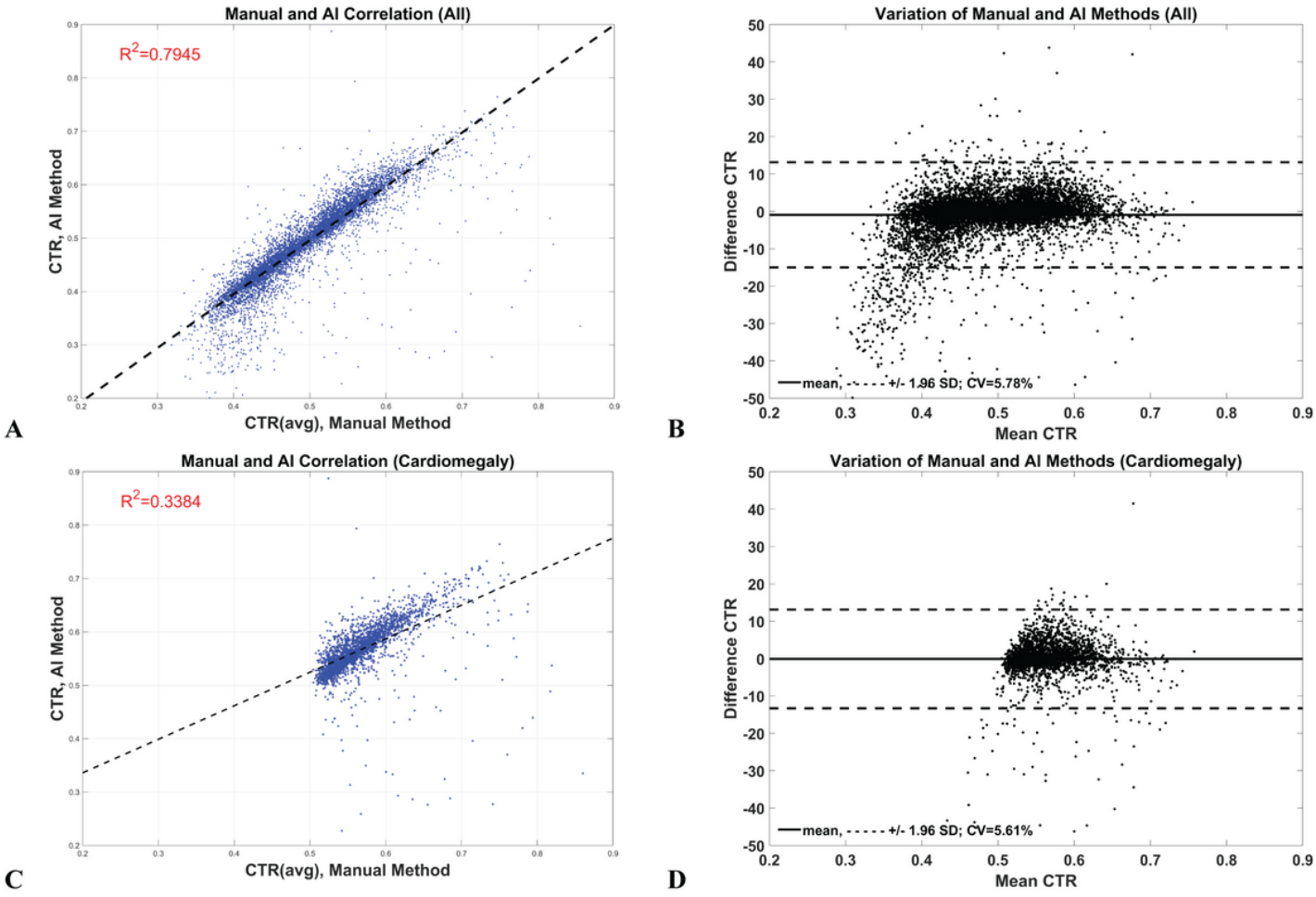

Figure 4

Linear correlation (A and C) and Bland-Altman (B and D) plots of Manual and AI method on all (the first row) and cardiomegaly (the last row) data. Note: Even these comparisons had high variation on both types of data, their R-squared were interestingly different as in good (0.7945) and poor (0.3384) in all and cardiomegaly data, respectively.

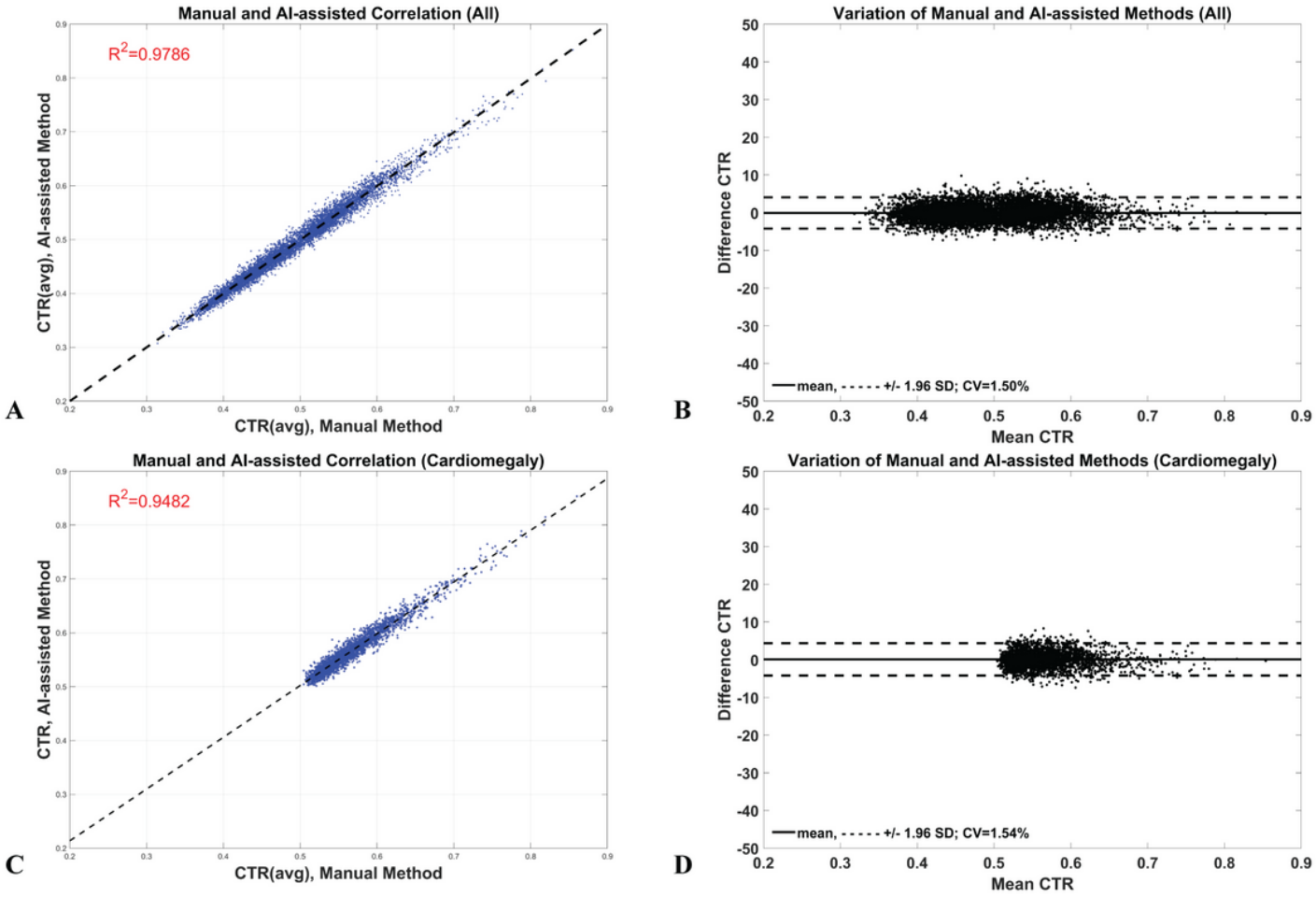

Figure 5

Linear correlation (A and C) and Bland-Altman (B and D) plots of Manual and AI-assisted method on all (the first row) and cardiomegaly (the last row) data. Note: These comparisons had low variation on both types of data and unlike from Manual and AI method (Figure 4), their R-squared were consistency at good category (0.978 and 0.9482).