RESEARCH

Reliability of a computer-aided system in the evaluation of indeterminate ultrasound images of thyroid nodules

J L Reverter[1,2], **L Ferrer-Estopiñan**[1,2], **F Vázquez**[1,2], **S Ballesta**[1,2], **S Batule**[1,2], **A Perez-Montes de Oca**[1,2], **C Puig-Jové**[1,2] and **M Puig-Domingo**[1,2]

[1]Endocrinology and Nutrition Service, Germans Trias i Pujol Hospital and Research Institute, Badalona, Spain
[2]Department of Medicine, Autonomous University of Barcelona, Barcelona, Spain

Correspondence should be addressed to J L Reverter: reverter.germanstrias@gencat.cat

Abstract

Introduction: Computer-aided diagnostic (CAD) programs for malignancy risk stratification from ultrasound (US) imaging of thyroid nodules are being validated both experimentally and in real-world practice. However, they have not been tested for reliability in analyzing difficult or unclear images.
Methods: US images with indeterminate characteristics were evaluated by five observers with different experience in US examination and by a commercial CAD program. The nodules, on which the observers widely agreed, were considered concordant and, if there was little agreement, not concordant or difficult to assess. The diagnostic performance of the readers and the CAD program was calculated and compared in both groups of nodule images.
Results: In the group of concordant thyroid nodules (n = 37), the clinicians and the CAD system obtained similar levels of accuracy (77.0% vs 74.2%, respectively; P = 0.7) and no differences were found in sensitivity (SEN) (95.0% vs 87.5%, P = 0.2), specificity (SPE) (45.5 vs 49.4, respectively; P = 0.7), positive predictive value (PPV) (75.2% vs 77.7%, respectively; P = 0.8), nor negative predictive value (NPV) (85.6 vs 77.7, respectively; P = 0.3). When analyzing the non-concordant nodules (n = 43), the CAD system presented a decrease in accuracy of 4.2%, which was significantly lower than that observed by the experts (19.9%, P = 0.02).
Conclusions: Clinical observers are similar to the CAD system in the US assessment of the risk of thyroid nodules. However, the AI system for thyroid nodules AmCAD-UT® showed more reliability in the analysis of unclear or misleading images.

Key Words
- thyroid nodules
- CAD system
- risk classification

Introduction

The incidence of thyroid nodules, which affect up to 60% of the general population in certain countries, continues to increase (1). Considering that about 5% prove to be malignant, this high and increasing figure of thyroid nodules has enhanced the need for endocrinologists to be able to define the risk of malignancy as accurately as possible (2). Ultrasound (US) is the first-line method to identify malignant thyroid nodules with the advantages of accessibility, cost-effectiveness, and no radiation exposure (3). US suspicious findings are well-established.

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0023

Features such as the presence of microcalcifications, irregular or spiculated margins, absence of halo, marked hypoechogenicity with mostly solid composition, and a taller-than-wider shape have been associated with an increased risk of malignancy. On the other hand, a round shape, isoechogenicity, presence of peripheral vascularity, a spongiform appearance, smooth margins, and a cystic composition are characteristics suggesting benign lesions (4). Based on these characteristics, US risk stratification systems (RSS) have been designed to standardize reports and to make the clinical decision-making process easier (5). Nevertheless, the presence of interoperator variation is inevitable and the diagnostic performance of US is highly related with the experience of the clinician who performs the image acquisition (6). To improve the diagnostic accuracy and efficiency, machine learning-based computer-aided diagnosis (CAD) systems are being introduced in the diagnosis process with the aim of achieving a more objective interpretation of the sonographic features due to the robustness of the computational analysis of US lesion characteristics. Recently, our group and others have assessed the performance of CAD systems in the evaluation of thyroid lesions and concluded that diagnostic performance is comparable to that of experienced observers (7, 8, 9, 10). In order to maximize the advantages of a CAD system, it is important that the images have a high-quality definition. This is a pre-requisite to make a correct assessment of the characteristics of a suspicious lesion, but despite having good quality US images, there are nodules that are difficult to categorize.

The aim of this study was to evaluate the reliability and consistency of a CAD system in the evaluation of US images of thyroid nodules that are difficult to analyze and that their nature was confirmed by histologic examination.

Materials and methods

This was a cross-sectional and retrospective study conducted following the Declaration of Helsinki and approved by the institutional review board and ethics committee at Germans Trias i Pujol Hospital in Badalona, Spain. In this institution, all patients undergoing an US examination provide their written consent for the subsequent use of their clinical and radiologic data, as well as any surplus tissue samples for research purposes.

Images of thyroid nodules were obtained from image archives of thyroid US studies previously conducted by our clinical and research group from January 2018 to December 2019. These images had not been included in the cohort of previous studies. The medical records and US images from the institutional electronic clinical history of patients with thyroid nodules who had undergone thyroidectomy and who had a pathologic diagnosis of benign or malignant thyroid disease were reviewed. The following inclusion criteria used in patient selection were: age older than 18 years at the time of diagnosis, having undergone total or nearly total thyroidectomy or lobectomy based on cytological data, nodule size and symptoms, having undergone preoperative US evaluation of thyroid nodules, and to have an available pathologic sample. These images were reviewed by a senior endocrinologist with long experience in US-based thyroid diagnostic and therapeutic procedures and those images showing confusing, mixed benign, and pathologic components, thus falling into a subjective category of 'difficult-to-interpret' or indeterminate, were included in the study. Most of the features contributing to this category included a variable degree of hypogenicity, the differentiation of microcalcifications vs colloid artifacts, apparently spongiform vs mixed nodules or the degree of definition of the lesion.

Computer-assisted evaluation

Each nodule image was analyzed using the AmCAD-UT software, with the program blinded to patient data. AmCAD-UT® is a software application that is designed to facilitate the detection, visualization, and characterization of thyroid nodule features in sonographic images using artificial intelligence and computational vision recognition and quantification algorithms. The program evaluates nodule shape, margins, and anechoic and hyperechoic areas, and then, applying seven scoring systems, it analyzes the risk of malignancy based on the quantified results obtained. The program performs the analyses from an external USB drive device and is easy to operate through a user-friendly interface. AmCAD-UT® can work with either JPEG or DICOM images, though in the present work, as noted above, the former was used. For each of the images used in the study, the program analyzed nodule size and all related features and then classified them for risk of malignancy according to TI-RADS classification system (11).

Study design

The study included images that presented indeterminate mixed criteria selected from a set of 300 images from our institutional repository. Sonographic images were acquired for each patient by an experienced endocrinologist (JLR)

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0023

using a US device equipped with a 5- to 15-MHz linear transducer (General Electric Logiq E9®, GE Healthcare) prior to the surgical procedure. The area of the US image featuring the nodule was then separated from the rest of the thyroid image and then stored as a specific JPEG file to which a random code number was assigned. These images were analyzed for the risk of malignancy based on TI-RADS classification by five endocrinologists with 1–5 years expertise in US thyroid explorations, and by the AmCAD-UT® system. The clinical evaluators were blinded to histologic information at the time of the imaging study performance. Subsequently, the images in which three or more observers agreed in the diagnosis were considered as the set of 'concordant' images and were used as a control group, while those in which two or fewer observers agreed were considered as 'not concordant' and were used as a problem group. The selected images were included as they followed into the categories of maximal agreement ($n=37$, control group) between clinical observers or maximal disagreement ($n=43$, problem group). After selection and risk assessment, confirmation of the nature of the lesions was performed by reviewing the cytological and histological reports. Study design flowchart is presented in Fig. 1.

Statistical analysis

Descriptive statistics were applied to all collected variables expressed as frequency tables for categorical data or mean values ± S.D. for continuous data. The χ^2 or Fisher's exact tests were used to compare differences between groups in terms of categorical variables and Student's *t*-test was used to compare quantitative variables. Nodule features and suspicion of malignancy as reported by the observers and the AmCAD-UT® program were compared with the histological diagnosis results. Sensitivity (SEN), specificity (SPE), positive predictive value (PPV), and negative predictive value (NPV) were calculated using the diagnostic test 2 × 2 contingency tables and for the comparisons the McNemar test was used. The area under the receiver operating characteristic curve (ROC) was also generated, and the area under the curve (AUC) was calculated to determine diagnostic performance. For statistical analysis, nodules were classified as 'likely benign' (TI-RADS <4) or 'likely malignant' (TI-RADS ≥4). A P value ≤0.05 was considered statistically significant. We used 80 images according to the sample calculation, considering an expected probability of agreement among human observers of 0.6 (obtained from a pilot study with similar samples not included in the main study), and an expected probability of agreement of 0.8 on the CAD method, with a bilateral alpha risk of 0.05 and a beta risk of 0.2 without dropouts. Statistical analysis was performed using a software package (IBM SPSS Statistics 24; IBM Corp.).

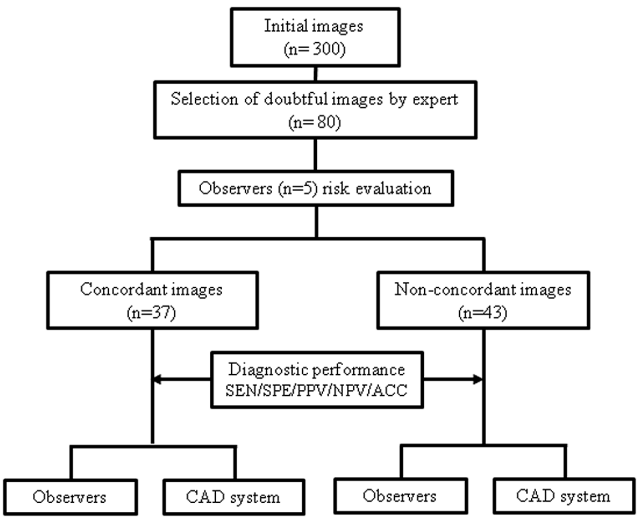

Figure 1
Study design flowchart (see description in the text). SEN: sensitivity, SPE: specificity, PPV: positive predictive value, NPV: negative predictive value, ACC: accuracy.

Results

Demographic, analytical, cytological, and histopathological characteristics

The study included a total of 80 thyroid undefined nodules from 80 patients (82% women). The mean age of patients was 56 ± 13 years. The mean maximum nodule diameter was 2.9 ± 0.2 cm for benign nodules and 3.3 ± 0.9 cm for malignant nodules (P < 0.01). No differences in thyroid stimulating hormone (TSH) levels were observed between groups (2.7 ± 1.2 mU/mL vs 2.8 ± 0.9 mU/mL, respectively; P= 0.6). Cytological diagnoses were as follows: 93% of benign nodules were Bethesda category II and 7% were category III, whereas 9% of malignant nodules were Bethesda category III, 84% were category IV, and 7% were category V. The benign histopathologic diagnoses were 80% nodular hyperplasia and 20% follicular adenoma, and the malignant lesions were 85% papillary thyroid carcinoma and 15% follicular thyroid cancer. No differences were found in the proportion of malignant nodules between concordant (control group) and non-concordant (problem group) images (25% vs 38%, P= 0.2).

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0023

Diagnostic performance

Table 1 shows the results obtained for SEN, SPE, PPV, NPV, and accuracy by the clinicians in comparison with the classification of risk offered by the CAD program in the two sets of images.

Concordant nodule images

In the group of thyroid nodules in which the clinicians and the software widely agreed (control group), both readers and the CAD system obtained similar levels of accuracy (77.0% vs 74.2%, respectively; P= 0.7), No differences were found between the clinicians and the CAD system regarding SEN (95.0% vs 87.5%, P= 0.2), SPE (45.5 vs 49.4, respectively; P= 0.7), PPV (75.2% vs 77.7%, respectively; P= 0.8), and NPV (85.6 vs 77.7, respectively; P= 0.3).

Non-concordant nodule images

In the problem group with high disagreement between clinical observers, the diagnostic performance of both the CAD system and reader groups decreased. This decline in accuracy was significantly more pronounced in clinical observers than with the use of the AmCAD-UT® system (19.9% vs 4.2% reduction, P= 0.02). The CAD program demonstrated a significant better accuracy in comparison to clinical observers in the problem group (70.0% vs 57.1%, respectively; P= 0.02). In addition, there was no statistical reduction of accuracy for CAD (74.2% vs 70%; P= 0.6) when concordant and non-concordant images were analyzed, while in the clinical observer group this change was statistically significant (77.0 vs 57.1; P= 0.05) (Table 1). The SPE and NPV values obtained in this group of nodules in which clinician readers do not agree and therefore could be considered difficult to assess were not significantly different in the CAD system compared to clinical readers, whereas the CAD system showed better values in SEN and PPV (Table 1).

By performing ROC analysis, AUC value obtained in the group of concordant nodules for the risk categories defined by the clinical observers was 0.81 and in the non-concordant nodules 0.69. The respective AUC values for the automated system were 0.72 and 0.70.

In the analysis of the images, those that presented the highest degree of disagreement between the clinicians and the CAD system were those that presented the poorer definition and higher echoic heterogeneity and the difficulty of defining margins and microcalcifications. The images in which the concordance was greater but the risk classification was erroneous were mostly isoechoic with poor definition of the nodule limits. Figure 2 shows the US characteristics and the risk classification analyzed by the CAD system in three nodules of the study in which the final biopsy results were in agreement with the risk pattern.

Discussion

In our study, the CAD system for risk of malignancy classification of thyroid nodules based in US images showed similar diagnostic performance to those of clinicians in images where the degree of agreement between clinicians is higher. But when these images may be considered difficult to evaluate and the judgment of the clinicians disagrees, the CAD system obtained greater accuracy. As far as we know, this is the first study evaluating the reliability of an IA-based system in the risk stratification of nodule images that present undefined characteristics or low definition, a situation that is not infrequent in clinical practice. In our institutional series, this corresponds to a 10–25% of all thyroid nodule explorations. The evaluation of these difficult nodules produced a significant drop in diagnostic performance, and this decline was more pronounced for the clinicians' judgment regardless of their previous experience, than in the AI-based program which demonstrated a marked reliability.

Table 1 Diagnostic performance of the group of readers and the CAD program in the set of concordant (n = 37) and non-concordant (n = 43) nodules.

	SEN (%)	**SPE** (%)	**PPV** (%)	**NPV** (%)	**Accuracy** (%)
AmCAD					
Concordant	87.5	45.5	77.7	77.7	74.2
Non-concordant	80.7	50.0	71.4	72.4	70.0
Clinical observers					
Concordant	95.0	49.4	75.2	85.6	77.0
Non-concordant	74.0[a]	43.5	46.2[a,b]	66.1	57.1[b]

[a]P < 0.05 with respect to concordant; [b]P < 0.05 with respect to AmCAD.
CAD, computer-aided diagnostic; NPV, negative predictive value; PPV, positive predictive value; SEN, Sensitivity; SPE, Specificity.

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0023

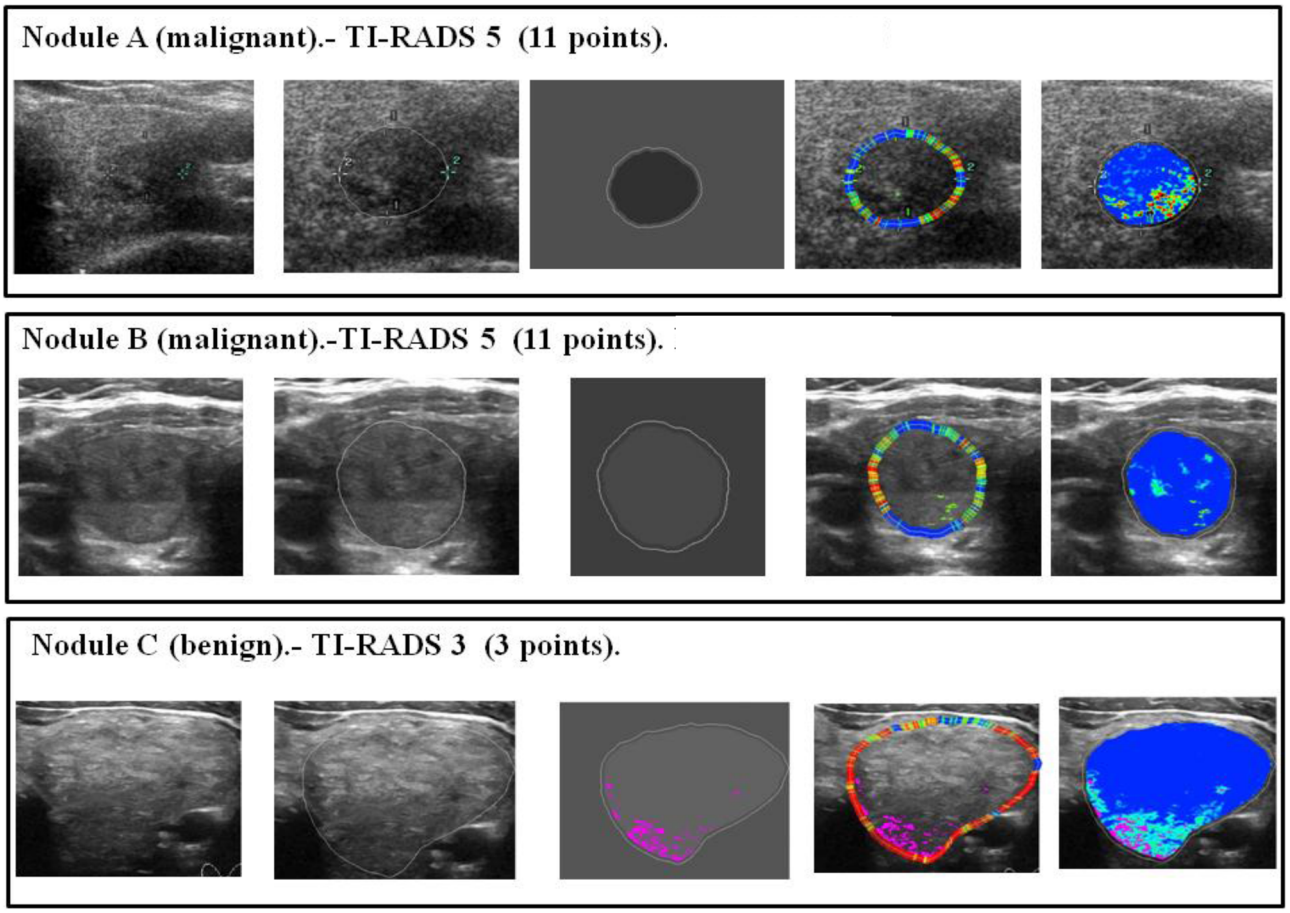

Figure 2
Risk classification offered by the CAD system adequate for the pathological results in three undefined nodules from the study. The ultrasound image, delimitation and shape, echogenicity and hypochoic foci, margins and echogenic foci, and heterogeneity are presented in each nodule.

Thyroid nodules are a very frequent referral to thyroid specialists in endocrinology and US exploration allows an accurate evaluation of the lesions and the selection of those requiring a fine-needle aspiration diagnostic procedure (12). There is a significant known interobserver variability in the US technique (13), not only due to operator expertise but also because of special characteristics sometimes present in the image evaluated.

Interobserver variability plays an important role in the interpretation, validation, and diagnosis of images obtained by different techniques such as X-rays or USs. For this reason and to overcome it, enhanced IA-based CAD programs have been developed. These systems are based in computational vision, which use mathematical algorithms to analyze images to reduce reader's subjectivity with significant enhanced diagnostic performance (7, 8), similar to those of expert observers. These CAD systems have been demonstrated to be cost-effective, easy to use, and can analyze throughput images at a fast rate. Our group recently evaluated the first commercial CAD program (AmCAD-UT®) in the assessment of the risk of malignancy in thyroid nodules and the level of diagnostic performance was almost comparable to that of the clinical endocrinologist (10). In clinical practice, a substantial number of US images are doubtful due to poorly defined characteristics of features such as echogenic foci or insufficient definition of margins, thus resulting in a lack of clear phenotypic imaging. In these cases, the ability to precisely define the risk of malignancy can be significantly reduced and the discrepancy between observers increased. In our study, we challenged the observers and the IA system to analyze nodules that were difficult to classify. We found that clinical observers decreased their diagnostic accuracy by a fifth, while in the case of the CAD program the decrease was slight and non-significant, and the final accuracy was better than with the observers. This decrease in diagnostic performance was also observed in the AUC, which in the case of human observers suffered a marked

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0023

decrease from a high precision point, while the CAD system remained stable.

Previous reports on the application of the CAD systems in thyroid nodular disease dealt with the power to detect malignant characteristics (14), how they compare with physicians regarding different degrees of expertise (15) and with other AI programs based on different algorithms (16), or their useful educational purposes (17). To our knowledge, there are no data available regarding the reliability of the CAD system exposed to misleading and complex images of thyroid nodules not allowing assigning a defined category. In our study, both the physicians and the CAD system obtained good and comparable results for non-discordant cases in SEN and NPV. In the difficult nodule analysis, the percentage reduction in all diagnostic performance parameters shown by the CAD system was only less than 5%, demonstrating significant reliability. Noteworthy, the CAD system maintained good levels of SEN and PPV, that is, its capability to detect malignant lesions. One of the uses of the CAD systems in thyroidology is to screen the lesions to serve as a filter, ruling out those highly suggestive of benignity. This would greatly reduce the number of nodules to assess FNAB and also the anxiety of patients due to the risk of suffering a malignant lesion. For this, the NPV should be as high as possible but it is also important that it be maintained even in those images that are difficult to interpret. Our study shows that CAD remains consistent with an adequate NPV value and therefore can be used as a screening tool. It should be noted that the CAD system maintains good levels of SEN and PPV, that is, its ability to detect malignant lesions.

The strengths of our article are the careful selection of the images analyzed, the collaboration of up to five observers with different experience in thyroid US, and the use of a CAD program previously validated by our group. The main weakness is the number of nodules evaluated, although it is considered sufficient to draw conclusions according to the design of the study.

In conclusion, although an expert observer is still superior to the AI system in the US evaluation of the risk of thyroid nodules malignancy, when facing images that are challenging due to their poor quality, subtle changes or misleading appearance, the program maintains a robust capacity to detect benign nodules.

Declaration of interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

Funding

This work did not receive any specific grant from any funding agency in the public, commercial or not-for-profit sector.

Author contribution statement

J L R conceived the study, performed CAD evaluation, and wrote the paper. L F-E wrote the manuscript. S B, F V, S B, A P-M O, and C P-J performed ultrasound evaluation. J L R, L F-E, and F V analyzed the data. J L R, F V, and M P-D reviewed the paper. All authors discussed the results and commented on the manuscript.

Acknowledgements

The authors thank AmCAD BioMed Corporation for providing a free version of the AmCAD-UT system for use in this study.

References

1 Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ & Cooper DS. The diagnosis and management of thyroid nodules: a review. *JAMA* 2018 **319** 914–924. (https://doi.org/10.1001/jama.2018.0898)

2 Burman KD & Wartofsky L. Thyroid nodules. *New England Journal of Medicine* 2015 **373** 2347–2356. (https://doi.org/10.1056/NEJMcp1415786)

3 Gharib H, Papini E, Garber JR, Duik DS, Harrell RM, Hegedüs L, Paschke R, Valcavi R, Vitti P & AACE/ACE/AME Task Force on Thyroid Nodules. AACE/ACE/AME Task Force on thyroid nodules American Association of Clinical Endocrinologists, American College of Endocrinology, and Associazione Medici Endocrinologi Medical guidelines for clinical practice for the diagnosis and management of thyroid nodules–2016. *Endocrine Practice* 2016 **22** 622–639. (https://doi.org/10.4158/EP161208.GL)

4 Remonti LR, Kramer CK, Leitão CB, Pinto LC & Gross JL. Thyroid ultrasound features and risk of carcinoma: a systematic review and meta-analysis of observational studies. *Thyroid* 2015 **25** 538–550. (https://doi.org/10.1089/thy.2014.0353)

5 Ha EJ, Baek JH & Na DG. Risk stratification of thyroid nodules on ultrasonography: current status and perspectives. *Thyroid* 2017 **27** 1463–1468. (https://doi.org/10.1089/thy.2016.0654)

6 Hoang JK, Middleton WD, Farjat AE, Teefey SA, Abinanti N, Boschini FJ, Bronner AJ, Dahiya N, Hertzberg BS, Newman JR, *et al.* Interobserver variability of sonographic features used in the American College of Radiology thyroid imaging reporting and data system. *American Journal of Roentgenology* 2018 **211** 162–167. (https://doi.org/10.2214/AJR.17.19192)

7 Gitto S, Grassi G, De Angelis C, Monaco CG, Sdao S, Sardanelli F, Sconfienza LM &, Mauri G. A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound. *Radiologia Medica* 2019 **124** 118–125. (https://doi.org/10.1007/s11547-018-0942-z)

8 Reverter JL, Vázquez F & Puig-Domingo M. Diagnostic performance evaluation of a computer-assisted imaging analysis system for ultrasound risk stratification of thyroid nodules. *American Journal of Roentgenology* 2019 **213** 169–174. (https://doi.org/10.2214/AJR.18.20740)

9 Xia S, Yao J, Zhou W, Dong Y, Xu S, Zhou J & Zhan WA. A computer-aided diagnosing system in the evaluation of thyroid nodules-experience in a specialized thyroid center. *World Journal of Surgical Oncology* 2019 **17** 210. (https://doi.org/10.1186/s12957-019-1752-z)

10 Yoo YJ, Ha EJ, Cho YJ, Kim HL, Han M & Kang SY. Computer-aided diagnosis of thyroid nodules via ultrasonography: initial clinical experience. *Korean Journal of Radiology* 2018 **19** 665–672. (https://doi.org/10.3348/kjr.2018.19.4.665)
11 Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, *et al.* ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *Journal of the American College of Radiology* 2017 **14** 587–595. (https://doi.org/10.1016/j.jacr.2017.01.046)
12 Popoveniuc G & Jonklaas J. Thyroid nodules. *Medical Clinics of North America* 2012 **96** 329–349. (https://doi.org/10.1016/j.mcna.2012.02.002)
13 Sych YP, Fadeev VV, Fisenko EP & Kalashnikova M. Reproducibility and interobserver agreement of different thyroid imaging and reporting data systems (TIRADS). *European Thyroid Journal* 2021 **10** 161–167. (https://doi.org/10.1159/000508959)
14 Wei X, Zhu J, Zhang H, Gao H, Yu R, Liu Z, Zheng X, Gao M & Zhang S. Visual interpretability in computer-assisted diagnosis of thyroid nodules using ultrasound images. *Medical Science Monitor* 2020 **26** e927007. (https://doi.org/10.12659/MSM.927007)
15 Chung SR, Baek JH, Lee MK, Ahn Y, Choi YJ, Sung TY, Song DE, Kim TY & Lee JH. Computer-aided diagnosis system for the evaluation of thyroid nodules on ultrasonography: prospective non-inferiority study according to the experience level of radiologists. *Korean Journal of Radiology* 2020 **21** 369–376. (https://doi.org/10.3348/kjr.2019.0581)
16 Xu L, Gao J, Wang Q, Yin J, Yu P, Bai B, Pei R, Chen D, Yang G, Wang S, *et al.* Computer-aided diagnosis systems in diagnosing malignant thyroid nodules on ultrasonography: a systematic review and meta-analysis. *European Thyroid Journal* 2020 **9** 186–193. (https://doi.org/10.1159/000504390)
17 Fresilli D, Grani G, De Pascali ML, Alagna G, Tassone E, Ramundo V, Ascoli V, Bosco D, Biffoni M, Bononi M, *et al.* Computer-aided diagnostic system for thyroid nodule sonographic evaluation outperforms the specificity of less experienced examiners. *Journal of Ultrasound* 2020 **23** 169–174. (https://doi.org/10.1007/s40477-020-00453-y)

Received in final form 25 August 2021
Accepted 7 September 2021
Accepted Manuscript published online 7 September 2021