RESEARCH

External validation of AIBx, an artificial intelligence model for risk stratification, in thyroid nodules

Kristine Z Swan[1], **Johnson Thomas**[2], **Viveque E Nielsen**[3], **Marie Louise Jespersen**[4] and **Steen J Bonnema**[5]

[1]Department of ORL, Head- and Neck Surgery, Aarhus University Hospital, Aarhus, Denmark
[2]Department of Endocrinology, Mercy Hospital, Springfield, Missouri, USA
[3]Department of ORL, Head- and Neck Surgery, Odense University Hospital, Odense, Denmark
[4]Department of Pathology, Aarhus University Hospital, Aarhus, Denmark
[5]Department of Endocrinology, Odense University Hospital, Odense, Denmark

Correspondence should be addressed to K Z Swan: kristineswan@dadlnet.dk

Abstract

Background: Artificial intelligence algorithms could be used to risk-stratify thyroid nodules and may reduce the subjectivity of ultrasonography. One such algorithm is AIBx which has shown good performance. However, external validation is crucial prior to clinical implementation.
Materials and methods: Patients harboring thyroid nodules 1–4 cm in size, undergoing thyroid surgery from 2014 to 2016 in a single institution, were included. A histological diagnosis was obtained in all cases. Medullary thyroid cancer, metastasis from other cancers, thyroid lymphomas, and purely cystic nodules were excluded. Retrospectively, transverse ultrasound images of the nodules were analyzed by AIBx, and the results were compared with histopathology and Thyroid Imaging Reporting and Data System (TIRADS), calculated by experienced physicians.
Results: Out of 329 patients, 257 nodules from 209 individuals met the eligibility criteria. Fifty-one nodules (20%) were malignant. AIBx had a negative predictive value (NPV) of 89.2%. Sensitivity, specificity, and positive predictive values (PPV) were 78.4, 44.2, and 25.8%, respectively. Considering both TIRADS 4 and TIRADS 5 nodules as malignant lesions resulted in an NPV of 93.0%, while PPV and specificity were only 22.4 and 19.4%, respectively. By combining AIBx with TIRADS, no malignant nodules were overlooked.
Conclusion: When applied to ultrasound images obtained in a different setting than used for training, AIBx had comparable NPVs to TIRADS. AIBx performed even better when combined with TIRADS, thus reducing false negative assessments. These data support the concept of AIBx for thyroid nodules, and this tool may help less experienced operators by reducing the subjectivity inherent to thyroid ultrasound interpretation.

Key Words
- thyroid nodules
- ultrasound
- artificial intelligence
- TIRADS

Introduction

Risk stratification of thyroid nodules uses ultrasound features predictive of benign or malignant disease to identify nodules that should undergo biopsy. Biopsy is an invasive procedure and may not yield a final diagnosis one out of seven times (1). Thus, reducing unnecessary biopsies may have a clinical

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0129

impact by reducing the number of diagnostic surgical procedures.

Systems used to classify thyroid nodules include, for example, Thyroid Imaging Reporting and Data System (TIRADS) created by the American College of Radiology (ACR), the American Thyroid Association classification system, the French TIRADS, k-TIRADS, and the EU-TIRADS (2, 3, 4, 5, 6). These systems are based on a subjective assessment of the nodule and have sub-optimal specificity and positive predictive values (PPV) (7). In addition, TIRADS may be inferior to the personal judgment by experts (8), and the same nodule may yield different risk estimates across different systems (9). A reliable, explainable, less subjective, and noninvasive technique to address this problem is desirable.

AIBx is an artificial intelligence (AI) model which might overcome these challenges (10). The AIBx algorithm retrieves images from an image library similar to the test image, and the associated diagnosis is displayed to the physician for decision-making. In the initial internal validation study of AIBx, the negative predictive value (NPV) of the AIBx was 93.2%, while sensitivity, specificity, PPV, and accuracy of the model were 87.8, 78.5, 65.9, and 81.5%, respectively (10). When compared to TIRADS, AIBx had comparable NPV with better sensitivity, specificity, and PPV (7, 11).

AIBx was developed using images from a single healthcare system and different ultrasound machine manufacturers. However, AI software created in one healthcare system may not provide similar results when applied to different populations and imaging machines (12). In addition, unrecognized biases and confounding factors could influence the results. Hence, it is essential to demonstrate the robustness of healthcare AI systems in different settings (13).

The primary aim of this study was to evaluate the performance of AIBx for risk stratification of thyroid nodules based on ultrasound images collected retrospectively from a different institution. The secondary aim was to assess the performance of AIBx applied to thyroid nodules with indeterminate cytology.

Materials and methods

This was a single-center retrospective study wherein existing ultrasound images of thyroid nodules and corresponding histological diagnosis were used to validate AIBx. The original images were collected for a prospective study previously described (clinicaltrial.gov registration no: NCT02150772) (14). In brief, adult patients undergoing thyroid surgery were included between January 2014 and February 2016 at the Department of Otorhinolaryngology, Head & Neck Surgery, Aarhus University Hospital, Denmark. Preoperatively, the majority of patients underwent fine-needle aspiration of the nodule. Bethesda categories (I–VI) were used to describe the cytopathological diagnosis (15). A histologically verified diagnosis, according to the WHO classification (16), was obtained in all patients based on the surgical specimen.

Image acquisition

For the present study, nodules were excluded if the dimension was <10 mm or >40 mm, or if the whole nodule was not completely visible in one ultrasound image. Ultrasound images containing annotations, markings, writings, or crosshairs within the nodule were excluded. Multinodular goiters without a separable nodule on the ultrasound image, medullary thyroid cancer, metastasis from other cancers, thyroid lymphomas, and purely cystic nodules were also excluded (Fig. 1).

All ultrasound images were acquired by two experienced physicians using SuperSonic Aixplorer (Supersonic, Aix en Provence, France). Images were obtained in the transverse plane and stored as JPEG files. Based on ultrasound features suggestive of malignancy, a TIRADS score was prospectively assigned, based on the EU-TIRADS criteria (4).

AIBx

One anonymized B-mode image from each nodule was transferred from Aarhus University Hospital, Denmark, to Mercy Hospital, USA. The images were analyzed using AIBx, while blinding the investigators toward the cytological and histological diagnoses, as well as the TIRADS score. In total, 2025 images were available in the reference library obtained on ultrasound machines manufactured by GE, Siemens, Philips, and Sonosite (10). Diagnosis of the first similar image by AIBx was considered as the diagnostic output of the algorithm.

Outcomes

The AIBx results were returned to Aarhus University Hospital for comparison with the true diagnoses (i.e. histopathological results) and the TIRADS scores. Accuracy, sensitivity, specificity, PPV, NPV, and area under the curve

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0129

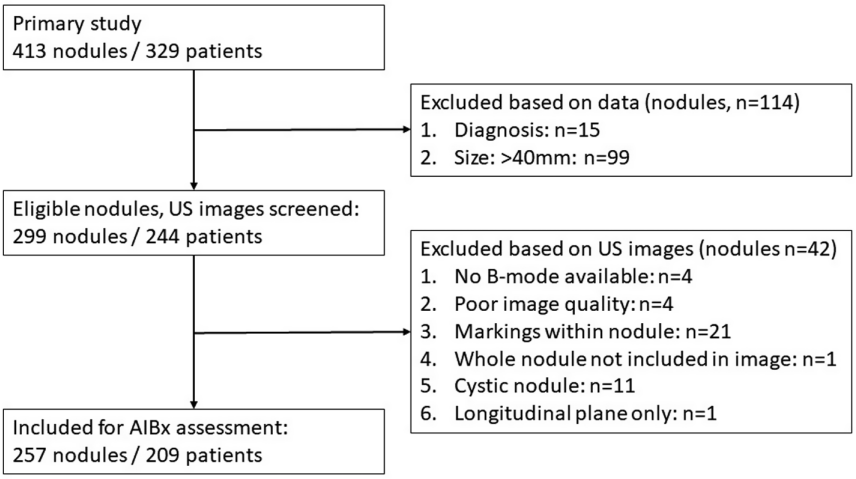

Figure 1
Patient selection flowchart. US, ultrasonography.

(AUC) were calculated from a confusion matrix using python programming language. A subgroup analysis was done on cytologically indeterminate nodules.

The study complies with the World Medical Association Declaration of Helsinki. The prospective collection of ultrasound images was approved by the Ethics committee of Region Midt, Denmark, for which the participants gave their written informed consent. Approval was given by the Ethics committee of Region Midt, Denmark to pass on the anonymous images from the prospective study, without obtaining further consent from the participants. The prospective study was registered at ClinicalTrials.gov (NCT02150772). A waiver of Consent/Assent and waiver of Health Insurance Portability and Accountability Act Authorization were granted for this study from Mercy Hospital. Institutional review board approval was obtained from Mercy Hospital.

Results

The original dataset contained 413 nodules from 329 patients. In total, 257 nodules from 209 patients were eligible for the study (females/males: 161/48; age (mean ± S.D.) 55.7 ± 13.2 years). All but 51 nodules underwent a biopsy. There were 206 benign (80%) and 51 malignant nodules (20%). The latter were all differentiated thyroid carcinomas, including 10 follicular thyroid carcinoma (FTC), 9 follicular variant papillary thyroid carcinoma (FvPTC), 29 classical PTC, and 3 tall cell variant PTC. When data were grouped based on the Bethesda classification, 56% of the nodules that underwent biopsy were assigned an indeterminate diagnosis, that is category III, IV, or V, with the majority (32%) being category IV (suspicious for follicular neoplasm) (Fig. 2).

Diagnostic performance

Diagnostic assessments are shown in Table 1 alongside the results of AIBx and TIRADS assessment. Overall, AIBx performed with an NPV of 89.2%. Sensitivity, specificity, and PPV were 78.4, 44.2, and 25.8%, respectively, while AUC was 0.61. For TIRADS, 190 nodules (74%) were in TIRADS 5 category, and 24 nodules (9%) were in TIRADS 4, resulting in a higher false-positive rate and a lower accuracy than obtained by AIBx. When restricting the analyses to PTC, AIBx had an NPV of 96% and an AUC of 0.65, while the corresponding values for TIRADS were 93% and 0.55, respectively. If only TIRADS 5 nodules were considered malignant, the NPV was 89.6%.

Concordance rates

The concordance rate between AIBx and TIRADS was 58%. Eleven malignant nodules (five FTC, two FvPTC, three

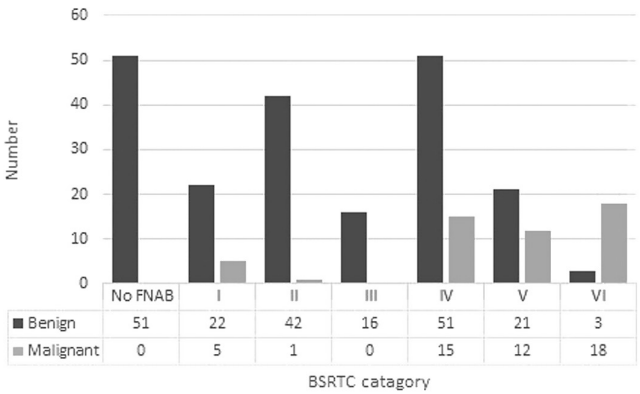

Figure 2
BSRTC category according to histological diagnosis. BSRTC, Bethesda system for reporting thyroid cytopathology; FNAB, fine-needle aspiration biopsy.

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0129

Table 1 Diagnostic and ultrasonographic assessment of included nodules.

Test	Result						
	Malignant, *n* (%)	Benign, *n* (%)	Sens (%)	Spec (%)	NPV (%)	PPV (%)	Accuracy (%)
All included nodules (*n* = 257)							
Histology[a]	51 (20)	206 (80)	–	–	–	–	–
AIBx	155 (60)	105 (40)	78.4	44.2	89.2	25.8	51.0
TIRADS	214 (83)	(43 (17)	94.1	19.4	93.0	22.4	34.2
Nodules with an indeterminate cytological diagnosis (BSRTC: III, IV, V) (*n* = 115)							
Histology[a]	27 (23)	88 (77)	–	–	–	–	–
AIBx	61 (53)	54 (47)	63.0	50.0	81.5	27.9	53.0
TIRADS	103 (90)	12 (10)	96.3	12.5	91.7	25.2	32.2

[a]Reference diagnosis.
BSRTC, Bethesda system for reporting thyroid cytopathology; NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; Spec, specificity.

classical PTC, and one tall cell PTC) were falsely negative by AIBx but correctly diagnosed by TIRADS. The biopsies were categorized as Bethesda IV (*n* = 7), V (*n* = 3), or VI (*n* = 1). Three malignant nodules, all PTC and classified as Bethesda category I, II, or V, were falsely negative by TIRADS but correctly diagnosed by AIBx. No malignant nodules were falsely negative by both AIBx and TIRADS. If both methods predicted the nodule to be malignant, the prevalence of malignancy was 28%.

Nodules with indeterminate cytological diagnoses

There were 115 nodules assigned Bethesda category III, IV, or V, representing indeterminate cytological diagnoses. The diagnoses for this subgroup assessed by histology, AIBx, and TIRADS score are shown in Table 1. The 27 malignant nodules included ten FTC, eight follicular variant PTC, and nine classical PTC. All nodules in the Bethesda category III (*n* = 16) were falsely predicted to be malignant by TIRADS (1 nodule assigned TIRADS 4 and 15 nodules assigned TIRADS 5). In contrast, only 25% of nodules in the Bethesda category III were predicted to be malignant by AIBx. Considering only TIRADS 5 nodules as malignant, the accuracy increased to 37% with a decrease in sensitivity and NPV to 82 and 81%, respectively.

Discussion

AIBx showed comparable performance to existing risk stratification methods as reflected by an NPV of 89%, increasing to 96% when restricting the analysis to PTC cancers only. On comparing AIBx to the assigned TIRADS score, AIBx performed with higher specificity and PPV but lower sensitivity and NPV. However, when including only PTC in the malignant group, the NPV of AIBx was higher than achieved by TIRADS. In our study, the thyroid ultrasound and the assessment of the TIRADS score were made by experts. It is likely, but remains to be proven, that AIBx will perform even better than conventional risk stratifications made by non-experts (17).

Overall, AIBx categorized fewer benign nodules as malignant (false positive) than TIRADS, thus potentially reducing the number of biopsies needed. On the other hand, AIBx categorized a higher fraction of malignant nodules as benign (false negative) compared with TIRADS, potentially overlooking more cancers. This pinpoints the major challenge in the risk stratification of thyroid nodules, of which only few are malignant (3). Therefore, clinicians must balance the risk of overlooking malignancy while reducing the number of unnecessary biopsies in order to avoid overtreatment of benign nodules. If a biopsy is categorized as indeterminate, this usually leads to diagnostic surgery or molecular testing due to a malignancy risk in the range of 6–60% (15).

The false-negative rates of AIBx and TIRADS were 22 and 6%, respectively. Importantly, no malignant nodule was overlooked when both methods deemed it benign. Thus, combining AIBx with TIRADS may have clinical relevance in order to avoid unnecessary biopsies. The majority of nodules presenting with false-negative AIBx results were of follicular origin, but these were all correctly assessed by TIRADS. On the contrary, all three PTC classified as benign by TIRADS were assessed malignant by AIBx. Two of these were also cytopathologically overlooked (Bethesda I or II). The shortcomings of TIRADS were confirmed in a large study, in which ACR TIRADS misclassified 32% of malignant nodules (18).

In the indeterminate categories (Bethesda III, IV, and V), NPV of TIRADS was similar to that found in the whole cohort, while NPV of AIBx decreased to 81.5%. This is probably explained by the relatively higher fraction

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0129

of FTC and FvPTC in this subgroup, while PTC was the predominant malignancy in the reference library (10). Thus, AIBx performed better when assessing PTC only. In Bethesda III nodules, all being histopathologically benign, AIBx performed superior to TIRADS, categorizing only 25% as malignant, while TIRADS deemed all nodules malignant.

In the internal validity study (10), the accuracy of AIBx was 81.5%, as compared with 51.0% in this study. This is mostly explained by a higher rate of false-positive results for AIBx in this external cohort, as reflected by higher specificity and PPV in the internal validity study. The NPV and sensitivity were also slightly higher in the internal validity study (10), but not to the same extent. The fraction of FTC in the present cohort was higher than in the reference library, and this type of thyroid carcinoma is generally more difficult to identify by ultrasound, as compared to PTC (3, 4, 19).

Facial recognition technology trained on a subset of population fails to recognize faces of another subset (20). Similarly, medical algorithms based on images from one machine may not work well on images obtained from another machine (12). The images from the SuperSonic Aixplorer machine used in the present study had different textures and were generally larger than the images used in the initial study (10). In addition, iodine status, environmental factors, and access to healthcare could affect the size and morphology of the thyroid nodule, and at which stage it comes to medical attention. Even with these differences, AIBx demonstrated a good NPV. It might be possible that the performance could be further improved by adding images from the SuperSonic Aixplorer or other machines to the AIBx reference library.

A recent review suggested that AI algorithms are able to match the accuracy provided by radiologists and pathologists (21). AIBx has several advantages. Unlike other automated black box algorithms, the operators are involved in each step of the algorithm, from image selection to assessing similar image categories. The operator can thereby counterbalance the potential problems posed by AI models. The system uses a technology similar to face recognition (10), which has some analogy with the pattern recognition upon which the various TIRADS are designed. Similar images from a large database are presented to the operator, thus supporting decision-making. For the less skilled operator, the library compensates for the lack of experience by presenting similar images and may prove to be a valuable teaching tool.

There are a few limitations of our study. This was a retrospective study, and only one image of each nodule was available. The image selected was the most representative of the nodule. However, another image from the same nodule might present different ultrasound features, and thus be scored differently by AIBx. Preferably, an image obtained in the longitudinal plane, and not only in the transversal plane, should be included for risk assessment. In addition, the performance of AIBx might have been further improved by including dynamic cine loops for the selection of representative images. In a Chinese retrospective study (22), a deep convolutional neural network model, that included sets of thyroid images, showed improved specificity in identifying thyroid cancer patients, as compared with a judgment by skilled radiologists. As for AIBx, future studies are needed to clarify the ideal setup.

Our patients represented a selected surgical cohort with a relatively high fraction of cancer. The NPV would probably increase in a cohort of patients harboring more benign nodules (23). However, to confirm the benign nature surgical removal of all nodules would be needed, which is rarely indicated in unselected patients. Microcarcinomas were not included in our study and should probably be investigated separately from larger cancers.

Ultrasonography is a method increasingly used for the morphological evaluation of thyroid nodules (3). This results in the identification of many nodules that otherwise would remain undetected. Most of these lesions are benign but will elicit a diagnostic work up to rule out malignancy. The huge number of thyroid images makes it possible to train and sophisticate the AI algorithms. Some of the initial algorithms were based on features extracted by physicians (21). Subsequently, thyroid ultrasound images were used directly to train deep learning models. Further studies are needed to identify the optimal use of AIBx and to increase its performance. This involves adding images from other institutions and ultrasound machines to the reference database. Optimal workflow for incorporating AIBx into clinical use is another issue that needs to be addressed.

We conclude that AIBx had comparable NPV to TIRADS, when applied to ultrasound images obtained in a different setting than used for training. AIBx performed even better than TIRADS in Bethesda category III and PTC. Combining AIBx with TIRADS may be highly valuable in clinical practice, by reducing unnecessary biopsies while still identifying thyroid cancer with high accuracy. Our study proves the concept of AIBx for thyroid nodules, and this tool may help less experienced operators by

https://etj.bioscientifica.com
https://doi.org/10.1530/ETJ-21-0129

reducing the subjectivity inherent to thyroid ultrasound interpretation.

Declaration of interest

J T Owns intellectual property rights to AIBx algorithm. The other authors have nothing to disclose.

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

All authors participated in the planning of the study and writing the manuscript. Image collection was performed by V E N and K Z S, pathological analysis was performed by M L J, and AIBx analysis was performed by J T.

References

1 Ali SZ, Siperstein A, Sadow PM, Golding AC, Kennedy GC, Kloos RT & Ladenson PW. Extending expressed RNA genomics from surgical decision making for cytologically indeterminate thyroid nodules to targeting therapies for metastatic thyroid cancer. *Cancer Cytopathology* 2019 **127** 362–369. (https://doi.org/10.1002/cncy.22132)

2 Grant EG, Tessler FN, Hoang JK, Langer JE, Beland MD, Berland LL, Cronan JJ, Desser TS, Frates MC, Hamper UM, *et al.* Thyroid ultrasound reporting lexicon: white paper of the ACR thyroid imaging, reporting and data system (TIRADS) committee. *Journal of the American College of Radiology* 2015 **12** 1272–1279. (https://doi.org/10.1016/j.jacr.2015.07.011)

3 Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, *et al.* 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer. *Thyroid* 2016 **26** 1–133. (https://doi.org/10.1089/thy.2015.0020)

4 Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R & Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *European Thyroid Journal* 2017 **6** 225–237. (https://doi.org/10.1159/000478927)

5 Kwak JY, Han KH, Yoon JH, Moon HJ, Son EJ, Park SH, Jung HK, Choi JS, Kim BM & Kim EK. Thyroid imaging reporting and data system for US features of nodules: a step in establishing better stratification of cancer risk. *Radiology* 2011 **260** 892–899. (https://doi.org/10.1148/radiol.11110206)

6 Russ G, Royer B, Bigorgne C, Rouxel A, Bienvenu-Perrard M & Leenhardt L. Prospective evaluation of thyroid imaging reporting and data system on 4550 nodules with and without elastography. *European Journal of Endocrinology* 2013 **168** 649–655. (https://doi.org/10.1530/EJE-12-0936)

7 Grani G, Lamartina L, Ascoli V, Bosco D, Biffoni M, Giacomelli L, Maranghi M, Falcone R, Ramundo V, Cantisani V, *et al.* Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the 'right' TIRADS. *Journal of Clinical Endocrinology and Metabolism* 2019 **104** 95–102. (https://doi.org/10.1210/jc.2018-01674)

8 Solymosi T, Hegedus L, Bonnema SJ, Frasoldati A, Jambor L, Kovacs GL, Papini E, Rucz K, Russ G, Karanyi Z, *et al.* Ultrasound-based indications for thyroid fine-needle aspiration: outcome of a TIRADS-based approach versus operators' expertise. *European Thyroid Journal* 2021 **10** 416–424. (https://doi.org/10.1159/000511183)

9 Huang BL, Ebner SA, Makkar JS, Bentley-Hibbert S, McConnell RJ, Lee JA, Hecht EM & Kuo JH. A multidisciplinary head-to-head comparison of American College of Radiology thyroid imaging and reporting data system and American Thyroid Association ultrasound risk stratification systems. *Oncologist* 2020 **25** 398–403. (https://doi.org/10.1634/theoncologist.2019-0362)

10 Thomas J & Haertling T. AIBx, artificial intelligence model to risk stratify thyroid nodules. *Thyroid* 2020 **30** 878–884. (https://doi.org/10.1089/thy.2019.0752)

11 Ahmadi S, Oyekunle T, Jiang X', Scheri R, Perkins J, Stang M, Roman S & Sosa JA. A direct comparison of the Ata and TI-RADS ultrasound scoring systems. *Endocrine Practice* 2019 **25** 413–422. (https://doi.org/10.4158/EP-2018-0369)

12 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ & Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine* 2018 **15** e1002683. (https://doi.org/10.1371/journal.pmed.1002683)

13 Bini F, Pica A, Azzimonti L, Giusti A, Ruinelli L, Marinozzi F & Trimboli P. Artificial intelligence in thyroid field-a comprehensive review. *Cancers* 2021 **13** 4740. (https://doi.org/10.3390/cancers13194740)

14 Swan KZ, Bonnema SJ, Jespersen ML & Nielsen VE. Reappraisal of shear wave elastography as a diagnostic tool for identifying thyroid carcinoma. *Endocrine Connections* 2019 **8** 1195–1205. (https://doi.org/10.1530/EC-19-0324)

15 Cibas ES & Ali SZ. The 2017 Bethesda system for reporting thyroid cytopathology. *Thyroid* 2017 **27** 1341–1346. (https://doi.org/10.1089/thy.2017.0500)

16 Bychkov A. Thyroid & parathyroid: Thyroid - general; WHO classification. Bingham Farms, MI, USA: PathologyOutlines, 2017. (available at: https://www.pathologyoutlines.com/topic/thyroidwho.html)

17 Kim SH, Park CS, Jung SL, Kang BJ, Kim JY, Choi JJ, Kim YI, Oh JK, Oh JS, Kim H, *et al.* Observer variability and the performance between faculties and residents: US criteria for benign and malignant thyroid nodules. *Korean Journal of Radiology* 2010 **11** 149–155. (https://doi.org/10.3348/kjr.2010.11.2.149)

18 Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM & Desser TS. Comparison of performance characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TIRADS, and American Thyroid Association guidelines. *American Journal of Roentgenology* 2018 **210** 1148–1154. (https://doi.org/10.2214/AJR.17.18822)

19 Lee JH, Han K, Kim EK, Moon HJ, Yoon JH, Park VY & Kwak JY. Risk stratification of thyroid nodules with atypia of undetermined significance/follicular lesion of undetermined significance (AUS/FLUS) cytology using ultrasonography patterns defined by the 2015 ATA guidelines. *Annals of Otology, Rhinology, and Laryngology* 2017 **126** 625–633. (https://doi.org/10.1177/0003489417719472)

20 Buolamwini J & Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pp. 77–91. Eds AF Sorelle & W Christo. Proceedings of Machine Learning Research: PMLR, 2018.

21 Thomas J, Ledger GA & Mamillapalli CK. Use of artificial intelligence and machine learning for estimating malignancy risk of thyroid nodules. *Current Opinion in Endocrinology, Diabetes, and Obesity* 2020 **27** 345–350. (https://doi.org/10.1097/MED.0000000000000557)

22 Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, Xin X, Qin C, Wang X, Li J, *et al.* Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet: Oncology* 2019 **20** 193–201. (https://doi.org/10.1016/S1470-2045(18)30762-9)

23 Ferris RL, Baloch Z, Bernet V, Chen A, Fahey TJ, 3rd, Ganly I, Hodak SP, Kebebew E, Patel KN, Shaha A, *et al.* American Thyroid Association statement on surgical application of molecular profiling for thyroid nodules: current impact on perioperative decision making. *Thyroid* 2015 **25** 760–768. (https://doi.org/10.1089/thy.2014.0502)

Received in final form 24 January 2022
Accepted 3 February 2022
Accepted Manuscript published online 3 February 2022