RESEARCH

Prediction of thyroid nodule histopathology by expert ultrasound evaluation

Olav Inge Håskjold[1], Henrik Stenestø Foshaug[2], Therese Benedikte Iversen[3], Helga Charlotte Kjøren[3] and **Vegard Heimly Brun[1,2]**

[1]Department of Breast and Endocrine Surgery, University Hospital of North Norway, Tromsø, Norway
[2]UiT – The Arctic University of Norway, Institute of Clinical Medicine, Tromsø, Norway
[3]Department of Radiology, University Hospital of North Norway, Harstad, Norway

Correspondence should be addressed to V H Brun: vegard.h.brun@uit.no

Abstract

Objective: The basis of thyroid nodule diagnostics is ultrasound-guided fine needle biopsy with cytological evaluation (FNC) if ultrasound appearance is not clearly benign. The aim of this study was to investigate the predictive potential of dedicated, expert high-resolution ultrasound, to see if histopathological entities of thyroid nodules can be diagnosed without invasive FNC biopsies.
Design: Prospective case-cohort study.
Methods: 187 patients with 221 thyroid nodules were examined with ultrasound and prospectively assigned to the expected histopathological diagnosis: colloid nodule, adenomatoid colloid nodule, follicular adenoma, follicular carcinoma, follicular variant of papillary thyroid carcinoma, papillary thyroid carcinoma, or other thyroid cancer. In 101 of these, we later obtained histopathological reports for comparison.
Results: Overall accuracy for classification into discrete histopathological categories by expert ultrasound was 71.3% and Cohen's Kappa was 0.62. The sensitivity and specificity for detecting malignancy were 97.3% and 78.1%. The diagnostic accuracy for malignancy was 85.1%. ACR-TIRADS scores for the same nodules had a sensitivity of 97.3%, specificity of 26.6%, and accuracy of 52.5%.
Conclusion: Dedicated expert high-resolution ultrasound without FNC can reliably distinguish benign vs malignant nodules, but also differentiate between several histopathological entities in thyroid nodules. There is potential for a reduction in the number of invasive FNC biopsies and diagnostic operations.

Key Words
- thyroid nodule
- cytology
- ultrasound
- overtreatment
- overdiagnosis
- accuracy

Endocrine Connections (2021) **10**, 776–781

Introduction

Surgery of the thyroid gland can be indicated because of compression symptoms, thyrotoxicosis, or cancer. However, typically 35–40% of surgeries in Europe are made when examination including fine-needle cytology (FNC) biopsy fail to exclude malignancy, according to the European database EUROCRINE (www.eurocrine.eu). In most of these cases, the final histology is benign, indicating overdiagnosis and overtreatment of thyroid nodules (1, 2).

In order to reduce overtreatment of thyroid nodules, improvements in the diagnostic workup are continuously being made. Ultrasound (US) is the primary radiological modality for the evaluation of thyroid nodules and has undergone significant technical development. While US was considered unsuitable to distinguish between benign and malignant thyroid nodules 1 or 2 decades ago (3, 4), now it has a central place in several risk stratification systems (RSS) for detection of thyroid malignancy (5). Among the RSS, there are several versions of Thyroid Imaging Reporting and Data Systems (TIRADS) published in the last 5 years (6, 7, 8). The RSS are helpful to standardize and

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-21-0192

improve the radiological examination, but consensus on a common system is hindered by low inter-rater reliability and variation in the malignancy risk associated with each category (9, 10, 11, 12).

FNC by ultrasound guidance is the required standard for thyroid nodule cytology, in contrast to FNC guided by palpation. Although US is accepted as necessary for the selection of suspicious nodules, US characteristics of thyroid nodules are often considered unspecific. Consequently, the decision about surgery is predominantly based on cytological evaluation, and the US evaluation in itself is not emphasized in the diagnostic workup. This is true despite the fact that cytology has a diagnostic accuracy of only 70% (13). Some studies show, however, that in cases of uncertain cytology, US evaluation can help to decide when diagnostic surgery is warranted (14, 15). In addition, emerging results from artificial intelligence (AI) algorithms show that the US images themselves contain enough information to predict malignancy with diagnostic accuracy comparable to cytological evaluation (16, 17).

We wanted to explore the diagnostic potential of recognition-based thyroid US, without cytology, when performed by a dedicated radiologist. The aim was to evaluate if the US evaluation alone can predict subtypes of benign and malignant histopathological entities.

Materials and methods

Patient inclusion and prospective ultrasound evaluation

187 patients referred to our clinic were consecutively enrolled in this study between February 2018 and April 2019. The reasons for referral were either symptomatic thyroid nodules or incidentally discovered thyroid nodules. The same expert thyroid ultrasound operator with more than a decade of thyroid-specific experience from University clinics evaluated 221 thyroid nodules. None of the patients had previous biopsies or cytology reports, but some had US examinations at referring clinics. Thirty-six of the nodules were also assessed and scored by a less experienced ultrasound operator. This subset of nodules was pseudorandomly selected by the days both operators were present in the clinic. The two operators were never present in the examining room simultaneously, and the scorings were performed prior to any communication between the two to ensure independent assessments. All exams were performed using a Philips Epiq 5G ultrasound machine (Phillips Ultrasound Inc, Bothell, WA, USA), with a L12-5 linear array transducer. We used a pre-determined custom scoring template with categories corresponding to the most common histological diagnoses: colloid nodule, adenomatoid colloid nodule, follicular adenoma, follicular carcinoma, follicular variant of papillary thyroid carcinoma, papillary thyroid carcinoma, or other thyroid cancer. The operator also indicated the confidence of the assignment to any category on a scale from 1 to 5 (1 = very uncertain, 2 = uncertain, 3 = neutral, 4 = certain, 5 = very certain).

The position of each evaluated nodule was indicated on a drawing of the thyroid gland, and its size was recorded, to ensure that later comparison with histology would be of the very same nodule. In cases of uncertainty, multiple categories were sometimes marked, but the category with the highest score was used for all analyses. After scoring, further diagnostic workup and treatment followed a clinical routine, including cytological biopsies if indicated. Evaluation for surgery was carried out according to national and international guidelines by surgeons and the multidisciplinary team who had no knowledge about the patient's participation in this study.

Histological evaluation

When patient inclusion was complete, we searched the hospital records of the patients and found that 98 of the assessed nodules were removed by surgery and 3 had undergone large needle biopsy (all anaplastic thyroid carcinomas). This allowed a comparison between initial ultrasound assessment and final histology for 101 nodules in total. Histological reports were manually categorized to harmonize with the pre-determined histological classes. For example, benign Hürthle cell adenomas were assigned to the follicular adenoma category. For most reports, no harmonization was necessary. In cases of multiple entities in the histological specimen, for example, the incidental finding of a papillary microcarcinoma (mPTC) embedded in colloid nodules, care was made to match the nodule evaluated by ultrasound to the histological description of the same area. Thus, histological mPTCs < 5 mm were disregarded in the analysis if found incidentally after surgery for larger nodules of another entity.

ACR-TIRADS classification

Ultrasound images were stored offline as video files containing transversal cine loops with approximately 250–500 frames per cine loop from each thyroid lobe. The scan was made from the submandibular gland to

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-21-0192

the supraclavicular fossa in a standardized fashion. To characterize our data set, two consultant general radiologists with previous experience of ACR-TIRADS and thyroid examinations blindly scored all 221 video recordings in the study retrospectively. No additional information about the patient was made available, except that 101 of the 221 nodules had required surgery and that some of them were cancers. The radiologists were instructed to score the most suspicious nodule in the video file according to ACR-TIRADS (6). Only the lobule of interest was scored. The average score of the two radiologists was used to classify the nodules according to ACR-TIRADS. Interobserver reliability was calculated as linearly weighted Cohen's kappa coefficients.

Statistical analysis

All data were organized in Microsoft Excel (Microsoft Corporation) and processed in MATLAB R2020b (MathWorks). IBM SPSS Statistics (IBM) was used for statistical tests.

Ethics approval/consent to participate

The study was approved by the Data Protection Official of the University Hospital of North Norway (approval 02050). The Norwegian Regional Ethics Committee North waived the need for written patient consent (ref 225025).

Results

Of all evaluated nodules (Table 1), about 4 out of 5 were in women. The age of patients with benign and malignant disease was not statistically different (ANOVA $F(2,218)=2.29$, $P=0.10$). Benign operated nodules were larger than unoperated presumably benign nodules or operated malignant nodules (ANOVA $F(2,215)=18.02$, $P<0.001$, *post hoc* contrasts $t(215)=5.90$, $P=<0.001$) while malignant and unoperated nodules were of equal size ($t(215)=0.83$, $P=0.41$).

We first evaluated the ability of expert US to distinguish between benign and malignant thyroid nodules, without any TIRADS scoring. Nodules were assigned to predicted histopathological entities, and the confidence level was scored as described in the Materials and Methods section. Of total 101 nodules, 78 US classifications were scored as either confident or very confident (Fig. 1A). Twenty-three of 101 nodules had lower confidence scores (value 1–3) indicating uncertainty, and these nodules were classified as 'Uncertain'. To calculate the ability of ultrasound to detect malignancy, 'Certain malignant' and 'Uncertain' ultrasound evaluation was both considered a positive test, as both categories demand further diagnostic or therapeutic action. With this definition, ultrasound detected cancer with a sensitivity of 97.3% (36/37) and the specificity was 78.1% (50/64). The positive predictive value was 72.0% while the negative predictive value was 98.0% (50/51), yielding an accuracy score of 85.1%. The only overlooked cancer in our material (patient ID 94) was a 4 mm papillary microcarcinoma (mPTC) scored as a colloid nodule with a confidence score of 4. It was fine-needle biopsied nevertheless and a Bethesda category V scoring led to diagnostic hemithyroidectomy and the histological diagnosis of mPTC.

Table 1 Overview of all included thyroid nodules in the study.

	Malignant histology	Benign histology	No histology
Number of nodules	37	64	120
Males (%)	22%	17%	19%
Age (years)	56 ± 3	50 ± 2	58 ± 2
Nodule size (mm)	21 ± 3	38 ± 2	23 ± 2

As several risk stratification systems for thyroid nodules have been published and implemented recently, we wanted to compare the performance of expert pattern recognition with one of the most commonly used RSS, the ACR-TIRADS. Twenty-eight of the 101 nodules were scored in TIRADS category 2–3, while the remaining 73 nodules were scored TIRADS 4–5 (Fig. 1B). The inter-rater reliability was fair (weighted kappa 0.38). The sensitivity for malignancy was 97.3%, comparable to expert ultrasound, but specificity was only 26.6%. The positive predictive value was 43.4%, and the negative predictive value was 94.4%. Thus, TIRADS categorization successfully ruled in cancer patients but had a low ability to rule out cancer.

Recognition-based ultrasound assessment classified the later unoperated nodules as mostly benign, clearly different from later operated nodules (chi-square 34.1, $P<0.0001$, Fig. 1A vs 1C), while TIRADS categorization for the unoperated nodules did not differ from the operated nodules (chi-square 1.88, $P=0.60$, Fig. 1B vs 1D).

To evaluate the ability of dedicated US to predict histological outcomes in more detail, we compared predicted and actual histological outcomes using three benign and four malignant categories (Fig. 2). Overall accuracy was 71%, and a Cohen's Kappa was 0.62 indicating substantial agreement. The precision was good for colloid nodules (71%), follicular adenomas (76%),

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-21-0192

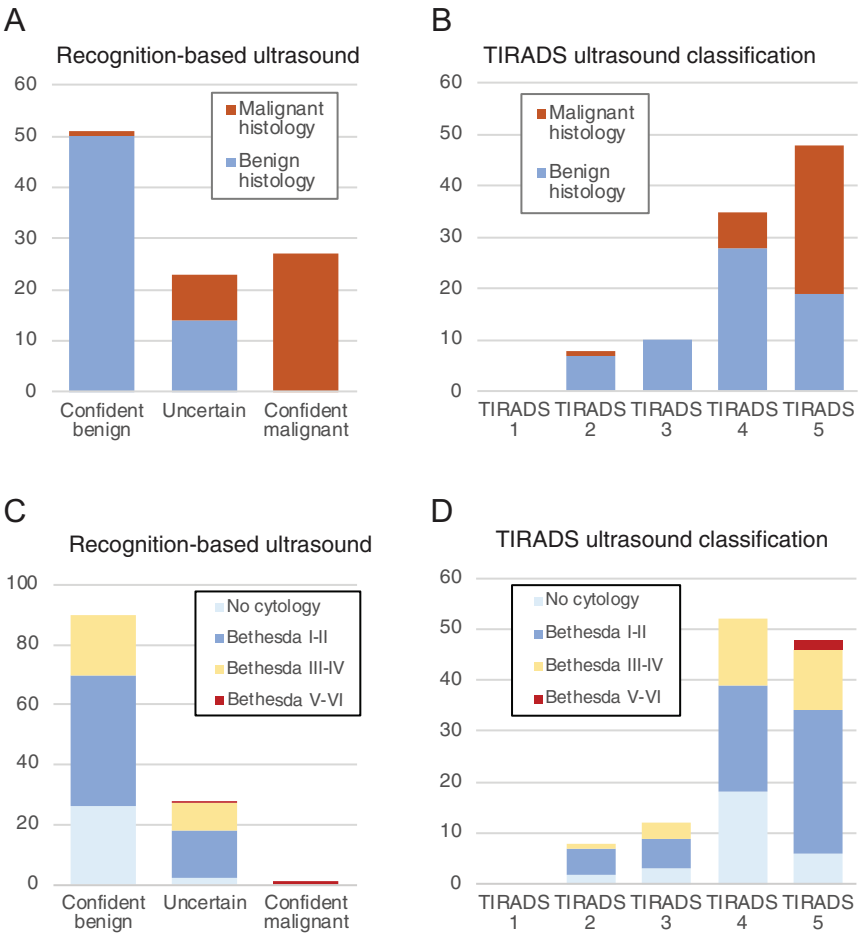

Figure 1
Classification of later operated (A and B) and unoperated (C and D) thyroid nodules according to recognition-based ultrasound (A and C) or ACR-TIRADS (B and D). For operated nodules, the final histopathology is shown in colors, while for unoperated nodules, the cytological assessment is color-coded.

papillary thyroid carcinomas (88%), while the other categories had too few observations to allow statistical assessment. We, therefore, re-categorized our data material into fewer categories, corresponding to clinically significant entities. Colloid and adenomatoid nodules were merged, all follicular neoplasias including follicular cancers were merged. Papillary thyroid cancers and anaplastic thyroid cancers remained as separate categories. The analysis showed that expert ultrasound evaluation could reliably differentiate between these four categories (Fig. 3). Overall accuracy was 82%, and the Cohen's Kappa was 0.73 indicating substantial agreement. The weighted average precision was 82%, while the weighted average sensitivity

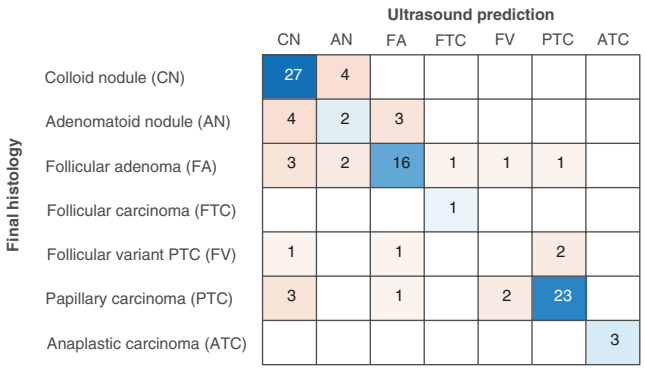

Figure 2
Confusion matrix comparing ultrasound prediction with actual histopathological diagnosis. Prediction was based on non-invasive ultrasound examination (without cytology). Overall accuracy was 71%.

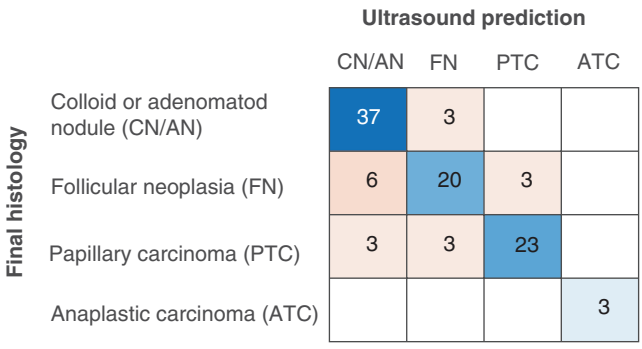

Figure 3
Confusion matrix comparing ultrasound prediction with actual histopathological diagnosis. Data are the same as in Fig. 2 but grouped into the major clinically relevant entities. Overall accuracy was 82%.

https://ec.bioscientifica.com
https://doi.org/10.1530/EC-21-0192

was 83%. The positive predictive value (precision) for each sub-category was 80% for colloid/adenomatoid nodules, 77% for follicular neoplasias, and 88% for papillary thyroid carcinomas. All three anaplastic cancers were correctly identified. We saw no medullary thyroid cancers in our study period.

To assess whether this single operator, recognition-based method can be replicated, a thyroid ultrasound trainee independently scored a subset of 36 nodules immediately before or after the expert examined the patient. The correlation of raw scoring matrixes was relatively high 0.72, and there was substantial agreement (weighted kappa value 0.70) on the main diagnosis when using the four diagnostic categories of Fig. 3.

Discussion

Our data indicate that thyroid US alone, without cytology, can have a high diagnostic value when performed by a dedicated operator. The results show the potential of expert thyroid US to help avoid invasive FNC and reduce the number of diagnostic thyroidectomies. The subjectivity of the method is inherent in all ultrasound-based diagnostic approaches including risk stratification systems for thyroid nodules.

The diagnostic performance of ultrasound in this study is similar to or better than the published sensitivity and specificity of cytological evaluation of thyroid nodules. However, a comparison between ultrasound and cytology is not reasonable, as the performance of cytology is directly dependent on the ultrasound evaluation. Cytological sampling is guided by ultrasound and, therefore, strongly biased by the quality of the ultrasound examination and by the precision of the ultrasound-guided FNC. When interpreting the results of this study, we also acknowledge that the accuracy of US is user- and equipment-dependent. The importance of this should be considered in the clinical settings when expert high volume US is not available, and the thyroid US and FNC are performed by low volume endocrine or ENT surgeons, sometimes not able to properly document images and FNC needle positions.

A perceived weakness in our study is the reliance on a single ultrasound operator. However, we do not aim to describe a systematic methodology for recognition-based thyroid nodule diagnosis but to demonstrate the potential of the ultrasound modality, without FNC. Expert recognition depends heavily on human pattern recognition and pattern completion processes. Such top-down processing normally outperforms feed-forward algorithms that summarize individual features (18). Recognition-based radiological diagnosis can be achieved by high volume experts but also using artificial intelligence (AI). While many researchers believe that the future of thyroid nodule diagnostics is found in molecular markers, several papers show that AI performs relatively well in detecting thyroid malignancy. Until now, human selection of the nodule of interest is still required in the AI algorithms (16). We are probably just seeing the beginning of AI in medical imaging and expect this field to expand considerably.

TIRADS scoring of our data set assigned the vast majority of nodules to category 4 and 5, yielding low specificity but a high negative predictive value (NPV). The tendency to score most nodules in the higher categories transmits most of the diagnostic differentiation to cytological evaluation, leaving US without significant contribution in the clinical decision-making. A downstaging of the TIRADS scores would probably not give a better selection but could result in low NPV consistent with other reports (19). We believe that the role of TIRADS is outside thyroid competence centers, where it can be very useful for the selection of patients that need referral or FNC. However, in dedicated thyroid centers or future AI-assisted diagnostics, US evaluation should be more ambitious and aspire to predict the pathological entity of the thyroid nodules. Expert ultrasound evaluation of thyroid nodules reaches beyond TIRADS and ATA algorithms, providing an excellent distinction between malignant and benign nodules, and high accuracy in predicting final histological diagnosis.

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
Vegard H Brun and Olav Inge Håskjold conceived the idea and designed the study. Olav Inge Håskjold performed prospective expert ultrasound evaluations, Vegard H Brun performed trainee ultrasound evaluations, while Therese B Iversen and Helga C Kjøren performed TIRADS scoring. Henrik Foshaug and Vegard H Brun analyzed the data. Vegard H Brun wrote the manuscript and supervised the project. All authors discussed the results and contributed to the final manuscript.

References

1 Mayson SE & Haugen BR. Molecular diagnostic evaluation of thyroid nodules. *Endocrinology and Metabolism Clinics of North America* 2019 **48** 85–97. (https://doi.org/10.1016/j.ecl.2018.10.004)

2 Schenke S, Klett R, Seifert P, Kreissl MC, Görges R & Zimny M. Diagnostic performance of different thyroid imaging reporting and data systems (Kwak-TIRADS, EU-TIRADS and ACR TI-RADS) for risk stratification of small thyroid nodules (≤10 mm). *Journal of Clinical Medicine* 2020 **9** 236. (https://doi.org/10.3390/jcm9010236)

3 Hegedüs L. Thyroid ultrasound. *Endocrinology and Metabolism Clinics of North America* 2001 **30** 339–360, viii–ix. (https://doi.org/10.1016/S0889-8529(05)70190-0)

4 Brito JP, Gionfriddo MR, Al Nofal A, Boehmer KR, Leppin AL, Reading C, Callstrom M, Elraiyah TA, Prokop LJ, Stan MN, *et al.* The accuracy of thyroid nodule ultrasound to predict thyroid cancer: systematic review and meta-analysis. *Journal of Clinical Endocrinology and Metabolism* 2014 **99** 1253–1263. (https://doi.org/10.1210/jc.2013-2928)

5 Trimboli P & Durante C. Ultrasound risk stratification systems for thyroid nodule: between lights and shadows, we are moving towards a new era. *Endocrine* 2020 **69** 1–4. (https://doi.org/10.1007/s12020-020-02196-6)

6 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)

7 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)

8 Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, Lim HK, Moon WJ, Na DG, Park JS, *et al.* Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology Consensus Statement and Recommendations. *Korean Journal of Radiology* 2016 **17** 370–395. (https://doi.org/10.3348/kjr.2016.17.3.370)

9 Pandya A, Caoili EM, Jawad-Makki F, Wasnik AP, Shankar PR, Bude R, Haymart MR & Davenport MS. Retrospective cohort study of 1947 thyroid nodules: a comparison of the 2017 American College of Radiology TI-RADS and the 2015 American Thyroid Association classifications. *American Journal of Roentgenology* 2020 **214** 900–906. (https://doi.org/10.2214/AJR.19.21904)

10 Phuttharak W, Boonrod A, Klungboonkrong V & Witsawapaisan T. Interrater reliability of various Thyroid Imaging Reporting and Data System (TIRADS) classifications for differentiating benign from malignant thyroid nodules. *Asian Pacific Journal of Cancer Prevention* 2019 **20** 1283–1288. (https://doi.org/10.31557/APJCP.2019.20.4.1283)

11 Seifert P, Görges R, Zimny M, Kreissl MC & Schenke S. Interobserver agreement and efficacy of consensus reading in Kwak-, EU-, and ACR-thyroid imaging recording and data systems and ATA guidelines for the ultrasound risk stratification of thyroid nodules. *Endocrine* 2020 **67** 143–154. (https://doi.org/10.1007/s12020-019-02134-1)

12 Hoang JK, Middleton WD & Tessler FN. Update on ACR TI-RADS: successes, challenges, and future directions, from the AJR special series on radiology reporting and data systems. *American Journal of Roentgenology* 2021 **216** 570–578. (https://doi.org/10.2214/AJR.20.24608)

13 Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L & Baloch ZW. The Bethesda system for reporting thyroid cytopathology: a meta-analysis. *Acta Cytologica* 2012 **56** 333–339. (https://doi.org/10.1159/000339959)

14 Gao LY, Wang Y, Jiang YX, Yang X, Liu RY, Xi XH, Zhu SL, Zhao RN, Lai XJ, Zhang XY, *et al.* Ultrasound is helpful to differentiate Bethesda class III thyroid nodules: a PRISMA-compliant systematic review and meta-analysis. *Medicine* 2017 **96** e6564. (https://doi.org/10.1097/MD.0000000000006564)

15 Moran C, Reyna R, Boots LS & Azziz R. Adrenocortical hyperresponsiveness to corticotropin in polycystic ovary syndrome patients with adrenal androgen excess. *Fertility and Sterility* 2004 **81** 126–131. (https://doi.org/10.1016/j.fertnstert.2003.07.008)

16 Koh J, Lee E, Han K, Kim EK, Son EJ, Sohn YM, Seo M, Kwon MR, Yoon JH, Lee JH, *et al.* Diagnosis of thyroid nodules on ultrasonography by a deep convolutional neural network. *Scientific Report* 2020 **10** 15245.

17 Park VY, Han K, Seong YK, Park MH, Kim EK, Moon HJ, Yoon JH & Kwak JY. Diagnosis of thyroid nodules: performance of a deep learning convolutional neural network model vs. radiologists. *Scientific Report* 2019 **9** 17843.

18 Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Ortega Caro J, Hardesty W, Cox D & Kreiman G. Recurrent computations for visual pattern completion. *PNAS* 2018 **115** 8835–8840. (https://doi.org/10.1073/pnas.1719397115)

19 Miao S, Jing M, Sheng R, Cui D, Lu S, Zhang X, Jing S, Zhang X, Shan T, Shan H, *et al.* The analysis of differential diagnosis of benign and malignant thyroid nodules based on ultrasound reports. *Gland Surgery* 2020 **9** 653–660. (https://doi.org/10.21037/gs.2020.04.03)

Received in final form 14 June 2021
Accepted 22 June 2021
Accepted Manuscript published online 22 June 2021