European Thyroid Journal

Clinical Thyroidology / Research Article

Eur Thyroid J 2021;10:416–424
DOI: 10.1159/000511183

Received: April 25, 2020
Accepted: August 25, 2020
Published online: November 25, 2020

Ultrasound-Based Indications for Thyroid Fine-Needle Aspiration: Outcome of a TIRADS-Based Approach versus Operators' Expertise

Tamas Solymosi[a] Laszlo Hegedüs[b] Steen Joop Bonnema[b] Andrea Frasoldati[c] Laszlo Jambor[d] Gabor Laszlo Kovacs[e] Enrico Papini[f] Karoly Rucz[g] Gilles Russ[h] Zsolt Karanyi[i] Endre V. Nagy[i]

[a]Endocrinology and Metabolism Clinic, Bugat Hospital, Gyöngyös, Hungary; [b]Department of Endocrinology and Metabolism, Odense University Hospital, Odense, Denmark; [c]Endocrinology Unit of Arcispedale S. Maria Nuova, Reggio Emilia, Italy; [d]Department of Radiology, Faculty of Medicine, University of Debrecen, Debrecen, Hungary; [e]1st Department of Medicine, Flohr Ferenc Hospital, Kerepestarcsa, Hungary; [f]Regina Apostolorum Hospital in Albano, Rome, Italy; [g]1st Department of Medicine, University of Pecs, Pecs, Hungary; [h]Unité Thyroïde et Tumeurs Endocrines – Pr Leenhardt Hôpital La Pitie Salpetriere, Sorbonne Université, Paris, France; [i]Division of Endocrinology, Department of Medicine, Faculty of Medicine, University of Debrecen, Debrecen, Hungary

Keywords
Thyroid nodule · Cytology · Thyroid carcinoma · Thyroid nodule image reporting and data systems · Ultrasound

Abstract
Background: Thyroid nodule image reporting and data systems (TIRADS) provide the indications for fine-needle aspiration (FNA) based on a combination of nodule sonographic features and size. We compared the TIRADS-based recommendations for FNA with those based on the personal expertise of qualified US investigators in the diagnosis of thyroid malignancy. ***Methods:*** Seven highly experienced ultrasound (US) investigators from 4 countries evaluated, online, the US video recordings of 123 histologically verified thyroid nodules. Technical resources provided the operators with a diagnostic approach close to the real-world practice. Altogether, 4,305 TIRADS scores were computed. The combined diagnostic potential of TIRADS (TIRSYS) and the personal recommendations of the investigators (PERS) were compared against 3 possible goals: to recognize all malignant lesions (allCA), nonpapillary plus non-pT1 papillary cancers (nPnT1PCA), or stage II-IV cancers (st2-4CA). ***Results:*** For allCA and nPnT1PCA, TIRSYS had lower sensitivity than PERS (69.8 vs. 87.2 and 83.5 vs. 92.6%, respectively, $p < 0.01$), while in st2-4CA the sensitivities were the same (99.1 vs. 98.6% and TIRSYS vs. PERS, respectively). TIRSYS had a higher specificity than PERS in all 3 types of cancers ($p < 0.001$). PERS recommended FNA in a similar proportion of lesions smaller or larger than 1 cm (76.9 vs. 82.7%; ns). ***Conclusions:*** Recommendations for FNA based on the investigators' US expertise demonstrated a better sensitivity for thyroid cancer in the 2 best prognostic groups, while TIRADS methodology showed superior specificity over the full prognostic range of cancers. Thus, personal experience provided more accurate diagnoses of malignancy, missing a lower number of small thyroid cancers, but the TIRADS approach resulted in a similar accuracy for the diagnosis of potentially aggressive lesions while

karger@karger.com
www.karger.com/etj

Tamas Solymosi
Endocrinology and Metabolism Clinic, Bugat Hospital
Dozsa Gy. ut 20-22
HU–3200 Gyöngyös (Hungary)
solymosi@thyrosite.com

sparing a relevant number of FNAs. Until it is not clearly stated what the goal of the US evaluation is, that is to diagnose all or only clinically relevant thyroid cancers, it cannot be determined whether one diagnostic approach is superior to the other for recommending FNA.

Introduction

For more than three decades, the cornerstones in the clinical management of patients with thyroid nodules have been ultrasound (US) and fine-needle aspiration (FNA) cytology [1–5]. Robust evidence demonstrates that the risk of malignancy in thyroid lesions is significantly correlated to the presence of specific US features, which include hypoechogenicity, microcalcifications, taller-than-wide shape, irregular or lobulated margins, and extra-thyroidal growth [1, 2, 6–11].

Several US thyroid nodule risk-classification systems have been proposed by scientific societies [10–16]. These thyroid nodule image reporting and data systems (TI-RADS) aim at providing indications for FNA, based on the combined results of the TIRADS malignancy risk scores and nodule size. Although all TIRADS are concordant in recommending against biopsy of nodules smaller than 1 cm [10–16], all of them allow considering diagnostic FNA if a nodule's US characteristics are highly suspicious. The latter approach is supported by long-term observational data demonstrating low-level aggressiveness of subcentimeter papillary carcinomas [17].

While the predictive value for malignancy of the single TIRADS systems has been reported as satisfactory in retrospective studies [10–16], recent prospective trials demonstrated only moderate diagnostic accuracy and considerable interobserver variation [18, 19]. Therefore, the actual diagnostic advantages of the TIRADS approach versus the operators' expertise need clarification in settings similar to the real-world clinical practice.

The present study did not aim at assessing yet another comparison of the validity of the different TIRADS systems for the diagnosis of thyroid malignancy. Focus of the current study was to compare the actual clinical advantages of the TIRADS-based approach for FNA indication with the recommendations from highly experienced US investigators and based on their personal expertise. The trial methodology was carefully designed to mimic the real-world conditions of thyroid US investigations. The outcome from the different approaches was evaluated in three predefined settings of malignancy. These, while seemingly arbitrarily chosen, represented different levels of aggressiveness of thyroid cancer: (a) the diagnosis of all malignant lesions whatever their size and histology (all-CA); (b) the diagnosis of the potentially more aggressive tumors, that is, nonpapillary cancers plus non-pT1 papillary cancers (nPnT1PCA); or (c) the diagnosis of the more advanced tumors only, that is, stage II-IV carcinomas (st2-4CA). In addition, the diagnostic performance of the investigators and TIRADS was compared in relation to the largest diameter of the nodules. The number of FNAs indicated by the two diagnostic approaches was also evaluated.

Materials and Methods

Patients and US Video Records

Between January 2014 and December 2016, the US examinations of 16,407 consecutive patients were video recorded and archived at the Thyroid Clinic of the Bugat Pal Hospital (Gyöngyös, Hungary) as part of the institutional routine record keeping. A Philips CX 50 US machine, equipped with a 12–5 MHz linear transducer, was used for thyroid US. Statistical power calculations have shown that a minimum of 102 cases, including at least 39 malignant cases, were required. We actually added 20% to the calculated numbers, suggesting that we have ample power in this study. In total, 709 cases had surgery for nodular goiter. From this chronological list of patients, the first patient (starting point) was randomly chosen, and the US video records of 47 (38.2%) subsequently operated patients with benign final histology, as well as 76 (61.8%) subsequently operated patients with malignant final histology, formed the sample of thyroid nodules included in the study. The indications for surgery were based on cytology in 79 patients (Bethesda IV in 19 patients, Bethesda V in 32 patients, and Bethesda VI in 28 patients), symptoms and/or signs of compression caused by the goiter in 35 cases, an autonomously functioning nodule causing hyperthyroidism in 5 patients, and patient wish in 4 cases. Final diagnoses were, in all cases, obtained by histological examination of the surgical samples. Relevant patient data appear in Table 1.

The most representative parts of each US video recording were presented to 7 investigators (see below), who were blinded to the data. In order to reproduce a setting similar to the real world, a short summary of the pre-US clinical data, including thyroid hormone and antibody levels, was provided. For the same reason, four US characteristics were pre-entered into the case report forms (CRFs): the three diameters of the nodule, the presence of taller-than-wide shape or pathological cervical lymph nodes, and the degree of nodular vascularization. A nodule presented taller-than-wide shape if the ratio of the anteroposterior to transverse diameter of the nodule was >1. A lymph node was defined as pathological if it had cystic change, heterogeneity, peripheral or abnormal vascularization, or calcifications. The intra-nodular vascularization was graded as absent, present and not extensive, or present and extensive. No clinically evident distant metastases were revealed in any patient.

Table 1. Nodule histology and tumor stage. All samples were verified by histology

Histology	*N*	Tumor status[1]	Tumor stage[1]	Male/female	Mean age (range), years
Benign	76	na	na	18/58	51.1 (24–75)
No nodule	4	na	na	0/4	49.3 (38–60)
Hyperplastic nodule	43	na	na	10/33	54.8 (30–75)
Adenoma	29	na	na	8/21	45.9 (24–67)
Malignant	47	T1 *n* = 25 T2 *n* = 8 T3 *n* = 2 T4 *n* = 11, na = 1		13/34	44.6 (18–89)
Papillary carcinoma	37	T1 *n* = 22 T2 *n* = 7 T3 *n* = 1 T4 *n* = 7	Stage I *n* = 33 Stage II *n* = 2 Stage III *n* = 1 Stage IV *n* = 1	11/26	42.5 (18–67)
Follicular carcinoma	3	T1 *n* = 2 T3 *n* = 1	Stage I *n* = 3	0/3	51 (21–53)
Poorly differentiated cancer	1	T1 *n* = 1	Stage I *n* = 1	0/1	42
Anaplastic carcinoma	2	T4 *n* = 2	Stage IV *n* = 2	2/0	69.5 (57–82)
Medullary carcinoma	2	T2 *n* = 1 T4 *n* = 1	Stage II *n* = 1 Stage IV *n* = 1	0/2	65.5 (32–89)
B-cell lymphoma	1	na	Stage II *n* = 1	0/1	49
Parathyroid carcinoma	1	T4 *n* = 1	Stage IV *n* = 1	0/1	54
Total	123				

na, not applicable. [1] TNM, Classification of Malignant Tumors (ref. [20]).

Evaluation Phase

The expert evaluations were performed online, using a website developed for this purpose. Seven investigators, from different thyroid centers in four European countries, with at least 15 years of experience in thyroid US (SB, AF, LJ, GK, EP, KR, and GR) analyzed the US video recordings of the 123 histologically verified thyroid lesions. The investigators were aware that all lesions had been operated upon but were blinded to the final histopathology and the benign to malignant ratio of the series of nodules under examination.

During a training phase, immediately preceding the study, ten nodules (not included in the 123 cases series) were analyzed by the 7 investigators in order to obtain acquaintance with the study methodology and resolve any questions before launch of the present study. The steering committee (L.H., T.S., and E.V.N.) resolved any issue raised by the US investigators. No further communication among investigators or with the steering committee was allowed.

The 123 cases were presented separately, and in random order, to each investigator. The transducer orientation above the upper, middle, or lower as well as the medial or lateral lobe regions was indicated. A static image of the whole gland was also included and the position of the nodule to be studied shown. The videos (median duration 43 s, range 20–73 s) allowed slow-motion assessment, repeat evaluation, and image-freezing, without time constraints. After the analysis of each video, investigators answered the 16 questions present in the electronic CRF. Fifteen of these were TIRADS-related questions, while the last question assessed the personal opinion of the investigator in relation to whether he would or would not perform an FNA assessment as part of his diagnostic nodule workup. The 15 questions enabled the generation of the 5 TIRADS systems' final score (see online suppl. Table 1; see www.karger.com/doi/10.1159/000511183, for all online suppl. material). To simulate a real-life-like evaluation, investigators could not modify their answers or rereview the video recordings once completed. Four-weeks were allowed for completion of the 123 cases. TIRADS scores were generated subsequently and were not accessible to the investigators during the evaluation phase.

Generating TIRADS Scores

The CRF enabled the electronic generation of 5 major TIRADS scores: the American Thyroid Association (ATA) [12], the American College of Radiology (ACR) [14], the American Association of Clinical Endocrinologists/American College of Endocrinology/Associazione Medici Endocrinologi (AACE/ACE/AME) [13], the European Thyroid Association (ETA) [15], and the Korean Society of Thyroid Radiology (KSTR) [16] scores. The generation of the 5 TIRADS scores, using the nodule characteristics entered into the CRFs by the 7 investigators, resulted in 35 TIRADS scores per nodule (5 different TIRADS scores by each of the 7 investigators). These scores were automatically translated to recommendation "for" or "against" the use of FNA according to the criteria of the different TIRADS (Table 2). The diagnostic performance of the TIRADS classifications was evaluated against final histology. Thus, a score of 35 "for FNA" and 0 "against FNA" (100%) of the total 35 TIRADS scores, in case of a histological malignant nodule, would indicate that TIRADS performs perfectly in nodule selection for FNA. In contrast, a score of 30 for FNA and 5 against FNA (85.7%)

Table 2. Indications of FNA cytology according to the different TIRADS systems

TIRADS category	AACE/ACE/AME, mm[a]	ATA, mm	ACR, mm	ETA, mm	KSTR, mm
1	Not indicated	Not indicated	Not indicated	Not indicated	Not indicated
2	>20	Not indicated[b]	Not indicated	Not indicated	Not indicated
3	>10[c]	≥15	≥25	>20	≥20
4	Not applicable	≥10	≥15	>15	≥15
5	Not applicable	>10[c]	≥10[c]	>10[c]	≥10[c]

TIRADS, thyroid imaging reporting and data system; AACE, American Association of Clinical Endocrinologists; ACE, American College of Endocrinology; AME, Associazione Medici Endocrinologi [13]; ATA, American Thyroid Association [12]; ACR, American College of Radiology [14]; ETA, European Thyroid Association [15]; KSTR, Korean Society of Thyroid Radiology [16]; FNA, fine-needle aspiration. [a] AACE/ACE/AME uses 3 categories. [b] ATA allows the consideration of FNA if ≥2 cm. [c] Between 5 and 10 mm, all TIRADS suggest the consideration of either FNA or active surveillance depending on the clinical setting and patient preference.

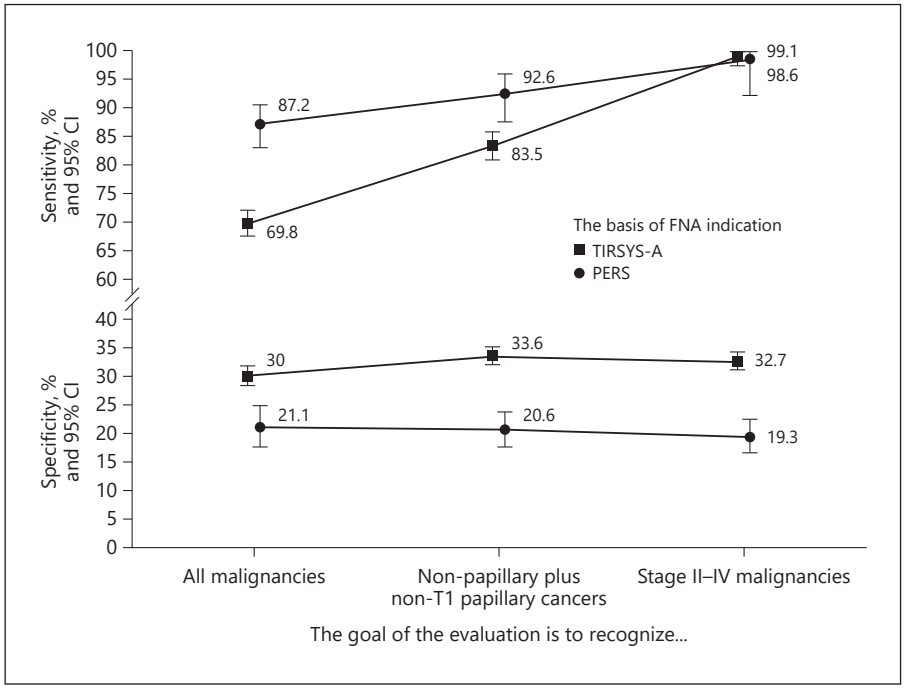

Fig. 1. Comparison of the TIRADS systems and the investigators' opinion. Sensitivities and specificities are dependent on which cancer phenotypes are to be found. TIRSYS-A, st2-5 TIRADS combined; PERS, investigators' judgment; st2-4CA, stage II-IV malignancies.

would indicate a somewhat poorer performance. This places each technique somewhere in the 0–100% range. As the scope of the trial was focused on the comparison of the TIRADS-guided methodology versus the traditional expertise-guided approach, only the combined performance obtained from the 5 TIRADS, defined as TIRSYS score, and not the individual performance of each of the 5 TIRADS systems, was evaluated in this study.

With the aim of analyzing the consequences of the possibility for the experts to indicate FNA for subcentimeter thyroid lesions, 2 options were used. The first approach (defined as TIRSYS-A) calculated the TIRADS' diagnostic outcomes without considering the indication for FNA in nodules <10 mm, even if in the high-risk US categories, nor for lesions >20 mm in the very low-suspicion groups of AACE/ACE/AME [13]. The second approach (defined as TIRSYS-B) included the indication for FNA in subcentimeter nodules in the high-risk TIRADS categories, as well as in nodules >20 mm in the lowest US suspicion group.

Indication for FNA

The outcome of TIRSYS was evaluated in 3 malignancy settings. These represented, even if arbitrarily chosen, different levels of aggressiveness of thyroid cancer: (i) the diagnosis of allCA whatever their size and histology; (ii) the diagnosis of the potentially more aggressive tumors (i.e., all nonpapillary malignancies and all non-T1 papillary cancers found in the thyroid), nPnT1PCA; or (iii) the diagnosis of the more advanced tumors only, that is, st2-4CA. The expertise-guided PERS of the investigators, with respect to FNA, were also analyzed. In addition, we compared the diagnos-

tic performance of PERS and TIRSYS methodologies in relation to the largest diameter of the nodules.

Statistical Methods

Sensitivities, specificities, and the 95% confidence intervals of the approaches tested (TIRSYS and PERS) were calculated against the 3 potential goals of the examination (diagnosing allCA, nPnT-1PCA, or st2-4CA) by package "epiR" [21] in R project [22]. The 2 approaches were compared by the χ^2 test.

Results

All investigators completed the full set of examinations and reported all the requested data in the online CRF. The 15 characteristics for each of the 123 nodules resulted in 12,915 answers and the calculation of 4,305 TIRADS scores. Data reported below in parenthesis refer to the frequency of the given nodule characteristics, as selected by the investigators from the CRF choices (not the number of nodules).

Based on TIRSYS-A and PERS approaches, FNA was indicated in 69.9 and 82.1% ($p < 0.001$), respectively, of the whole series of nodules whatever their size (range: 5–85 mm). According to the same approach, FNA was indicated in 78.2 and 82.7% ($p = 0.02$), respectively, of the nodules >10 mm. The comparison of TIRSYS-A and PERS, for the indication of FNA, is presented in Figure 1. The sensitivity of PERS exceeded that of TIRSYS-A by 17.4% in allCA and by 9.1% in nPnT1PCA histological categories ($p > 0.01$). PERS had a lower specificity than TIRSYS-A in both allCA (21.1 vs. 30.0%, $p < 0.01$) and nPnT1PCA histological categories (20.6 vs. 33.6%, $p < 0.01$). The better specificity of TIRSYS-A was evident for all 3 cancer categories, including st2-4CA (Fig. 1; Table 3). In the s2-4CA category, the sensitivity of both TIRSYS-A (99.1%) and PERS (98.6%) was excellent in patient selection for FNA, while the specificity of TIRSYS-A (32.7%) was higher than that of PERS (19.3%) ($p < 0.01$).

The results of TIRSYS-B were similar to PERS, with no significant difference in sensitivity or specificity in any of the malignancy phenotype categories. Significant differences were observed between TIRSYS-A and TIRSYS-B, mimicking the differences described between TIRSYS-A and PERS (Table 3).

TIRSYS-A does not recommend FNA in nodules <10 mm. Both PERS and TIRSYS-B were allowed to recommend FNA in such nodules. There was no statistically significant difference in FNA recommendation in nodules smaller or >10 mm. Employing PERS the rate was 76.9% (70/91) in nodules <10 mm and 82.7% (637/770) (ns) in nodules ≥10 mm ($p = 0.17$). Using TIRSYS-B, FNA was recommended in 77.1% (351/455) of nodules smaller than 10 mm and in 82.3% (3,169/3,850) of those ≥10 mm ($p = 0.07$).

Irrespective of the nodule size, the sensitivity of PERS and TIRSYS-B tended to be higher and the specificity lower when compared with TIRSYS-A (see Table 4). Finally, after exclusion of the papillary microcarcinomas from the analysis, the sensitivity of PERS was not significantly better than that of TIRSYS-A (91.5 vs. 88.7%; $p = 0.18$). PERS indicated FNA in 82.1% (707/861) of cases, TIRSYS-A in 69.9% (3,010/4,305) of cases, and TIRSYS-B in 82.0% (3,530/4,305) of cases (TIRSYS-A vs. PERS and TIRSYS-A vs. TIRSYS-B, $p < 0.01$ for both comparisons).

Discussion

Several studies, mostly from single centers, have evaluated the role of individual TIRADS systems in predicting thyroid malignancy [23–31]. Even if the interobserver agreement between different centers appears lower than that assessed in single-center trials [19], all TIRADS systems demonstrated satisfactory outcomes. However, the available trials analyzed the performance of the various TIRADS in settings different from the real-world practice. The majority of the studies used static US images [23–28, 32], were performed in single centers, and relied on FNA and not on final histology as the "gold standard." For these reasons, we used a video-based approach, which is superior to static images, is much closer to the real-world situation, and bypasses the potential bias due to inappropriate image preselection [33]. Histology was used as gold standard, since FNA as a reference would have distorted the calculations, because of the risk of false-positive and false-negative FNA results and of the uncertain diagnoses due to indeterminate cytology [25]. The few studies which compared TIRADS based on histologically verified nodules were based on static images [29, 30, 34, 35]. The use of FNA instead of final histology and the bias caused by the retrospective selection of still images were considered major limitations of the available studies in a recent meta-analysis by Castellano et al. [36]. In aggregate, these trials have analyzed the comparison between the different TIRADS categories. However, they did not address the factual advantages of TIRADS methodology versus the traditional diagnostic performance based on the expertise of US operators, for recommending FNA. Table 5 summarizes the main differences between the present and previous studies.

Table 3. Sensitivities and specificities of the TIRADS systems (TIRSYS) and investigators' opinion (PERS) according to the cancer phenotypes to be found

	Cancer phenotype					
	allCA		nPnT1PCA		st2-4CA only	
	sensitivity, % (CI)[3]	specificity, % (CI)[3]	sensitivity, % (CI)[3]	specificity, % (CI)[3]	sensitivity, % (CI)[3]	specificity, % (CI)[3]
TIRSYS-A[1]	69.8 (67.6–72.1)	30.0 (28.3–31.8)	83.5 (80.9–85.9)	33.6 (32–35.2)	99.1 (97.5–99.8)	32.7 (31.2–34.2)
TIRSYS-B[1]	86.7 (84.9–88.3)	21.3 (19.7–22.9)	90.5 (88.4–92.4)	20.2 (18.8–21.6)	99.7 (98.4–100)	19.8 (18.6–21.1)
PERS[2]	87.2 (83.1–90.6)	21.1 (17.7–24.8)	92.6 (87.6–96)	20.6 (17.6–23.8)	98.6 (92.3–100)	19.3 (16.6–22.3)
p (TIRSYS-A vs. PERS) (χ^2)	<0.001 (41.8)	<0.001 (17.5)	0.002 (9.4)	<0.001 (44.9)	0.65	<0.001 (55.3)
p (TIRSYS-B vs. PERS) (χ^2)	0.81	0.89	0.39	1.0	0.20	0.76
p (TIRSYS-A vs. TIRSYS-B) (χ^2)	<0.001 (139.1)	<0.001 (52.6)	<0.001 (18.8)	<0.001 (149.7)	0.31	<0.001 (169.2)

In TIRSYS-A, the stricter suggestion of TIRADS was taken into account; FNA was indicated neither in nodules <1 cm nor in the very low-suspicion group of ATA. In TIRSYS-B, the permissive suggestions of TIRADS have been considered; FNA was indicated both in those subcentimeter nodules which belonged to the highest TIRADS category and in the very low-suspicion group of ATA if the nodule was >2 cm. allCA, all malignant lesions; nPnT1PCA, nonpapillary plus non-pT1 papillary cancers; st2-4CA, stage II-IV cancers; FNA, fine-needle aspiration; TIRADS, thyroid image reporting and data systems; ATA, American Thyroid Association. [1] TIRSYS means the combined results of TIRADS of 5 professional societies, that is, the AACE/ACE/AME [13], the ATA [12], the ACR [14], the ETA [15], and the KSTR [16]. [2] Personal suggestions of the investigators. [3] CI means 95% confidence interval.

Eur Thyroid J 2021;10:416–424
DOI: 10.1159/000511183

Table 4. Sensitivities and specificities of the TIRADS systems (TIRSYS) and investigators' opinion (PERS) according to the size of the nodules

	Maximal diameter of the lesion							
	<1 cm		1–2 cm		2–3 cm		≥3 cm	
	sensitivity, % (CI)[3]	specificity, % (CI)	sensitivity, % (CI)	specificity, % (CI)	sensitivity, % (CI)	specificity, % (CI)	sensitivity, % (CI)	specificity, % (CI)
TIRSYS-A[1]	0 (0–1)	100.0 (96.5–100)	84.8 (81.9–87.5)	44.5 (41.1–47.8)	96.4 (93.5–98.3)	15.5 (12.7–18.6)	90.0 (86.4–92.9)	19.6 (17.3–22.1)
TIRSYS-B[1]	74.3 (69.4–78.8)	13.3 (7.5–21.4)	84.8 (81.9–87.5)	44.5 (41.1–47.8)	97.9 (95.4–99.2)	8.9 (6.7–11.5)	94.0 (91–96.2)	10.2 (8.5–12.2)
PERS[2]	71.4 (59.4–81.6)	5.5 (0.1–23.8)	88.0 (81.2–93.0)	31.4 (24.6–38.9)	92.9 (82.7–98.0)	21.8 (14.8–30.4)	97.1 (90.1–99.7)	13.8 (9.5–19.1)
p (TIRSYS-A vs. PERS)	<0.001	<0.001	0.35	0.01	0.22	0.09	0.05	0.05
p (TIRSYS-B vs. PERS)	0.62	0.27	0.35	0.01	0.04	<0.001	0.29	0.12
p (TIRSYS-A vs. TIRSYS-B)	0.001	<0.001	1.0	1.0	0.31	<0.001	0.05	<0.001

In TIRSYS-A, the stricter suggestion of TIRADS was taken into account; FNA was indicated neither in nodules <1 cm nor in the very low-suspicion group of ATA. In TIRSYS-B, the permissive suggestions of TIRADS have been considered; FNA was indicated both in those subcentimeter nodules which belonged to the highest TIRADS category and in the very low-suspicion group of ATA if the nodule was >2 cm. [1] TIRSYS means the combined results of TIRADS of 5 professional societies, that is, the AACE/ACE/AME [13], the ATA [12], the ACR [14], the ETA [15], and the KSTR [16]. [2] Personal suggestions of the investigators. [3] CI means 95% confidence interval.

Table 5. Aims and tools used by historical studies compared to the present study

Reference	Tool of the analysis	Was the biological reference histology?	Goal of the study regarding TIRADS	Size of the nodule taken into account as suggested by the respective TIRADS
Middleton et al. [23], Lauria Pantano et al. [24], Grani et al. [25], Wu et al. [26], Ha et al. [27], Xu et al. [28]	Image	No	To determine the distribution of benign and malignant cases among TIRADS categories	No
Gao et al. [29], Skowronska et al. [30], Chng et al. [34], Trimboli et al. [35]	Image	Yes		No
Current study	Video	Yes	To analyze the usefulness of TIRADS in recommendation of FNA	Yes

TIRADS, thyroid imaging reporting and data system; FNA, fine-needle aspiration.

For the above reasons, the outcomes of the composite TIRADS score, here defined as TIRSYS, and those of the indications based on the personal experience of the examiners, here defined as PERS, were blindly assessed, simulating conditions close to real clinical practice. TIRADS systems recommend against the use of FNA for thyroid lesions <10 mm, whatever their US risk level; however, all of them allow consideration of diagnostic FNA in highly suspicious cases [16–20]. TIRSYS-A did not consider for FNA thyroid nodules <10 mm, while TIRSYS-B evaluated the indication for FNA of suspicious lesions <10 mm. The outcomes were analyzed considering 3 different diagnostic goals. We arbitrarily distinguished 3 clinical conditions that may represent different levels of harm for the patients: (i) the detection of allCA, (ii) the diagnosis of the potentially more aggressive tumors only, with the exclusion of nPnT1PCA, or (iii) the diagnosis of the more advanced neoplasia, that is, st2-4CA.

TIRSYS-A had a higher specificity than PERS for all 3 cancer categories. As for sensitivity, PERS outperformed TIRSYS-A in the allCA and nPnT1PCA groups. Thus, TIRSYS-A performed better if the goal was to find only the cancers that are clinically more relevant. Conversely, if the aim of the indication for FNA is to diagnose all malignant lesions, regardless of their size and potential aggressiveness, the PERS approach appeared superior to TIRSYS-A. The better sensitivity of PERS might partly be explained by the fact that the knowledge of the presence or absence of pathological lymph nodes provided an advantage to the investigators; of the 5 TIRADS, only the AACE/ACE/AME considers this information. The sensitivity of TIRSYS-A was 69.8% for allCA while the specificity was 30.0%. Both values are in the range published in the literature; 54–87 and 28–64%, respectively, for sensitivity and specificity [36].

The investigators' expertise and the TIRSYS-B methodology recommended FNA in a similar proportion of nodules, larger or smaller than 10 mm. The use of a size cutoff is, presently, still controversial. In a recent study, 36% of malignant thyroid lesions did not qualify for FNA [37] by TIRADS; we found a similar rate by using TIRSYS-A (30.2%).

The cancer phenotype categorization used in the present study, that is, allCA, nPnT1PCA, and st2-4CA, is arbitrary, but based on the potential clinical relevance of the malignancies under examination. At one end of the evaluation spectrum, the diagnostic methodology was mainly addressed identifying st2-4CA because these lesions are associated with risk of disease-specific mortality. At the other end, the diagnostic workup aimed at recognizing allCA, regardless of their aggressiveness, and including subgroups at very low mortality risk. Finally, the intermediary goal was to recognize all nPnT1PCA, that is, lesions characterized by potential growth and extra-thyroidal spread. This intermediary goal would decrease the risk of overlooking aggressive thyroid malignancies and may provide a cost-effective approach to balance the timely diagnosis of harmful lesions and the utilization of resources [17, 38]. Indeed, we find that the cancer phenotype grouping used in our study allowed drawing clinical conclusions.

There is currently a pronounced incentive in the international thyroid community to develop one universal TIRADS system [39]. Based on the present work, there are obvious factors which need to be in focus when developing a universal TIRADS: (i) lack of a defined goal which should state if all or only the potentially clinically relevant tumors need to be identified, (ii) the personal judgment, based on clinical expertise, that still calls for the use of

more FNAs than those indicated on the basis of TIRADS, (iii) the different size limits currently proposed by the 5 TIRADS for recommending FNA, and (iv) the controversial approach of TIRADS regarding the subcentimeter nodules with high-risk US features. Notably, the 10-mm limit for FNA, provided by some of the TIRADS classifications, is at least in part arbitrary, as the TNM staging system does not use the 10 mm size cutoff for pT1 differentiated carcinomas [20]. Although the main tables of all 5 TIRADS clearly state that subcentimeter nodules are not candidates for FNA, all 5 TIRADS permit this in "certain cases" [12–16]. While this ambiguity is perfectly understandable, it also causes uncertainty.

The limitations of our study include the relatively low number of nodules in the sample set, particularly in specific subgroups of carcinomas. The benign to malignant ratio in our study has been intentionally set to accommodate the purposes of the investigation and is not suited to draw conclusions on epidemiology of thyroid nodules or thyroid cancer. In the real-world practice, the clinician applies one specific US risk-stratification system and uses his/her own clinical judgment, which cannot be fully simulated in an experimental context. Further, a few data were a priori provided and could not be changed by the investigators. The strengths of the study are that only nodules verified by histology were included, and only highly experienced investigators, from different countries participated in the trial. Due to the recent demonstration of a rather low level of interobserver agreement between different centers [19], this consideration appears extremely relevant. Finally, the methodology of the study was as close as possible to the real-world clinical practice with the use of video records and the possibility of freezing images and reevaluating them in slow motion. This may imply that our results offer a best-case scenario and that they cannot be directly compared with that of others'.

Recommendations for FNA, based on investigators US expertise, demonstrated a better sensitivity for thyroid cancer if all cancers were to be identified. TIRADS methodology was comparably sensitive for advanced tumors and was more specific over the full range of cancers. Thus, personal experience provided a more accurate diagnosis for thyroid malignancies as a whole, missing a lower number of small thyroid cancers, but the TIRADS approach performed with a similar accuracy for the diagnosis of clinically relevant or potentially aggressive lesions, while sparing a relevant number of FNA. Experienced investigators seem to prioritize finding all cancers, even when smaller than 1 cm.

Statement of Ethics

This study was approved by the Regional and Institutional Ethics Committee of the University of Debrecen (number of approval 5350/2019). This study was carried out in accordance with the Declaration of Helsinki. Written informed consent has been obtained from each patient after full explanation of the purpose and nature of all procedures used.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

The authors did not receive any funding.

Author Contributions

Tamas Solymosi: study design, preparation of the 123 video records involved in the study, steering, and manuscript writing. Laszlo Hegedüs: study design, steering, manuscript preparation, manuscript preparation, and final manuscript approval. Steen Bonnema: study design, evaluation of the cases, and approval of the manuscript. Andrea Frasoldati: study design, evaluation of the cases, and approval of the manuscript. Laszlo Jambor: study design, evaluation of the cases, and approval of the manuscript. Gabor L. Kovacs: study design, evaluation of the cases, and approval of the manuscript. Enrico Papini: study design, evaluation of the cases, and approval of the manuscript. Gilles Russ: study design, evaluation of the cases, and approval of the manuscript. Zsolt Karanyi: development of the eCRF and the online evaluation system, and statistical analysis. Endre V. Nagy: study design, steering, and manuscript preparation.

References

1 Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, et al. American Thyroid Association Guidelines Taskforce. Management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid. 2006;16:109–42.
2 Gharib H, Papini E, Valcavi R, Baskin HJ, Crescenzi A, Dottorini ME, et al. American Association of Clinical Endocrinologists and Associazione Medici Endocrinologi medical guidelines for clinical practice for the diagnosis and management of thyroid nodules. Endocr Pract. 2006;12(1):63–102.
3 Mazzaferri EL. Management of a solitary thyroid nodule. N Engl J Med. 1993;328(8):553–9.
4 Hegedüs L. The thyroid nodule. N Engl J Med. 2004;351(17):1764–71.
5 Burman KD, Wartofsky L. Thyroid nodules. N Engl J Med. 2016;374(13):1294–5.

6 Papini E, Guglielmi R, Bianchini A, Crescenzi A, Taccogna S, Nardi F, et al. Risk of malignancy in nonpalpable thyroid nodules: predictive value of ultrasound and color-Doppler features. J Clin Endocrinol Metab. 2002;87(5): 1941–1946. http: //dx.doi.org/10.1210/jcem.87.5.8504.
7 Tollin SR, Mery GM, Jelveh N, Fallon EF, Mikhail M, Blumenfeld W, et al. The use of fine-needle aspiration biopsy under ultrasound guidance to assess the risk of malignancy in patients with a multinodular goiter. Thyroid. 2000;10(3):235–41.
8 Frates MC, Benson CB, Doubilet PM, Kunreuther E, Contreras M, Cibas ES, et al. Prevalence and distribution of carcinoma in patients with solitary and multiple thyroid nodules on sonography. J Clin Endocrinol Metab. 2006;91(9):3411–7.
9 Orell SR, Philips J. The thyroid. Fine needle biopsy and cytological diagnosis of thyroid lesions. Basel: Karger; 1997.
10 Moon WJ, Baek JH, Jung SL, Kim DW, Kim EK, Kim JY, et al. Ultrasonography and the ultrasound-based management of thyroid nodules: consensus statement and recommendations. Korean J Radiol. 2011;12(1):1–14.
11 American Thyroid Association (ATA); Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, et al. Guidelines Taskforce on Thyroid Nodules and Differentiated Thyroid Cancer-Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid. 2009;19:1167–214.
12 Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, 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):1–133.
13 Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedüs L, et al. 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 update. Endocr Pract. 2016;22(5):622–39.
14 Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. J Am Coll Radiol. 2017;14(5): 587–95.
15 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. Eur Thyroid J. 2017; 6(5):225–37.
16 Shin JH, Baek JH, Chung J, Ha EJ, Kim JH, Lee YH, et al. Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology consensus statement and recommendations. Korean J Radiol. 2016;17(3):370–95.
17 Ito Y, Miyauchi A, Oda H. Low-risk papillary microcarcinoma of the thyroid: a review of active surveillance trials. Eur J Surg Oncol. 2018;44(3):307–15.
18 Persichetti A, Di Stasio E, Guglielmi R, Bizzarri G, Taccogna S, Misischi I, et al. Predictive value of malignancy of thyroid nodule ultrasound classification systems: a prospective study. J Clin Endocrinol Metab. 2018;103(4): 1359–68.
19 Persichetti A, Di Stasio E, Coccaro C, Graziano F, Bianchini A, Di Donna V, et al. Inter- and intraobserver agreement in the assessment of thyroid nodule ultrasound features and classification systems: a blinded multicenter study. Thyroid. 2020 Feb;30(2):237–42.
20 Tuttle M, Morris LF, Haugen B, Shah J, Sosa JA, E R, et al. Thyroid differentiated and anaplastic carcinoma. 8th ed. In: Amin MB, Edge SB, Greene F, Byrd D, Brookland RK, Washington MK, et al., editors. AJCC cancer staging manual. New York, NY: Springer International Publishing; 2017.
21 Stevenson M, Nunes T, Heuer C, Marshall J, Sanchez J, Thornton R, et al. epiR: tools for the analysis of epidemiological data. R package version 1.0-2. 2019. Available from: https: //CRAN.R-project.org/package=epiR.
22 R Core Team. R. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2019. Available from: https://www.R-project.org/.
23 Middleton WD, Teefey SA, Reading CC, Langer JE, Beland MD, Szabunio MM, et al. Comparison of performance characteristics of American College of Radiology TI-RADS, Korean Society of Thyroid Radiology TI-RADS, and American Thyroid Association guidelines. AJR Am J Roentgenol. 2018; 210(5):1148–54.
24 Lauria Pantano A, Maddaloni E, Briganti SI, Beretta Anguissola G, Perrella E, Taffon C, et al. Differences between ATA, AACE/ACE/AME and ACR TI-RADS ultrasound classifications performance in identifying cytological high-risk thyroid nodules. Eur J Endocrinol. 2018;178(6):595–603.
25 Grani G, Lamartina L, Ascoli V, Bosco D, Biffoni M, Giacomelli L, et al. Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the "right" TIRADS. J Clin Endocrinol Metab. 2019;104:95–102.
26 Wu XL, Du JR, Wang H, Jin CX, Sui GQ, Yang DY, et al. Comparison and preliminary discussion of the reasons for the differences in diagnostic performance and unnecessary FNA biopsies between the ACR TIRADS and 2015 ATA guidelines. Endocrine. 2019;65(1): 121–31.
27 Ha SM, Baek JH, Na DG, Suh CH, Chung SR, Choi YJ, et al. Diagnostic performance of practice guidelines for thyroid nodules: thyroid nodule size versus biopsy rates. Radiology. 2019;291(1):92–9.
28 Xu T, Wu Y, Wu RX, Zhang YZ, Gu JY, Ye XH, et al. Validation and comparison of three newly-released Thyroid Imaging Reporting and Data Systems for cancer risk determination. Endocrine. 2019;64(2):299–307.
29 Gao L, Xi X, Jiang Y, Yang X, Wang Y, Zhu S, et al. Comparison among TIRADS (ACR TI-RADS and KWAK- TI-RADS) and 2015 ATA Guidelines in the diagnostic efficiency of thyroid nodules. Endocrine. 2019;64(1):90–6.
30 Skowrońska A, Milczarek-Banach J, Wiechno W, Chudziski W, Zach M, Mazurkiewicz M, et al. Accuracy of the European Thyroid Imaging Reporting and Data System (EU-TI-RADS) in the valuation of thyroid nodule malignancy in reference to the post-surgery histological results. Pol J Radiol. 2018;83:579–86.
31 Sahli ZT, Karipineni F, Hang JF, Canner JK, Mathur A, Prescott JD, et al. The association between the ultrasonography TIRADS classification system and surgical pathology among indeterminate thyroid nodules. Surgery. 2019;165(1): 69–74.
32 Hoang JK, Middleton WD, Farjat AE, Teefey SA, Abinanti N, Boschini FJ, et al. Interobserver variability of sonographic features used in the American college of radiology thyroid imaging reporting and data system. Am J Roentgenol. 2018;211(1):162–7.
33 Tessler N, William D, Middleton WD, Grant EG. Thyroid imaging reporting and data system (TI-RADS): a user's guide. Radiology. 2018;287:29–36.
34 Chng CL, Tan HC, Too CW, Lim WY, Chiam PPS, Zhu L, et al. Diagnostic performance of ATA, BTA and TIRADS sonographic patterns in the prediction of malignancy in histologically proven thyroid nodules. Singapore Med J. 2018;59(11):578–83.
35 Trimboli P, Ngu R, Royer B, Giovanella L, Bigorgne C, Simo R, et al. A multicentre validation study for the EU-TIRADS using histological diagnosis as a gold standard. Clin Endocrinol. 2019;91(2):340–7.
36 Castellana M, Castellane C, Treglia G, Giorgino F, Giovanella L, Russ G, et al. Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. A metaanalysis. J Clin Endocrinol Metab. 2020;105:1659–69.
37 Dobruch-Sobczak K, Adamczewski Z, Szczepanek-Parulska E, Migda B, Woliński K, Krauze A, et al. Histopathological verification of the diagnostic performance of the EU-TIRADS classification of thyroid nodules-results of a multicenter study performed in a previously iodine-deficient region. J Clin Med. 2019;8(11):1781.
38 Miyauchi A, Ito Y. Conservative surveillance management of low-risk papillary thyroid microcarcinoma. Endocrinol Metab Clin North Am. 2019;48(1):215–26.
39 Oncology Central. Exploring the Thyroid Imaging Reporting and Data System (TI-RADS): an interview with Franklin Tessler. Available from: https: //www.oncology-central.com/disease-area/endocrine/exploring-thyroid-imaging-reporting-data-system-ti-rads-interview-franklin-tessler/. Accessed 2018 Aug 7.