European Radiology (2021) 31:2877–2885
https://doi.org/10.1007/s00330-020-07384-6

HEAD AND NECK

Unnecessary thyroid nodule biopsy rates under four ultrasound risk stratification systems: a systematic review and meta-analysis

Pyeong Hwa Kim[1] · Chong Hyun Suh[1] · Jung Hwan Baek[1] · Sae Rom Chung[1] · Young Jun Choi[1] · Jeong Hyun Lee[1]

Received: 27 May 2020 / Revised: 27 August 2020 / Accepted: 6 October 2020 / Published online: 15 October 2020

Abstract

Objectives To summarize and compare unnecessary biopsy rates and diagnostic performance in the examination of thyroid nodules according to four representative US-based risk stratification systems.

Methods MEDLINE/PubMed and EMBASE databases were searched to identify original articles investigating unnecessary biopsy rates according to at least one of the following guidelines: ACR-TIRADS, ATA, EU-TIRADS, and K-TIRADS. The unnecessary biopsy rates for each risk stratification system were pooled using a random-effects model. Meta-regression analyses were performed to explore heterogeneity. Diagnostic odds ratios (DORs) for the appropriate selection of thyroid nodules for fine-needle aspiration were also pooled using a bivariate random-effects model.

Results Eight articles including 13,092 thyroid nodules met the eligibility criteria and were included. The pooled unnecessary biopsy rates of ACR-TIRADS, ATA, EU-TIRADS, and K-TIRADS were 25% (95% CI, 22–29%), 51% (95% CI, 44–58%), 38% (95% CI, 16–66%), and 55% (95% CI, 42–67%), respectively. The pooled unnecessary biopsy rate of ACR-TIRADS was significantly lower than that of ATA ($p < .001$) and K-TIRADS ($p < .001$), and also lower than that of EU-TIRADS, but not reaching statistical significance ($p = .087$). The pooled DORs of ACR-TIRADS, ATA, and K-TIRADS were 5.9 (95% CI, 3.6–9.6), 6.3 (95% CI, 4.5–8.8), and 4.5 (95% CI, 1.7–11.6), respectively, with the differences not being statistically significant.

Conclusions ACR-TIRADS showed a lower unnecessary biopsy rate than the other risk stratification systems albeit DOR was comparable between ACR-TIRADS, ATA, and K-TIRADS. Future revisions of each system should be made by referring to ACR-TIRADS to reduce unnecessary biopsy rates.

Key Points

- *The pooled unnecessary biopsy rates of ACR-TIRADS, ATA, EU-TIRADS, and K-TIRADS were 25% (95% CI, 22–29%), 51% (95% CI, 44–58%), 38% (95% CI, 16–66%), and 55% (95% CI, 42–67%), respectively.*
- *The pooled unnecessary biopsy rate of ACR-TIRADS was significantly lower than that of ATA ($p < .001$) and K-TIRADS ($p < .001$).*
- *The pooled DORs of ACR-TIRADS, ATA, and K-TIRADS were 5.9 (95% CI, 3.6–9.6), 6.3 (95% CI, 4.5–8.8), and 4.5 (95% CI, 1.7–11.6), respectively, with the differences not being statistically significant.*

Keywords Thyroid · Thyroid neoplasm · Ultrasonography · Biopsy · Meta-analysis

Abbreviations

ACR	American College of Radiology
ATA	American Thyroid Association
DOR	Diagnostic odds ratio
EU-TIRADS	2017 European Thyroid Association TIRADS
FNAB	Fine-needle aspiration biopsy
K-TIRADS	2016 Korean Thyroid Association/Korean Society of Thyroid Radiology (KTA/KSThR) TIRADS
TIRADS	Thyroid Imaging Reporting and Data System
US	Ultrasound

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-020-07384-6) contains supplementary material, which is available to authorized users.

✉ Jung Hwan Baek
radbaek@naver.com

[1] Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 86 Asanbyeongwon-Gil, Songpa-Gu, Seoul 05505, South Korea

Introduction

Ultrasonography (US) is the diagnostic modality of choice for the characterization of thyroid nodules [1]. To date, several international societies have developed US-based risk stratification systems, also known as Thyroid Imaging Reporting and Data Systems (TIRADS), to maximize the diagnostic performance of thyroid US and identify those thyroid nodules that should be biopsied [2–5]. In 2015, the American Thyroid Association (ATA) proposed a qualitative US-based five-tier risk stratification system [3]. The Korean Thyroid Association/Korean Society of Thyroid Radiology (KTA/KSThR) also proposed a risk stratification system (K-TIRADS), which is a pattern-based qualitative system defining four categories with different risks of malignancy [4]. In 2017, the American College of Radiology (ACR) proposed a five-tier risk stratification system (ACR-TIRADS) that was characterized by its quantitative scoring method [2]. In the same year, the European Thyroid Association also proposed a pattern-based qualitative system defining four categories (EU-TIRADS) [5].

Although fine-needle aspiration biopsy (FNAB) has a crucial role in the diagnosis of thyroid cancer, there has been an emphasis on reducing the number of excessive biopsies, which can lead to overdiagnosis and overtreatment, especially considering the less invasive nature of thyroid cancer [6–10]. In this regard, the emphasis in the evaluation of the current TIRADS has shifted from simply evaluating the diagnostic performance to the inclusion of unnecessary biopsy rates. However, there is considerable discordance in the recommended criteria for suspicious US patterns and size cut-offs for FNAB between the TIRADS [11, 12]. In this context, although many authors have attempted to evaluate and compare the unnecessary biopsy rates and diagnostic performance of each system [11–15], the presence of substantial between-study heterogeneity still remains which makes the interpretation difficult. Therefore, we considered it is timely and necessary to summarize the currently available data to provide valuable information for clinical practice and future revisions of the current TIRADS.

Thus, the present systematic review and meta-analysis aimed to evaluate the diagnostic performance and unnecessary thyroid nodule biopsy rates under four representative US-based risk stratification systems: ACR-TIRADS, ATA, EU-TIRADS, and K-TIRADS.

Materials and methods

This study was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [16].

Search strategy and eligibility criteria

A literature search of the MEDLINE/PubMed and EMBASE databases was conducted using pertinent MeSH or EMTREE terms with common keywords for relevant articles up until August 5, 2019. The search terms were as follows: ((thyroid)) AND ((thyroid imaging reporting and data system) OR (TIRADS) OR (TI-RADS) OR (guideline)) AND ((American Thyroid Association) OR (ATA) OR (American College of Radiology) OR (ACR) OR (Europe*) OR (EU-TIRADS) OR (Korea*) OR (K-TIRADS)). The search was limited to English-language publications, but was not limited by human or animal studies, or publication date.

After eliminating duplicates, articles were screened according to their title and abstract. Full-text articles were then thoroughly assessed according to the following eligibility criteria: (a) population: patients who underwent US examinations for thyroid nodules; (b) index test: US-based risk stratification systems according to at least one of the following guidelines: ACR-TIRADS [2], ATA [3], EU-TIRADS [5], and K-TIRADS [4]; (c) reference standard: pathological diagnosis or imaging follow-up; (d) outcomes: unnecessary biopsy rate; (e) study design: not limited. Studies were excluded if any of the following criteria were met: (a) studies including non-consecutive nodules; (b) studies not providing sufficient details to calculate the unnecessary biopsy rate; (c) review articles; (d) case reports or case series including fewer than ten patients; (e) conference abstracts; (f) letters, editorials, and comments; (g) animal studies; (h) studies with a partially overlapping patient cohort (for studies with an overlapping study population, the study with the largest population was selected); (i) studies conducted with a pediatric population; or (j) studies using a pathology reporting system other than the Bethesda classification system [17]. The literature search and application of the criteria were conducted independently by two authors (P.H.K. and C.H.S., with 3 and 8 years of experience in performing thyroid US and interventional procedures, respectively), and any discrepancies were resolved through discussion and consensus with a third author (J.H.B., with 21 years of experience in performing thyroid US and interventional procedures).

Data extraction and quality assessment

A standardized extraction form was used to obtain the following information from the selected studies: (a) study characteristics: institution, study period, study design (prospective vs. retrospective; single-center vs. multicenter), reference standard, and blinding to the reference standard; (b) demographic and clinical characteristics: total number of patients, total number of nodules and malignant nodules, mean age (range), and proportion of female patients; (c) unnecessary biopsy rates; and (d) diagnostic performance of each risk stratification

system in the form of a 2 × 2 table, with indication for FNA as the index test [1]. The quality of the selected studies was investigated using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) [18].

Data synthesis and analysis

The primary outcome of this meta-analysis was the unnecessary biopsy rate, defined as the proportion of benign nodules among the biopsied nodules. Meta-analytic pooling was based on the inverse variance method for calculating weights, and their 95% confidence intervals (CIs) were determined using DerSimonian–Laird random-effects modeling. Heterogeneity across studies was assessed using the Q test and I^2 statistic, with $I^2 > 50\%$ being taken to indicate the presence of heterogeneity [19–21].

The secondary outcome was the diagnostic odds ratio (DOR) of each system with indication for FNA as the index test. For the meta-analytic pooling of DOR, a bivariate random-effects model with two-by-two tables including true-positive (TP; nodules for which FNAB was indicated and the nodule was found to be malignant), false-positive (FP; nodules for which FNAB was indicated and the nodule was found to be benign), false-negative (FN; nodules for which FNAB was not indicated yet the nodule was found to be malignant), and true-negative (TN; nodules in which FNAB was not indicated and the nodule was found to be benign) findings was constructed for each study. In addition, the pooled sensitivity and specificity and their 95% CIs were calculated, and a coupled forest plot was constructed [20–24]. Indirect comparisons of unnecessary biopsy rates and DORs between the risk stratification systems were performed using a Wald-type chi-square test with multiplicity adjustment, and the regression coefficient was obtained to estimate the intervention effect from a reference group [25, 26]. Statistical analyses were conducted by one of the authors (C.H.S., with 8 years of experience in performing systematic reviews and meta-analyses) using the "metandi" and "midas" modules in Stata 15.0 (StataCorp), and the "meta", "metafor", and "mada" packages in R software (version 3.6.2.; R Foundation for Statistical Computing).

Results

Literature search

A flow chart summarizing the publication selection process is presented in Fig. 1. A total of 411 non-duplicate studies were identified. Of these, 307 articles were excluded on the basis of their titles and abstracts because they were not in the field of interest ($n = 232$), or they were guidelines ($n = 63$), reviews ($n = 8$), case reports ($n = 2$), an erratum ($n = 1$), or an animal study ($n = 1$). Subsequently, 104 potentially eligible full-text articles were assessed according to the eligibility criteria, and a further 96 studies were excluded because they included non-consecutive nodules ($n = 29$), did not provide sufficient details to calculate the unnecessary biopsy rate ($n = 29$), did not use any of the four risk stratification systems of interest (ACR-TIRADS, ATA, EU-TIRADS, or K-TIRADS; $n = 11$), used data included in subsequent articles ($n = 10$), were not in the field of interest ($n = 9$), included inseparable adult and pediatric patients ($n = 6$), used a histopathologic reporting system other than the Bethesda system ($n = 1$), or did not include histopathology as a reference standard ($n = 1$). Consequently, a total of eight articles including 13,092 thyroid nodules met the eligibility criteria and were included in the analysis [11, 13–15, 27–30].

Characteristics of the included studies

The detailed study characteristics are summarized in Table 1. One of eight studies was of a prospective design [28], and three were multicenter studies [14, 15, 27, 31–33]. The number of included patients ranged from 127 to 3190, and the mean patient age ranged from 44 to 55 years. The proportion of female patients in each study ranged from 61.2 to 86.6%, and the proportion of female patients in the pooled population was 77.7% (8280 out of 10,654; excluding Wu et al [30] in which the data was not available). The proportion of malignant nodules in each study varied from 13.2 to 53.0%, with the pooled proportion being 29.2% (3826 out of 13,092). Unnecessary biopsy rates according to ACR-TIRADS, ATA, EU-TIRADS, and K-TIRADS were reported in eight [11, 13–15, 27–30], five [13, 14, 27, 29, 30], two [11, 15], and five [11, 13–15, 27] studies, respectively.

Quality assessment

The results of the quality assessment based on QUADAS-2 criteria are shown in Supplementary Figure S1. Three studies [11, 14, 28] had an unclear risk of bias in the index test domain because of no or unclear blinding to the reference standard during the US examinations. All eight studies [11, 13–15, 27–30] had an unclear risk of bias in the reference standard domain because of no or unclear blinding to the index test during pathologic evaluation. Additionally, three studies [11, 14, 27] had a high risk and one study [28] an unclear risk of bias in the flow and timing domain because of inconsistency or unclear consistency on the reference standard for diagnosing benign nodules across the study population. Three studies [11, 15, 28] had a high concern on the applicability of the index test because of single or unreported numbers of readers for the US images. One study [28] had an unclear concern on the applicability of the reference standard because of no

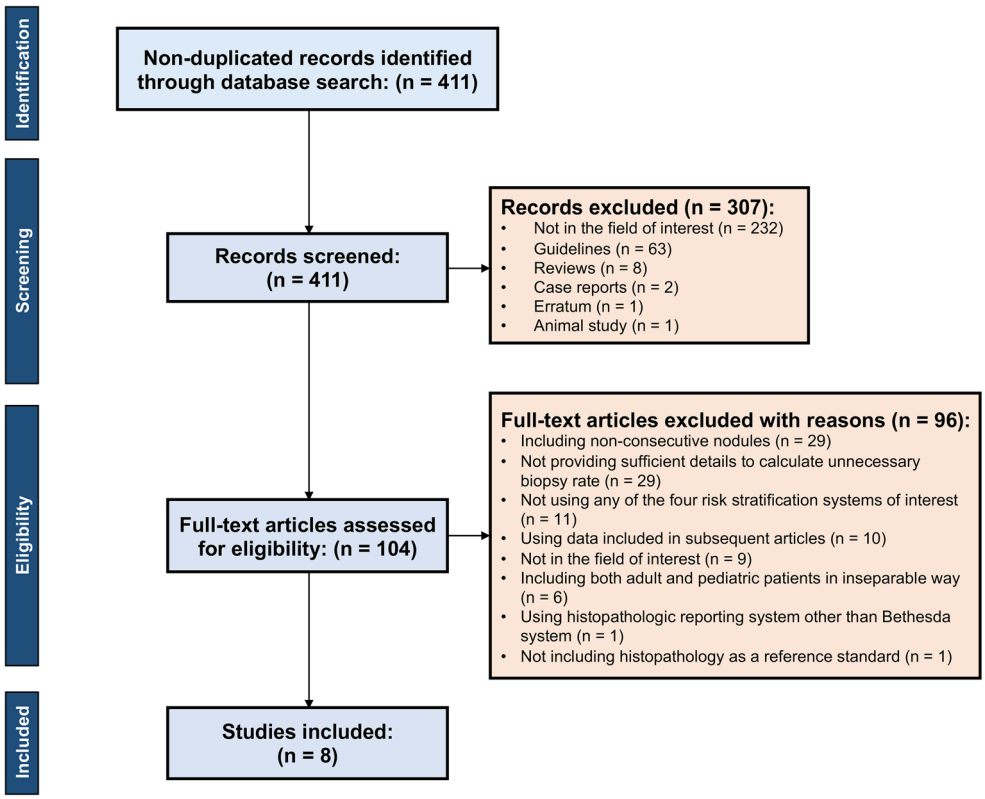

Fig. 1 Flow chart of the publication selection process

information on how the tissue specimens were examined. There were no concerns on the applicability of patient selection.

Unnecessary biopsy rates

The pooled unnecessary biopsy rates of ACR-TIRADS, ATA, EU-TIRADS, and K-TIRADS were 25% (95% CI, 22–29%), 51% (95% CI, 44–58%), 38% (95% CI, 16–66%), and 55% (95% CI, 42–67%), respectively (Fig. 2). There was substantial heterogeneity observed with all four risk stratification systems (I^2 > 50%). Meta-regression analysis identified that the pooled unnecessary biopsy rate of ACR-TIRADS was significantly lower than that of ATA (OR [95% CI], 1.29 [1.15–1.44]; p < .001) and K-TIRADS (OR [95% CI], 1.34 [1.20–1.49]; p < .001; Table 2), and also lower than that of EU-TIRADS, but not reaching statistical significance (p = .087).

Diagnostic performance

The pooled DORs of each system for selecting thyroid nodules for FNA are depicted in Fig. 3. Meta-analytic pooling was not possible for EU-TIRADS as data were available for only two studies [11, 15]. The pooled DORs of ACR-TIRADS, ATA, and K-TIRADS were 5.9 (95% CI, 3.6–9.6), 6.3 (95% CI, 4.5–8.8), and 4.5 (95% CI, 1.7–11.6), respectively. Substantial heterogeneity was observed with all three risk stratification systems (I^2 > 50%). Indirect comparisons showed that the DOR of ACR-TIRADS was not statistically different to that of ATA-TIRADS (p = .816) and K-TIRADS (p = .524). Sensitivity analysis excluding Xu T et al [15] due to its relatively lower DOR showed the modest decrease of heterogeneity in ACR-TIRADS (I^2, 95% to 76%) and marked decrease of heterogeneity in K-TIRADS (I^2, 97% to 0%), with the pooled DORs of ACR-TIRADS, ATA, and K-TIRADS to be 7.0 (95% CI, 5.3–9.2), 6.3 (95% CI, 4.5–8.8), and 6.3 (95% CI, 5.0–7.9), respectively. Indirect comparisons also showed that the DOR of ACR-TIRADS was not statistically different to that of ATA-TIRADS (p = .605) and K-TIRADS (p = .658). The pooled sensitivities of ACR-TIRADS, ATA, and K-TIRADS were 75% (95% CI, 61–84%), 93% (95% CI, 88–95%), and 91% (95% CI, 80–96%), respectively, while the pooled specificities were 67% (95% CI, 61–73%), 34% (95% CI, 26–42%), and 32% (95% CI, 25–39%), respectively. Of note, ACR-TIRADS showed significantly lower sensitivity compared with ATA (p < .01) and K-TIRADS (p < .01), but higher specificity compared with ATA (p < .01) and K-TIRADS (p < .01) (Supplementary Table S1).

Discussion

The present meta-analysis investigated the unnecessary biopsy rates of each thyroid nodule risk stratification system using

Table 1 Characteristics of the included studies

Author (year of publication)	Country	Study period	No. of patients	Mean age (range)	M:F	No. of nodules	Malignant nodules (%)	Study design	Minimum nodule size for inclusion (mm)	Risk stratification system				Reference standard		
										ACR	ATA	EU	K	Surgery	Biopsy	Follow-up[b]
Ha EJ et al (2018) [14]	South Korea	2010,1–2011,5	1802	51.2 (13–79)	415:1387	2000	22.7	Multicenter, retrospective	10	Yes	Yes	No	Yes	Yes	Yes[a]	Yes
Ha EJ et al (2018) [27]	South Korea	2013,6–2015,5	750	NA (9–81)	156:594	902	29.5	Multicenter, retrospective	5	Yes	Yes	No	Yes	Yes	Yes[a]	Yes
Ha SM et al (2019) [13]	South Korea	2013,1–2013,12	3190	43.5 (14–94)	673:2517	3323	25.8	Single-center, retrospective	All	Yes	Yes	No	Yes	Yes	Yes[a]	Yes
Jabar ASS et al (2019) [28]	India	2017,12–2018,8	127	NA	17:110	127	18.1	Single-center, prospective	NA	Yes	No	No	No	Yes	Yes	No
Ruan JL et al (2019) [29]	China	2016,5–2017,12	918	45.7 (14–78)	356:562	1001	39.2	Single-center, retrospective	5	Yes	Yes	No	No	Yes	No	No
Wu XL et al (2019) [30]	China	2016,4–2017,3	894	NA	NA	1000	53.0	Single-center, retrospective	All	Yes	Yes	No	No	Yes	Yes[a]	Yes
Xu T et al (2019) [15]	China	2014,1–2017,10	2031	47.7	415:1616	2465	40.8	Multicenter, retrospective	All	Yes	No	Yes	Yes	Yes	Yes	Yes
Yoon SJ et al (2019) [11]	South Korea	2011,1–2016,12	1836	55.1 (9–92)	342:1494	2274	13.2	Single-center, retrospective	10	Yes	No	Yes	Yes	Yes	Yes	No

ACR, 2017 American College of Radiology; *ATA*, 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer; *EU*, 2017 European Thyroid Association; *K*, 2016 Korean Thyroid Association/Korean Society of Thyroid Radiology; *FNA*, fine-needle aspiration; *NA*, not applicable

[a] Fine-needle aspiration biopsy (FNAB) and core needle biopsy (CNB) were used to obtain the specimen. In other included studies, only FNAB was used

[b] In all studies using follow-up as a reference standard, thyroid nodules with initial benign results on biopsy and decreased or stable size at follow-up US more than 12 months later were finally confirmed as benign

Table 2 Results of the meta-regression for unnecessary biopsy rates

Variables	Subgroups	OR	95% CI	*p* value
TIRADS	ACR	REF		
	ATA	1.29	1.15–1.44	< .001
	EU	1.14	0.98–1.33	.087
	K	1.34	1.20–1.49	< .001
Single vs. multicenter	Single center	REF		
	Multicenter	0.94	0.80–1.11	.44
Female proportion	< 78%	REF		
	≥ 78%	0.94	0.78–1.14	.51
Malignant nodule proportion	< 30%	REF		
	≥ 30%	1.00	0.99–1.01	.42
Inclusion of follow-up in reference standard	No	REF		
	Yes	0.89	0.75–1.06	.19

OR, odds ratio; *CI*, confidence interval; *ACR*, 2017 American College of Radiology; *ATA*, 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer; *EU*, 2017 European Thyroid Association; *K*, 2016 Korean Thyroid Association/Korean Society of Thyroid Radiology; *REF*, reference category

eight studies including 13,092 thyroid nodules. The unnecessary biopsy rate was lower with ACR-TIRADS (25%) than with ATA (51%), EU-TIRADS (38%), or K-TIRADS (55%), with this finding being confirmed in the meta-regression analysis. The DOR was comparable between the risk stratification systems. Considering our results and the clinical importance of the unnecessary biopsy rate in the workup of thyroid nodules, future revisions of each system to reduce

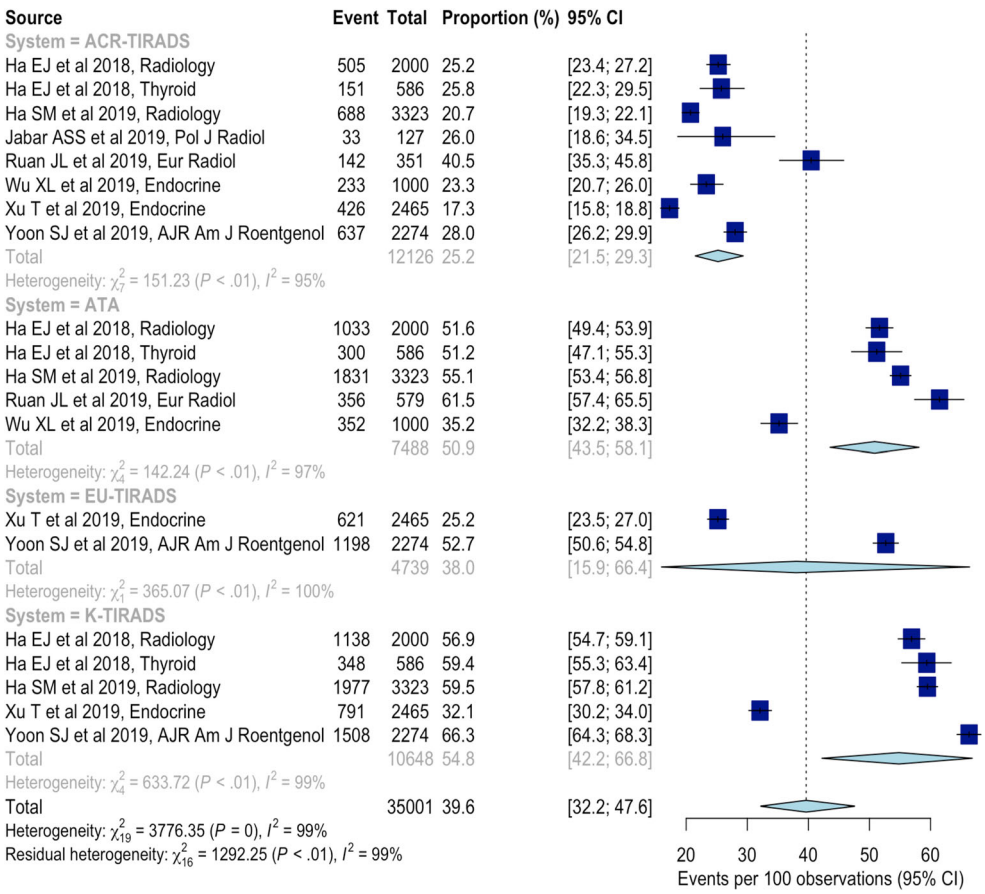

Fig. 2 Unnecessary biopsy rates of the four risk stratification systems

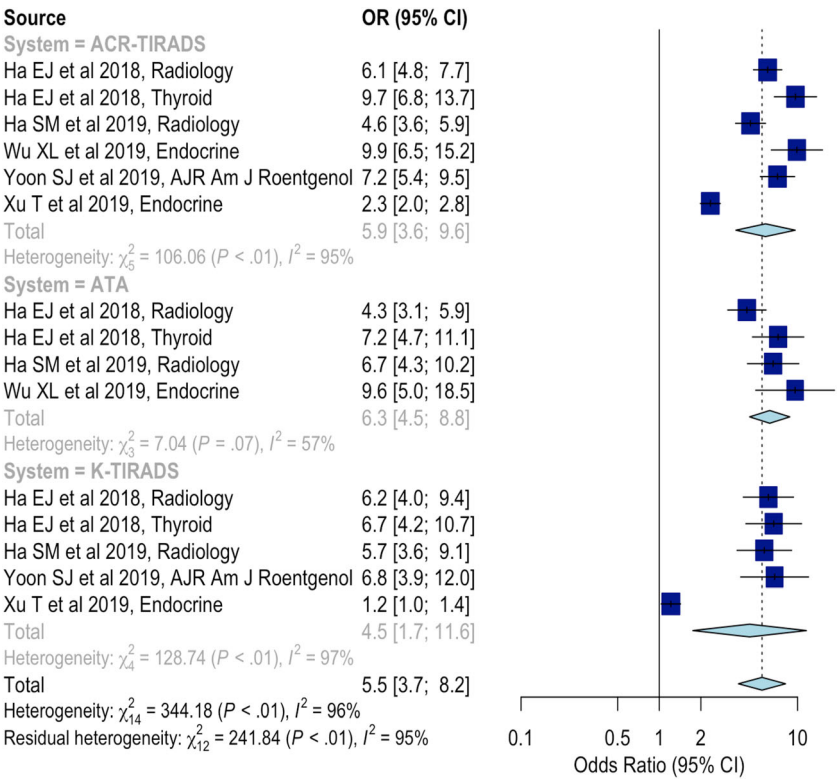

Fig. 3 Diagnostic odds ratios of the three risk stratification systems

unnecessary biopsy rates should be made by referring to ACR-TIRADS.

In our meta-analysis, ACR-TIRADS showed the lowest unnecessary biopsy rate among the four risk stratification systems, which is concordant with previous studies [12, 14, 32]. The reason for this low rate can be explained by the minimum FNAB-recommended nodule size with a discordant risk of malignancy in each category. Indeed, a simulation study conducted by Ha SM et al demonstrated that the unnecessary biopsy rates of ATA and K-TIRADS became similar to that of ACR-TIRADS (21%) when the ACR-TIRADS nodule size cut-offs were applied to each category (ATA, 55% to 20%; K-TIRADS, 60% to 26%) [13]. This indicates that unnecessary biopsy rates may be largely determined by the nodule size cut-off for FNAB. In detail, the risks of malignancy and size cut-offs for FNAB in nodules with intermediate suspicion are 5–20% and 15 mm for ACR-TIRADS, 10–20% and 10 mm for ATA, 6–17% and 15 mm for EU-TIRADS, and 15–50% and 10 mm for K-TIRADS [2–5, 12]. These data show that ACR-TIRADS, ATA, and EU-TIRADS assume similar risks of malignancy, but that ATA sets a smaller size cut-off for FNAB. K-TIRADS assumes a wide range in the risk of malignancy (15–50%) and a 10-mm size cut-off for FNAB. For low-suspicion nodules, the risks of malignancy and size cut-offs for FNAB are 5% and 25 mm for ACR-TIRADS, 5–10% and 15 mm for ATA, 3–15% and 15 mm for EU-TIRADS, and 2–4% and 20 mm for K-TIRADS, showing that the four systems assume a similar risk of malignancy, but that ACR-TIRADS has the largest size cut-off for FNAB. Furthermore, Yim Y et al reported a high concordance between ACR-TIRADS, ATA, and K-TIRADS for high- or intermediate-suspicion nodules, indicating that the size cut-off for FNAB is the main factor influencing diagnostic performance [31]. Therefore, an understanding of the impact of size cut-offs for each category seems necessary for future TIRADS.

Our analysis showed that ACR-TIRADS showed comparable DOR, but lower sensitivity and higher specificity to ATA and K-TIRADS. These differences were also reported in the previous studies [12, 32]. This can be at least partially explained by the nodule size cut-off for FNAB, as elucidated by the simulation study by Ha SM et al [13]. In their study, when similar nodule size cut-offs to those used in ACR-TIRADS were applied to each category, the sensitivity of ATA and K-TIRADS decreased, but the specificity and accuracy increased (ATA: sensitivity, 92% to 61%; specificity, 34% to 76%; accuracy, 44% to 73%; K-TIRADS: sensitivity, 94% to 64%; specificity, 29% to 69%; accuracy, 39% to 68%).

Recently, many efforts have been made to improve the risk stratification systems for thyroid nodules [11–13]. In current practice, the mortality rates of thyroid cancer have not changed, although there has been an increasing incidence of thyroid cancer [9, 10], implying a tendency to overdiagnosis. Therefore, an optimal risk stratification system requires both low rates of unnecessary biopsies and high discriminatory power to select nodules requiring FNAB, thereby reducing

patients' discomfort and anxiety, and reducing medical costs associated with excessive biopsies. Thus, we evaluated the current risk stratification systems in terms of unnecessary biopsy rates and DOR to measure the discriminatory power of the diagnostic tests. As the DOR is independent of the frequency of events in the study population (e.g., the proportion of malignant nodules in each study) [33, 34], it can minimize associated bias. Furthermore, DOR is a single indicator that makes comparisons between diagnostic tests simple. Indeed, the conventional indicators that have been used to evaluate TIRADS (e.g., sensitivity and specificity) explain only a part of the diagnostic performance and are thus not decisive by themselves, making it difficult to simply rank different TIRADS. Therefore, the use of DOR seems appropriate in our study, and it may also be useful in future research. Considering our results, future revisions should take reducing overdiagnosis into account, thus minimizing unnecessary biopsies by referring to ACR-TIRADS.

However, it should be also emphasized that just reducing unnecessary biopsy rates is not always a right answer. In other words, reducing unnecessary biopsy rates may adversely increase the risk of missed malignancy. Indeed, we showed that ACR-TIRADS demonstrated the lowest sensitivity (75%) among the risk stratification systems. Of course, the probability of malignancy among the examined nodules is low, and one retrospective study reported that only 1.2% (17/1382) of nodules in which FNAB was not required according to ACR-TIRADS was confirmed as malignancy [35]. However, to our knowledge, there is no large prospective study evaluating whether reducing unnecessary biopsy rates is indeed beneficial in terms of cost-effectiveness without a negative impact on survival. Further studies seem to be necessary to clarify this issue.

Our study has several limitations of note. First, all studies except one were retrospective, implying a potential misclassification due to unstandardized image acquisition during the examination. Second, the included studies presented heterogenous minimum nodule size cut-offs for inclusion, and therefore a study-level meta-analysis of nodules larger than 1 cm was not possible. In addition, national/institutional policies for biopsy might act as a confounder. Third, the included studies were performed in tertiary referral hospitals, and therefore the data presented in this study might not reflect the actual primary care setting. Fourth, the influence of interobserver variability and clinical expertise could not be evaluated. Finally, there were substantial heterogeneity noted both in the pooled unnecessary biopsy rates and DOR. To overcome this, we performed meta-regression and sensitivity analyses, but heterogeneity was not much resolved. Those might be due to inconsistent minimum nodule size cut-offs for the inclusion and heterogenous classification of the nodules between the studies. In particular, follicular neoplasms were regarded as indeterminate cytology and excluded from the analysis in the study by Wu et al [30] but were included and classified based on their surgical pathology in some studies [13, 14, 27]. These unresolved heterogeneities might affect the credibility of the results.

In conclusion, ACR-TIRADS showed a lower unnecessary biopsy rate than the other risk stratification systems albeit DOR was comparable between ACR-TIRADS, ATA, and K-TIRADS. Future revisions of each system should be made by referring to ACR-TIRADS to reduce unnecessary biopsy rates.

Funding The authors state that this work has not received any funding.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Jung Hwan Baek.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was not required for this study because this study is a systematic review and meta-analysis.

Ethical approval Institutional Review Board approval was not required for this study because this study is a systematic review and meta-analysis.

Methodology
- Meta-analysis
- Performed at one institution

References

1. Ha EJ, Lim HK, Yoon JH et al (2018) Primary imaging test and appropriate biopsy methods for thyroid nodules: guidelines by Korean Society of Radiology and National Evidence-Based Healthcare Collaborating Agency. Korean J Radiol 19:623–631
2. Tessler FN, Middleton WD, Grant EG et al (2017) ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee. J Am Coll Radiol 14:587–595
3. Haugen BR, Alexander EK, Bible KC et al (2016) 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 26:1–133
4. Shin JH, Baek JH, Chung J et al (2016) Ultrasonography diagnosis and imaging-based management of thyroid nodules: revised Korean Society of Thyroid Radiology Consensus statement and recommendations. Korean J Radiol 17:370–395
5. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L (2017) European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. Eur Thyroid J 6:225–237
6. Kim BW, Yousman W, Wong WX, Cheng C, McAninch EA (2016) Less is more: comparing the 2015 and 2009 American

Thyroid Association guidelines for thyroid nodules and cancer. Thyroid 26:759–764
7. Kim TY, Shong YK (2017) Active surveillance of papillary thyroid microcarcinoma: a mini-review from Korea. Endocrinol Metab (Seoul) 32:399–406
8. Oda H, Miyauchi A, Ito Y et al (2016) Incidences of unfavorable events in the management of low-risk papillary microcarcinoma of the thyroid by active surveillance versus immediate surgery. Thyroid 26:150–155
9. Davies L, Welch HG (2014) Current thyroid cancer trends in the United States. JAMA Otolaryngol Head Neck Surg 140:317–322
10. Ahn HS, Kim HJ, Welch HG (2014) Korea's thyroid-cancer "epidemic"–screening and overdiagnosis. N Engl J Med 371:1765–1767
11. Yoon SJ, Na DG, Gwon HY et al (2019) Similarities and differences between thyroid imaging reporting and data systems. AJR Am J Roentgenol 213:W76–W84
12. Grani G, Lamartina L, Ascoli V et al (2019) Reducing the number of unnecessary thyroid biopsies while improving diagnostic accuracy: toward the "Right" TIRADS. J Clin Endocrinol Metab 104:95–102
13. Ha SM, Baek JH, Na DG et al (2019) Diagnostic performance of practice guidelines for thyroid nodules: thyroid nodule size versus biopsy rates. Radiology 291:92–99
14. Ha EJ, Na DG, Baek JH, Sung JY, Kim JH, Kang SY (2018) US fine-needle aspiration biopsy for thyroid malignancy: diagnostic performance of seven society guidelines applied to 2000 thyroid nodules. Radiology 287:893–900
15. Xu T, Wu Y, Wu RX et al (2019) Validation and comparison of three newly-released Thyroid Imaging Reporting and Data Systems for cancer risk determination. Endocrine 64:299–307
16. Liberati A, Altman DG, Tetzlaff J et al (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 151:W65–W94
17. Cibas ES, Ali SZ (2017) The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid 27:1341–1346
18. Whiting PF, Rutjes AW, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529–536
19. Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560
20. Kim KW, Lee J, Choi SH, Huh J, Park SH (2015) Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-Part I. General guidance and tips. Korean J Radiol 16:1175–1187
21. Lee J, Kim KW, Choi SH, Huh J, Park SH (2015) Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-Part II. Statistical methods of meta-analysis. Korean J Radiol 16:1188–1196
22. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 58:982–990
23. Rutter CM, Gatsonis CA (2001) A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 20:2865–2884
24. Suh CH, Park SH (2016) Successful publication of systematic review and meta-analysis of studies evaluating diagnostic test accuracy. Korean J Radiol 17:5–6
25. Higgins JP, Thompson SG (2004) Controlling the risk of spurious findings from meta-regression. Stat Med 23:1663–1682
26. Knapp G, Hartung J (2003) Improved tests for a random effects meta-regression with a single covariate. Stat Med 22:2693–2710
27. Ha EJ, Na DG, Moon WJ, Lee YH, Choi N (2018) Diagnostic performance of ultrasound-based risk-stratification systems for thyroid nodules: comparison of the 2015 American Thyroid Association guidelines with the 2016 Korean Thyroid Association/Korean Society of Thyroid Radiology and 2017 American Congress of Radiology guidelines. Thyroid 28:1532–1537
28. Jabar ASS, Koteshwara P, Andrade J (2019) Diagnostic reliability of the thyroid imaging reporting and data system (TI-RADS) in routine practice. Pol J Radiol 84:274–280
29. Ruan JL, Yang HY, Liu RB et al (2019) Fine needle aspiration biopsy indications for thyroid nodules: compare a point-based risk stratification system with a pattern-based risk stratification system. Eur Radiol 29:4871–4878
30. Wu XL, Du JR, Wang H et al (2019) 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 65:121–131
31. Yim Y, Na DG, Ha EJ et al (2020) Concordance of three international guidelines for thyroid nodules classified by ultrasonography and diagnostic performance of biopsy criteria. Korean J Radiol 21:108–116
32. Castellana M, Castellana C, Treglia G et al (2020) Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. J Clin Endocrinol Metab 105:1659–1669
33. Eusebi P (2013) Diagnostic accuracy measures. Cerebrovasc Dis 36:267–272
34. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56:1129–1135
35. Koseoglu Atilla FD, Ozgen Saydam B, Erarslan NA et al (2018) Does the ACR TI-RADS scoring allow us to safely avoid unnecessary thyroid biopsy? single center analysis in a large cohort. Endocrine 61:398–402