Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

# Performance of EU-TIRADS in malignancy risk stratification of thyroid nodules. A meta-analysis.

Marco Castellana, MD [1], Giorgio Grani, PhD [2], Maija Radzina, PhD [3], Vito Guerra, BSc [1], Luca Giovanella, PhD [4,5], Maurilio Deandrea, MD [6], Rose Ngu, MD [7], Cosimo Durante, PhD [2], Pierpaolo Trimboli, MD [4,8]

**Affiliations**

[1] National Institute of Gastroenterology "S. de Bellis", Research Hospital, Castellana Grotte, Italy.

[2] Department of Translational and Precision Medicine, Sapienza University of Rome, Rome, Italy.

[3] Diagnostic Radiology Institute Paula Stradina Clinical University Hospital, University of Latvia, Radiology Research laboratory, Riga Stradins University, Riga, Latvia.

[4] Clinic for Nuclear Medicine and Competence Center for Thyroid Diseases, Imaging Institute of Southern Switzerland, Ente Ospedaliero Cantonale, Bellinzona, Switzerland.

[5] Clinic for Nuclear Medicine, University Hospital and University of Zurich, Zurich, Switzerland.

[6] Endocrinology, Diabetes and Metabolic Disease Unit, Azienda Ospedaliera Ordine Mauriziano, Torino, Italy.

[7] Head Neck and Thyroid Imaging, Department of Radiology, Guy's and St Thomas' Hospitals NHS Foundation Trust, London, United Kingdom.

[8] Faculty of Biomedical Sciences, Università della Svizzera Italiana (USI), Lugano, Switzerland.

**Corresponding author**

Pierpaolo Trimboli, MD

Clinic for Nuclear Medicine and Competence Center for Thyroid Diseases, Imaging Institute of Southern Switzerland, Ente Ospedaliero Cantonale.

Via Ospedale 12, 6500 Bellinzona, Svizzera

Tel: +41 (0)91 811 85 46

Email: pierpaolo.trimboli@eoc.ch

**Short title**

Performance of EU-TIRADS.

**Key words**

Thyroid nodule; ultrasound; EU-TIRADS; thyroid cancer; meta-analysis.

**Word count:** 3557

**ORCID**

| | |
|---|---|
| Marco Castellana | 0000-0002-1175-8998 |
| Giorgio Grani | 0000-0002-0388-1283 |
| Maija Radzina | 0000-0002-9518-4855 |
| Luca Giovanella | 0000-0003-0230-0974 |
| Maurilio Deandrea | 0000-0001-9217-8815 |
| Cosimo Durante | 0000-0002-1791-5915 |
| Pierpaolo Trimboli | 0000-0002-2125-4937 |

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

**ABSTRACT**

**Objective.** Several thyroid imaging reporting and data systems (TIRADS) have been proposed to stratify the malignancy risk of thyroid nodule by ultrasound. The TIRADS by the European Thyroid Association, namely EU-TIRADS, was the last one to be published.

**Design.** We conducted a meta-analysis to assess the prevalence of malignancy in each EU-TIRADS class and the performance of EU-TIRADS class 5 versus 2, 3 and 4 in detecting malignant lesions.

**Methods.** Four databases were searched until December 2019. Original articles reporting the performance of EU-TIRADS and adopting histology as reference standard were included. The number of malignant nodules in each class and the number of nodules classified as true/false positive/negative were extracted. A random-effects model was used for pooling data.

**Results.** Seven studies were included, evaluating 5,672 thyroid nodules. The prevalence of malignancy in each EU-TIRADS class was 0.5% (95%CI 0.0-1.3), 5.9% (95%CI 2.6-9.2), 21.4% (95%CI 11.1-31.7), and 76.1% (95%CI 63.7-88.5). Sensitivity, specificity, PPV, NPV, LR+, LR- and DOR of EU-TIRADS class 5 were 83.5% (95%CI 74.5-89.8), 84.3% (95%CI 66.2-93.7), 76.1% (95%CI 63.7-88.5), 85.4% (95%CI 79.1-91.8), 4.9 (95%CI 2.9-8.2), 0.2 (95%CI 0.1-0.3), and 24.5 (95%CI 11.7-51.0), respectively. A further improved performance was found after excluding two studies because of limited sample size and low prevalence of malignancy in class 5.

**Conclusions.** A limited number of studies generally conducted using a retrospective design was found. Acknowledging this limitation, the performance of EU-TIRADS in stratifying the risk of thyroid nodules was high. Also, EU-TIRADS class 5 showed moderate evidence of detecting malignant lesions.

## INTRODUCTION

Thyroid nodule is a common entity. The prevalence of palpable lesions is estimated in 5% in women and 1% in men living in iodine-sufficient parts of the world; this increases to 19%–68% in randomly selected individuals assessed by imaging, with higher frequencies reported in women, the elderly and subjects with metabolic syndrome (1,2). Ultrasound (US) is the first-line imaging tool for the assessment of malignancy risk of thyroid nodules. Hypoechogenicity, taller-than-wide shape, irregular margins, microcalcifications, extrathyroidal extension, are recognized as risk features (3,4). However, US reliability is affected by inter- and intra-operator variability of using these features as single parameters (5,6). To solve these weaknesses, several US risk stratification systems (i.e. thyroid imaging reporting and data system, TIRADS) have been developed to stratify the malignancy risk of a nodule and then suggest the need for fine-needle aspiration (FNA) (1,7-11). Among these, the proposal by the European Thyroid Association (ETA), namely EU-TIRADS, was the last one to be published (11). This system categorizes nodules into five classes, from 1 (no nodules) to 5 (high risk). The most remarkable difference with the other systems consists in the fact that the presence of at least one of four features of high suspicion (non-oval/round shape, irregular margins, microcalcifications or markedly hypoechogenicity in a solid nodule) defines a nodule at high risk of cancer (EU-TIRADS class 5) regardless of other US features. Following this approach, the ETA experts have estimated a risk of malignancy close to zero in EU-TIRADS class 2, 2-4% in EU-TIRADS class 3, 6-17% in EU-TIRADS class 4, and ranging from 26 to 87% in EU-TIRADS class 5 (11).

A number of original papers have attempted to evaluate the performance of TIRADSs, including EU-TIRADS (12). In those articles there were two specific outcomes, represented by the risk of malignancy of each class and the reliability in indicating FNA. However, most of those studies were retrospective and their results were heterogeneous, thus limiting the applicability of findings in clinical practice. Importantly, they enrolled nodules previously submitted to FNA even if this indication was not based on TIRADSs; therefore, a significant selection bias was present in those data (12). Finally, the majority of these studies used FNA as reference standard with the introduction of further significant bias; while cytology can detect papillary thyroid carcinoma (PTC), this is not true for follicular cancer (FTC) which is cytologically indistinguishable from its benign counterpart [follicular adenoma (FA)] and usually classified in the indeterminate category (13), or medullary cancer (MTC), which is missed by cytology in up to 50% of cases (14). On the contrary, evaluating the reliability of one TIRADS in a population of patients undergone surgery and using histology as gold standard could allow avoiding bias related to the final diagnosis, even if selection bias is still possible.

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

The present study was undertaken to achieve solid information on the performance of EU-TIRADS. In this order, we planned a systematic review to identify studies reporting histological data of nodules classified according to EU-TIRADS. Also, we performed a meta-analysis of available data to: 1) verify if the predicted/estimated risk of malignancy in each EU-TIRADS class is consistent with real data; 2) evaluate sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), likelihood ratio for positive results (LR+) and for negative results (LR-), diagnostic odds ratio (DOR) of EU-TIRADS class 5 versus the other classes in detecting malignant lesions.

## MATERIALS AND METHODS

The systematic review was registered in PROSPERO (CRD42020150843) and performed in accordance with the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (PRISMA-DTA) (Supplementary Table 1 and 2) (15).

### Search strategy

A six-step search strategy was planned. Firstly, sentinel studies were searched in PubMed. Secondly, keywords and MeSH terms were identified in PubMed. Thirdly, in order to test the strategy, the terms "European" AND "TIRADS" and "EU-TIRADS" were searched in PubMed. Fourthly, PubMed, CENTRAL, Scopus and Web of Science were searched. Fifthly, studies reporting histological data of nodules classified according to EU-TIRADS were selected. Finally, references of included studies were screened for additional papers. The last search was performed on December 5th, 2019. Articles in all languages were accepted and with no restriction to the year they were published. Two investigators (MC, PT) independently searched the papers, screened titles and abstracts of the retrieved articles, reviewed the full-texts and selected articles for their inclusion.

### Data extraction

The following information was extracted independently and in duplicate by two investigators (MC, PT) in a piloted form: 1) general information on the study (author, year of publication, country, study type, number of patients, number of nodules, final diagnosis); 2) number of malignant lesions in each EU-TIRADS class; 3) number of nodules classified as true/false positive/negative. For the purpose of diagnostic performance meta-analysis, EU-TIRADS class was the index test and histology was the reference standard. A benign nodule was

considered as true negative if it was classified as EU-TIRADS class 2, 3 or 4. A benign nodule was considered as false positive if it was classified as EU-TIRADS class 5. A malignant nodule was considered as true positive if it was classified as EU-TIRADS class 5. A malignant nodule was considered as false negative if it was classified as EU-TIRADS class 2, 3 or 4. The main paper and supplementary data were searched; if data was missing, authors were contacted via email. Data were cross-checked and any discrepancy was discussed.

**Study quality assessment**

The risk of bias of included studies was assessed independently by two reviewers (MC, PT) through the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool for the following aspects: patient selection; index test; reference standard; flow and timing. Risk of bias and concerns about applicability were rated as low, high or unclear (16).

**Data analysis**

The characteristics of included studies were summarized. Then, separate analyses were performed according to the following steps. First, a proportion meta-analysis was carried to obtain the pooled rate with 95% confidence interval (95%CI) of malignancy among all histologically proven nodules within a specific EU-TIRADS class. For statistical pooling of data, a random-effects model was used. Second, a diagnostic performance meta-analysis of EU-TIRADS class 5 versus the other classes considered as a whole (i.e. 2, 3 and 4) in selecting malignant nodules was carried out. Summary operating points including sensitivity, specificity, NPV, PPV, LR+, LR-, and DOR, with 95%CI, were estimated. DOR provides a single measure of test performance; it is equal to LR+/LR- and corresponds to the odds of the EU-TIRADS class 5 in a malignant nodule compared with the odds of the EU-TIRADS class 5 in a benign one. The value ranges from zero to infinity, with higher values indicating higher performance. LR+ is the likelihood that EU-TIRADS class 5 would be expected in a malignant nodule (true positive) compared to the likelihood in a benign one (false positive). A LR+ greater than 10 means strong evidence, between 5 and 10 moderate evidence and less than 5 weak evidence. LR- is the likelihood that EU-TIRADS class 2, 3 or 4 would be expected a malignant nodule (false negative) compared to the likelihood in a benign one (true negative). A LR- less than 0.1 means strong evidence, between 0.1 and 0.2 moderate evidence and higher than 0.2 weak evidence. A bivariate random-effects model was used for the pooled analysis of sensitivity and specificity; a random-effects model was used for the pooled analysis of the remaining metrics (17,18). All analyses were performed on a per lesion basis and carried out using

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

OpenMeta[Analyst] (Rockville, Maryland, United States), StatsDirect statistical software (StatsDirect Ltd; Altrincham, UK) and GraphPad Prism version 7 (La Jolla California, United States). Heterogeneity between studies was assessed by using $I^2$, with 50% or higher values regarded as high heterogeneity. For the proportion meta-analysis, the Egger's test was carried out to evaluate the possible presence of significant publication bias. For the diagnostic performance meta-analysis, publication bias was not evaluated, because of uncertainty about the determinants for diagnostic accuracy studies and the inadequacy of tests for detecting funnel plot asymmetry (18). A sensitivity analysis by excluding those studies with specific characteristics was performed. A $p < 0.05$ was regarded as significant.

## RESULTS

A total of 74 papers were found, of which 25 were on PubMed, 27 were on Scopus, 18 were on Web of Science and 4 were on CENTRAL. After removal of 35 duplicates, 39 articles were analyzed for title and abstract; 27 records were excluded (guideline, meta-analysis, TIRADS other than EU-TIRADS, study including only specific groups of nodules [e.g. benign nodules, indeterminate nodules, PET/CT focal thyroid incidentalomas], poster, case report). The remaining 12 papers were retrieved in full-text and 7 articles were finally included in the systematic review (Figure 1) (19-25). No additional study was retrieved after screening the references of these papers.

### Qualitative analysis (systematic review)

The characteristics of the included articles are summarized in Table 1. The papers were published between 2018 and 2019, had sample sizes ranging from 48 to 1,612 thyroid nodules. Participants were adult outpatients who had undergone either thyroid surgery or parathyroid surgery and with US images available. Both nodules on which surgical indication was based (either compressive symptoms or cancer risk) and other nodules in the same patients were included, with a mean number of 1.4 nodules per patient ranging from 1.0 to 2.7 (19,22). Five studies were retrospective, and two prospective cohorts. Two studies were carried out in China, two in Poland, one in Korea, one in Italy and one multicenter study in France, Switzerland and the United Kingdom. All studies assessed EU-TIRADS with histology as the gold standard for both malignant and benign diagnosis. The prevalence of malignancy ranged from 6% to 75% (19,21). Overall, 2,533 malignant and 3,139 benign nodules were included in the present review.

**Quantitative analysis (meta-analysis)**

First, the pooled prevalence of malignancy among all nodules was assessed. It corresponded to 0.5% (95%CI 0.0 to 1.3; $I^2$=0%) in EU-TIRADS class 2, 5.9% (95%CI 2.6 to 9.2; $I^2$=88%) in EU-TIRADS class 3, 21.4% (95%CI 11.1 to 31.7; $I^2$=96%) in EU-TIRADS class 4, and 76.1% (95%CI 63.7 to 88.5; $I^2$=98%) in EU-TIRADS class 5 (Figure 2). There was no evidence of publication bias.

Second, a diagnostic performance meta-analysis of EU-TIRADS class 5 versus the other classes considered as a whole (i.e. 2, 3 and 4) in selecting malignant nodules was carried out. The number of true/false positive/negative in each study is shown in Table 2. The pooled sensitivity was 83.5% (95%CI 74.5 to 89.8), specificity was 84.3% (95%CI 66.2 to 93.7), PPV was 76.1% (95%CI 63.7 to 88.5), and NPV 85.4% (95%CI 79.1 to 91.8). Since these summary operating points are influenced by the prevalence of the disease in the population tested, we estimated the following parameters, which are independent of disease prevalence and thus characteristics of EU-TIRADS. The pooled LR+ was 4.9 (95%CI 2.9 to 8.2), LR- was 0.2 (95%CI 0.1 to 0.3), and DOR was 24.5 (95%CI 11.7 to 51.0). A high heterogeneity was found for all the outcomes (Table 3).

Among the included studies, there were one study with a limited sample size and one study in which the prevalence of malignancy among all nodules in EU-TIRADS class 5 differed significantly from the other studies (20,21). Therefore, we performed a sensitivity analysis by removing these studies. The prevalence of malignancy among all nodules was 1.6% (95%CI 0.0 to 3.8) in EU-TIRADS class 2, 5.5% (95%CI 2.2 to 8.7) in EU-TIRADS class 3, 20.6% (95%CI 8.2 to 33.0) in EU-TIRADS class 4, and 83.3% (95%CI 77.4 to 89.2) in EU-TIRADS class 5 (Supplementary Figure 1 and 2). When the performance of EU-TIRADS class 5 versus the other classes in selecting malignant nodules was assessed, the following results were found: sensitivity was 81.9% (95%CI 71.2 to 89.2), specificity was 90.4% (95%CI 77.0 to 96.4), PPV was 83.3% (95%CI 77.4 to 89.2), NPV was 86.3% (95%CI 78.6 to 94.0), LR+ was 7.2 (95%CI 4.2 to 12.5), LR- was 0.2 (95%CI 0.1 to 0.4), and DOR was 36.9 (95%CI 15.6 to 87.6). In this sensitivity, as in the overall analysis, a high heterogeneity was found for all the outcomes except for the prevalence of malignancy in the EU-TIRADS class 2 (Supplementary Table 3).

**Study quality assessment**

The risk of bias of the included studies is shown in Supplemental Table 4. Overall, we found a low risk of bias: in most studies patients included were consecutive ones and had a histological diagnosis in a specific time period; the classification according to EU-TIRADS was conducted before the final diagnosis or, in

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

retrospective studies, researchers were blinded to the final diagnosis. We rated flow and timing bias as low since thyroid cancer is a chronic condition. The only exception to the statements above included one study in which patient selection risk of bias was rated as unclear, with no information on a consecutive or random enrollment was reported (20). In the same study, the index test applicability concerns item was rated as high, with the prevalence of malignancy among those nodules classified as EU-TIRADS class 5 differing significantly from the other studies. This may be due to differences in technology, execution, or interpretation which affected the estimates of the diagnostic accuracy (20). Finally, five studies excluded nodules depending on size or composition, thus patient selection applicability concerns item was rated as high (20-23,25).

**DISCUSSION**

The aim of this systematic review was to identify the best available evidence on the performance of EU-TIRADS. Particularly, we aimed to assess if the prevalence of malignancy in each EU-TIRADS class was in line with the one estimated by the ETA experts and if EU-TIRADS class 5 was able to select the majority of malignant nodules. To avoid any bias related to the reference standard, we included only histologically proven lesions. To our knowledge, this is the first systematic review and meta-analysis on the topic. We believe this to be a significant contribution to the current understanding, since studies evaluating populations with different prevalence of malignancy could be interpreted together. An extensive database search was performed without time or language restrictions and inclusion criteria were defined prior to the database search. Seven studies were found, evaluating 2,533 malignant and 3,139 benign thyroid nodules.

The prevalence of malignancy was 0.5% in EU-TIRADS class 2, 5.9% in EU-TIRADS class 3, 21.4% in EU-TIRADS class 4, and 76.1% in EU-TIRADS class 5. These findings were very close to the ETA experts estimates, being these close to zero in EU-TIRADS class 2, 2-4% in EU-TIRADS class 3, 6-17% in EU-TIRADS class 4, and ranging from 26 to 87% in EU-TIRADS class 5, as stated (11). Interestingly, no heterogeneity was found in EU-TIRADS class 2. Therefore, EU-TIRADS should be considered as an accurate system to stratify the risk of malignancy of thyroid nodules and the recommendation of not performing FNA in nodules classified in EU-TIRADS class 2 unless compressive symptoms are complained is now supported by a high-level of evidence.

EU-TIRADS, as all other TIRADSs, was conceived to distinguish at US benign nodules that can be managed conservatively from those with suspicious or malignant features requiring further management, usually represented by FNA (11). We have previously reported that all the five most commonly used TIRADS have an

appropriate performance in the selecting malignant thyroid nodules for FNA, with some differences (12). The correlation between US presentation and cytological diagnosis is reported in the literature (26) and all the main systems for thyroid cytology reporting have been found to appropriately stratify the risk of malignancy (27-31). Then, we raised the question whether a high-risk US presentation could be deemed sufficient to submit the patient to surgery, without the need for a FNA confirmation. Particularly, what if all patients with nodules classified as EU-TIRADS class 5 would be submitted to surgery? Also, what if only those patients with nodules classified as EU-TIRADS class 5 would be submitted to surgery? (32). Accordingly, a diagnostic performance meta-analysis to evaluate the ability of EU-TIRADS class 5 versus the classes 2, 3 and 4 considered as a whole was performed. Sensitivity was 83.5%, specificity was 84.3%, PPV was 76.1%, NPV was 85.4%, LR+ was 4.9, LR- was 0.2, and DOR was 24.5. Of note, the performance of EU-TIRADS class 5 was further improved when two studies were excluded. These data provided with moderate evidence that EU-TIRADS class 5 is able to select malignant nodules. All analyses were performed on a nodule basis, then these findings can be applied to a hypothetical population of subjects having a single thyroid nodule. If all patients with an EU-TIRADS class 5 nodule were submitted to surgery, about 76% of patients in the overall analysis and 83% in the sensitivity analysis would have been found to have a malignant nodule. On the other hand, if only those patients with an EU-TIRADS class 5 nodule were submitted to surgery, about 17% of patients with malignancy in the overall analysis and 18% in the sensitivity analysis would have been missed, but the number of patients submitted to surgery would have been reduced by 55% in the overall analysis and 62% in the sensitivity analysis. Although significant, these results should only be interpreted as promising. Indeed, patients included in this meta-analysis were submitted to surgery because of cancer risk or compressive symptoms; in the latter patients, surgery would be still indicated, irrespective from EU-TIRADS class. Also, multinodular disease is a common finding. Therefore, the number of spared surgeries of our estimate is possibly overestimated and any inference possibly biased. However, future TIRADS should take this data into account. On the other hand, it is currently debated whether all malignant nodules should undergo surgery: small, low-risk malignancies may also be managed conservatively (33), and it is also to be taken into account along with malignancy itself.

The following two aspects reduced the consistency of our findings. First, the prevalence of malignancy in EU-TIRADS class 5 in the study conducted by Dobruch-Sobczak et al. differed significantly from the others, as stated (20). Second, a high heterogeneity for all summary operating points above was estimated (20). Concerning the former aspect, according to EU-TIRADS, a nodule should be classified as class 5 if at least one feature amongst non-oval/round shape, irregular margins, microcalcifications or markedly hypoechogenicity

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

(and solid) is found (11). In a multicenter study, there was no difference in the prevalence of malignancy in this class, when data from the three institutions was compared (23). Also, Skowrońska et al. performed a study in the same country of Dobruch-Sobczak et al. and found a prevalence of malignancy in EU-TIRADS class 5 close to the one estimated in the other studies (19). Therefore, the lower prevalence of malignancy reported by Dobruch-Sobczak et al. could be possibly due to US being an operator-dependent imaging modality, rather than characteristics of included patients (20). The same may hold true for the lack of homogeneity for those parameters known to be independent from the prevalence of the disease in the population tested (i.e. LR+, LR-, DOR), as previously reported (12). From another perspective, while the results of the sensitivity analysis can be considered representative of the performance of EU-TIRADS under ideal conditions, findings from the overall analysis may possibly be closer to the real-life data. Anyway, both performances were in line with the predicted risk of malignancy in each EU-TIRADS class estimated by the ETA experts and they were close to one another.

This review has several limitations. The first limitation relates to the design of studies. The majority of studies here included performed a retrospective review and re-classification of nodules which had been submitted to FNA, with possible selection bias. A second aspect leading to a selection bias was represented by the inclusion of patients who had undergone surgery only. It is worth underlining that this was planned to exclude any bias related to the reference standard. Also, it resulted in the inclusion of nodules other than the one on which surgery indication was based. The third limitation was the inter-exam agreement between real-time and retrospective US image interpretation for thyroid nodules. If the appropriate images were not captured during ultrasound examination, this would lead to an unreliable re-assessment of nodules included in retrospective cohort studies (20,22-25,34). Finally, the number of PTC, FTC, MTC and other malignancies in each EU-TIRADS class was generally not reported, the only exception being represented by Skowrońska et al. (19). Therefore, the performance of EU-TIRADS in classifying and detecting each histotype remains to be assessed.

The advantages of adopting TIRADSs in improving the selection of thyroid nodules is recognized and several options were reported in the literature. However, the implications for clinical practice of available studies evaluating the performance of these TIRADS were often limited by the inclusion of nodules with a cytological diagnosis only. EU-TIRADS is a pattern-based practical tool, allowing a rapid assessment in patients with uni- and multinodular goiter. In the present study, only histologically proven nodules were included and EU-TIRADS was found to be effective in stratifying their risk of malignancy. The risk of malignancy in EU-TIRADS class 2 was limited, then no further procedure is needed in these nodules unless symptomatic. On the other hand, a diagnostic and surgical workup is indicated in nodules classified as EU-TIRADS class 3 and above. Particularly,

moderate evidence was found for EU-TIRADS class 5 of selecting malignant lesions. Further prospective studies would be helpful to further support the performance of the EU-TIRADS.

 Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

## DECLARATION OF INTEREST, FUNDING AND ACKNOWLEDGEMENTS

### Declaration of interest

There is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

### Funding

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

### Author contribution statement

PT conceived the meta-analysis. All authors contributed to the development of the selection criteria, the risk of bias assessment strategy and data extraction criteria. MC developed the search strategy. MC and PT performed the database search, acquired the data, and analyzed the data. MC, GG and PT drafted the manuscript. All authors read, provided feedback, and approved the final manuscript.

### Acknowledgements

The authors thank Prof. Dong Gyu Na (Gangneung-si Gangwon-do, Korea), Prof. Xiao-Hong Wu (Nanjing, China), and Dr Katarzyna Dobruch-Sobczak (Warsaw, Poland) for providing the requested data.

### Data availability

The datasets generated during and/or analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request.

### Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

**REFERENCES**


1. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, Pacini F, Randolph GW, Sawka AM, Schlumberger M, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* 2016 **26** 1-133.
2. Guo W, Tan L, Chen W, Fan L, Chen Y, Du C, Zhu M, Wei H, Wang W, Gao M, et al. Relationship between metabolic syndrome and thyroid nodules and thyroid volume in an adult population. *Endocrine* 2019 **65** 357-364.
3. Hegedus L. Clinical practice. The thyroid nodule. *The New England Journal of Medicine* 2004 **351** 1764-1771.
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 & Metabolism* 2014 **99** 1253-1263.
5. Hoang JK, Middleton WD, Farjat AE, Teefey SA, Abinanti N, Boschini FJ, Bronner AJ, Dahiya N, Hertzberg BS, Newman JR, et al. Interobserver Variability of Sonographic Features Used in the American College of Radiology Thyroid Imaging Reporting and Data System. *American Journal of Roentgenology* 2018 **211** 162-167.
6. Grani G, Lamartina L, Cantisani V, Maranghi M, Lucia P & Durante C. Interobserver agreement of various thyroid imaging reporting and data systems. *Endocrine Connections* 2018 **7** 1-7.
7. Perros P, Boelaert K, Colley S, Evans C, Evans RM, Gerrard BG, Gilbert J, Harrison B, Johnson SJ, Giles TE, et al. British Thyroid Association. Guidelines for the management of thyroid cancer. *Clinical Endocrinology* 2014 **81** 1-122.
8. Gharib H, Papini E, Garber JR, Duick DS, Harrell RM, Hegedüs L, Paschke R, Valcavi R & Vitti P. 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. *Endocrine Practice* 2016 **22** 622-639.
9. 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. *Koreanan Journal of Radiology* 2016 **17** 370-395.

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

10. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, Cronan JJ, Beland MD, Desser TS, Frates MC, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *Journal of the American College of Radiology* 2017 **14** 587-595.
11. 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.
12. Castellana M, Castellana C, Treglia G, Giorgino F, Giovanella L, Russ G & Trimboli P. Performance of five ultrasound risk stratification systems in selecting thyroid nodules for FNA. A meta-analysis. *Journal of Clinical Endocrinology & Metabolism* 2019 doi: 10.1210/clinem/dgz170.
13. Grani G, Lamartina L, Durante C, Filetti S & Cooper DS. Follicular thyroid cancer and Hurthle cell carcinoma: challenges in diagnosis, treatment, and clinical management. *Lancet Diabetes & Endocrinology* 2018 **6** 500-514.
14. Trimboli P, Treglia G, Guidobaldi L, Romanelli F, Nigri G, Valabrega S, Sadeghi R, Crescenzi A, Faquin WC, Bongiovanni M, et al. Detection rate of FNA cytology in medullary thyroid carcinoma: a meta-analysis. *Clinical Endocrinology* 2015 **82** 280-285.
15. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, Cohen JF, Deeks JJ, Gatsonis C, Hooft L, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *The Journal of the American Medical Association* 2018 **319** 388-396.
16. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA & Bossuyt PM. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011 **155** 529-536.
17. Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflang M & Scholten R. 2013 Chapter 11: Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C, (eds) Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9. Vol. The Cochrane Collaboration.
18. European Network for Health Technology Assessment. Meta-analysis of Diagnostic Test Accuracy Studies. November 2014. https://www.eunethta.eu/wp-content/uploads/2018/01/Meta-analysis-of-Diagnostic-Test-Accuracy-Studies_Guideline_Final-Nov-2014.pdf.
19. Skowronska A, Milczarek-Banach J, Wiechno W, Chudzinski W, Zach M, Mazurkiewicz M, Miskiewicz P & Bednarczuk T. Accuracy of the European Thyroid Imaging Reporting and Data System (EU-TIRADS) in

the valuation of thyroid nodule malignancy in reference to the post-surgery histological results. *Polish Journal of Radiology* 2018 **83** e579-e586.

20. Dobruch-Sobczak K, Adamczewski Z, Szczepanek-Parulska E, Migda B, Wolinski K, Krauze A, Prostko P, Ruchala M, Lewinski A, Jakubowski W, 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. *Journal of Clinical Medicine* 2019 **8** 1781.
21. Grani G, Lamartina L, Ascoli V, Bosco D, Biffoni M, Giacomelli L, Maranghi M, Falcone R, Ramundo V, Cantisani V, et al. Reducing the Number of Unnecessary Thyroid Biopsies While Improving Diagnostic Accuracy: Toward the "Right" TIRADS. *Journal of Clinical Endocrinology & Metabolism* 2019 104 95-102.
22. Shen Y, Liu M, He J, Wu S, Chen M, Wan Y, Gao L, Cai X, Ding J & Fu X. 2019 Comparison of Different Risk-Stratification Systems for the Diagnosis of Benign and Malignant Thyroid Nodules. Front Oncol 9:378.
23. Trimboli P, Ngu R, Royer B, Giovanella L, Bigorgne C, Simo R, Carroll P & Russ G. A multicentre validation study for the EU-TIRADS using histological diagnosis as a gold standard. *Clinical Endocrinology* 2019 **91** 340-347.
24. Xu T, Wu Y, Wu RX, Zhang YZ, Gu JY, Ye XH, Tang W, Xu SH, Liu C & Wu XH. Validation and comparison of three newly-released Thyroid Imaging Reporting and Data Systems for cancer risk determination. *Endocrine* 2019 **64** 299-307.
25. Yoon SJ, Na DG, Gwon HY, Paik W, Kim WJ, Song JS & Shim MS. Similarities and Differences Between Thyroid Imaging Reporting and Data Systems. *American Journal of Roentgenology* 2019 **213** W76-W84.
26. Lee MJ, Hong SW, Chung WY, Kwak JY, Kim MJ & Kim EK. Cytological results of ultrasound-guided fine-needle aspiration cytology for thyroid nodules: emphasis on correlation with sonographic findings. *Yonsei Medical Journal* 2011 **52** 838-844.
27. 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.
28. Trimboli P, Deandrea M, Mormile A, Ceriani L, Garino F, Limone PP & Giovanella L. American Thyroid Association ultrasound system for the initial assessment of thyroid nodules: Use in stratifying the risk of malignancy of indeterminate lesions. *Head and Neck* 2018 **40** 722-727.

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

29. Trimboli P, Crescenzi A, Castellana M, Giorgino F, Giovanella L & Bongiovanni M. Italian consensus for the classification and reporting of thyroid cytology: the risk of malignancy between indeterminate lesions at low or high risk. A systematic review and meta-analysis. *Endocrine* 2019 **63** 430-438.

30. Trimboli P, Fulciniti F, Paone G, Barizzi J, Piccardo A, Merlo E, Mazzucchelli L & Giovanella L. Risk of Malignancy (ROM) of Thyroid FNA Diagnosed as Suspicious for Malignancy or Malignant: an Institutional Experience with Systematic Review and Meta-Analysis of Literature. *Endocrine Pathology* 2020 doi: 10.1007/s12022-019-09602-4.

31. Poller DN, Bongiovanni M & Trimboli P. Risk of malignancy in the various categories of the UK Royal College of Pathologists Thy terminology for thyroid FNA cytology: A systematic review and meta-analysis. *Cancer Cytopathology* 2020 **128** 36-42.

32. Trimboli P & Durante C. Ultrasound risk stratification systems for thyroid nodule: between lights and shadows, we are moving towards a new era. *Endocrine* 2020 doi: 10.1007/s12020-020-02196-6

33. Ramundo V, Sponziello M, Falcone R, Verrienti A, Filetti S, Durante C & Grani G. Low-risk papillary thyroid microcarcinoma: optimal management toward a more conservative approach. *Journal of Surgical Oncology* 2020 *in press*

**34.** Bae JM, Hahn SY, Shin JH & Ko EY. Inter-exam agreement and diagnostic performance of the Korean thyroid imaging reporting and data system for thyroid nodule assessment: Real-time versus static ultrasonography. *European Journal of Radiology* 2018 **98** 14-19.

**Table 1: Characteristic of included studies.**

| First Author, year | Country | Study design | Selection criteria of included study | Thyroid nodules (n) | Malignant nodules (n, %) | Maximum diameter (mm; mean, SD) |
|---|---|---|---|---|---|---|
| Skowrońska, 2018 (19) | Poland | PCS | Thyroid or parathyroid surgery | 143 | 8 (6) | 16.1 ± 17.1 |
| Dobruch-Sobczak, 2019 (20) | Poland | RCS | Thyroid surgery following Bethesda IV-VI FNA or nodular goiter with clinical symptoms. Patients with symptomatic purely cystic lesions were excluded | 842 | 229 (27) | 19.3 ± 12.8 |
| Grani, 2019 (21) | Italy | PCS | Subgroup of patients undergoing thyroid surgery. Patients with nodules < 10 mm were excluded | 48 | 36 (75) | 21.1 ± 11.3 |
| Shen, 2019 (22) | China | RCS | Thyroid surgery following FNA or the finding of highly suggestive features on US or nodular goiter with clinical symptoms. Patients with nodules < 5 mm were excluded | 1,612 | 773 (48) | 16.7 ± 11.7 |
| Trimboli, 2019 (23) | France, Switzerland and the United Kingdom | RCS | Thyroid surgery for all causes. Patients with nodules < 5 mm were excluded. | 1,058 | 257 (24) | 17.9 ± 12.9 |
| Xu, 2019 (24) | China | RCS | Thyroid surgery | 1,510 | 1,005 (66) | 16.5 ± 12.5 |
| Yoon, 2019 (25) | Korea | RCS | Subgroup of patients undergoing thyroid surgery. Patients with nodules < 10 mm were excluded | 459 | 225 (49) | 22.2 ± 9.9 |

Legend - FNA, fine-needle aspiration cytology; PCS, prospective cohort study; RCS retrospective cohort study; US, ultrasound. In Yoon et al., 2019 one benign and five malignant nodules were non-classifiable according to EU-TIRADS; they were not included in the following analyses

 Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

**Table 2: Classification of thyroid nodules for the purpose of diagnostic performance meta-analysis.**

| First Author, year | True positive | False negative | True negative | False positive |
|---|---|---|---|---|
| Skowrońska, 2018 (19) | 6 | 2 | 133 | 2 |
| Dobruch-Sobczak, 2019 (20) | 214 | 15 | 335 | 278 |
| Grani, 2019 (21) | 26 | 10 | 6 | 6 |
| Shen, 2019 (22) | 721 | 52 | 679 | 160 |
| Trimboli, 2019 (23) | 192 | 65 | 774 | 27 |
| Xu, 2019 (24) | 836 | 169 | 410 | 95 |
| Yoon, 2019 (25) | 164 | 56 | 174 | 59 |

**Table 3: Summary estimates of the diagnostic performance of EU-TIRADS class 5 versus the classes 2, 3, and 4 considered as a whole in selecting malignant nodules: results of the overall analysis based on the seven included studies.**

| Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) | LR+ (95% CI) | LR- (95% CI) | DOR (95% CI) |
|---|---|---|---|---|---|---|
| 83.5 (74.5-89.8) | 84.3 (66.2-93.7) | 76.1 (63.7-88.5) | 85.4 (79.1-91.8) | 4.9 (2.9-8.2) | 0.2 (0.1-0.3) | 24.5 (11.7-51.0) |

Legend: DOR, diagnostic odds ratio; LR+, likelihood ratio for positive results; LR-, likelihood ratio for negative results; NPV, negative predictive value; PPV, positive predictive value.

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

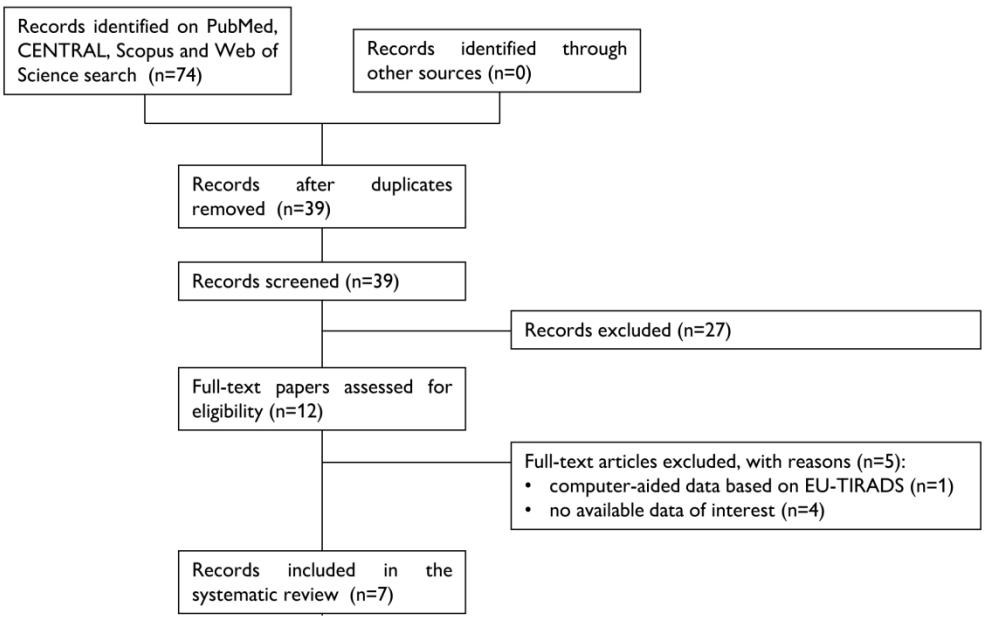



Figure 1: Flow chart of the systematic review.

191x119mm (300 x 300 DPI)

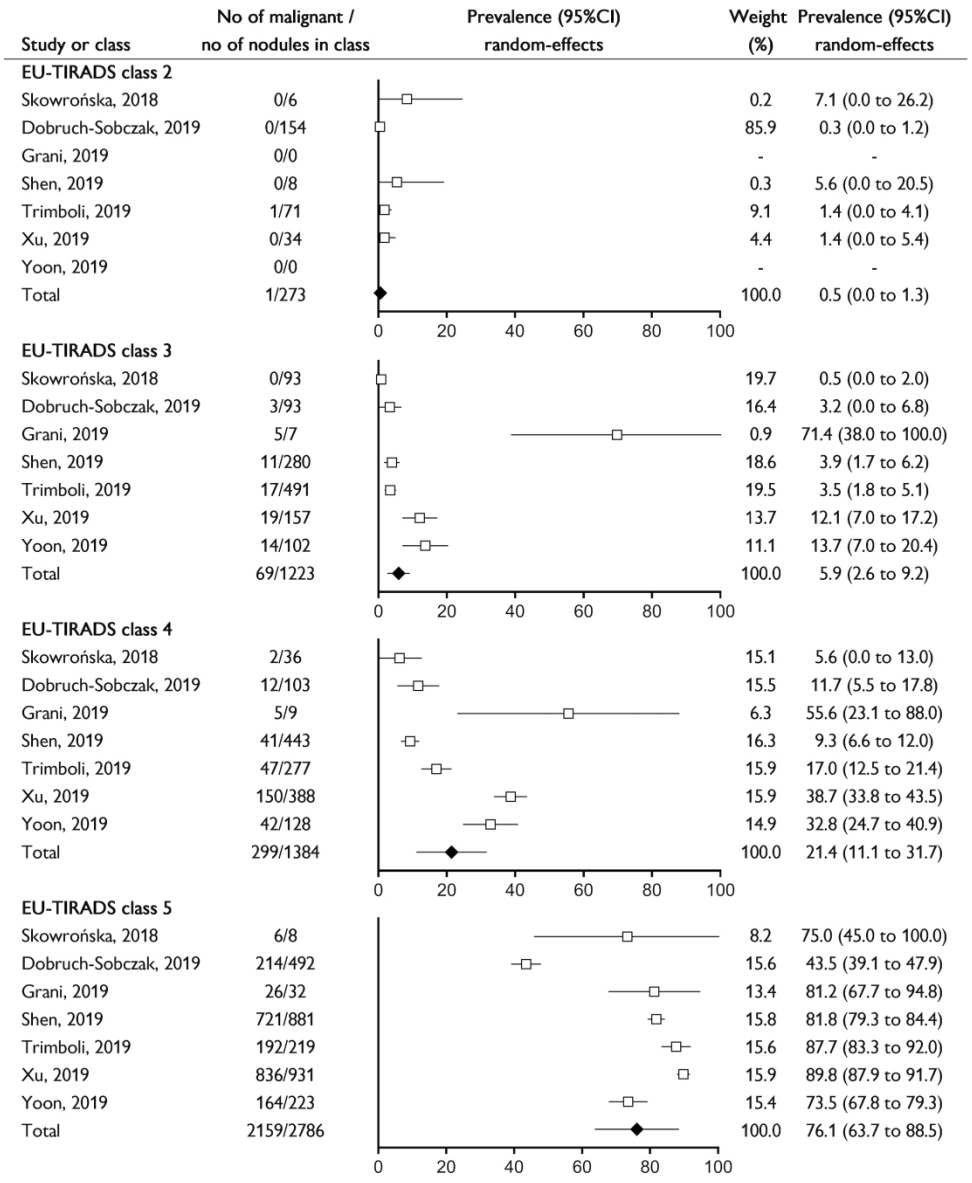



Figure 2. Prevalence of malignancy among all nodules in each EU-TIRADS class in the overall analysis.

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

**Supplementary Table 1: PRISMA-DTA for Abstract Checklist**

| Section/topic | # | PRISMA-DTA for Abstracts Checklist item | Reported on page # |
|---|---|---|---|
| **TITLE and PURPOSE** | | | |
| Title | 1 | Identify the report as a systematic review (+/- meta-analysis) of diagnostic test accuracy (DTA) studies. | 1 |
| Objectives | 2 | Indicate the research question, including components such as participants, index test, and target conditions. | 4 |
| **METHODS** | | | |
| Eligibility criteria | 3 | Include study characteristics used as criteria for eligibility. | 4 |
| Information sources | 4 | List the key databases searched and the search dates. | 4 |
| Risk of bias & applicability | 5 | Indicate the methods of assessing risk of bias and applicability. | 8 |
| Synthesis of results | A1 | Indicate the methods for the data synthesis. | 4 |
| **RESULTS** | | | |
| Included studies | 6 | Indicate the number and type of included studies and the participants and relevant characteristics of the studies (including the reference standard). | 4 |
| Synthesis of results | 7 | Include the results for the analysis of diagnostic accuracy, preferably indicating the number of studies and participants. Describe test accuracy including variability; if meta-analysis was done, include summary results and confidence intervals. | 4 |
| **DISCUSSION** | | | |
| Strengths and limitations | 9 | Provide a brief summary of the strengths and limitations of the evidence | 4 |
| Interpretation | 10 | Provide a general interpretation of the results and the important implications. | 4 |
| **OTHER** | | | |
| Funding | 11 | Indicate the primary source of funding for the review. | 16 |
| Registration | 12 | Provide the registration number and the registry name | 4 |

*Adapted From:* McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, The PRISMA-DTA Group (2018). Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. 2018 Jan 23;319(4):388-396. doi: 10.1001/jama.2017.19163.
For more information, visit: **www.prisma-statement.org**.

**Supplementary Table 2: PRISMA-DTA Checklist**

| Section/topic | # | PRISMA-DTA Checklist Item | Reported on page # |
|---|---|---|---|
| **TITLE / ABSTRACT** | | | |
| Title | 1 | Identify the report as a systematic review (+/- meta-analysis) of diagnostic test accuracy (DTA) studies. | 1 |
| Abstract | 2 | Abstract: See PRISMA-DTA for abstracts. | 4 |
| **INTRODUCTION** | | | |
| Rationale | 3 | Describe the rationale for the review in the context of what is already known. | 5-6 |
| Clinical role of index test | D1 | State the scientific and clinical background, including the intended use and clinical role of the index test, and if applicable, the rationale for minimally acceptable test accuracy (or minimum difference in accuracy for comparative design). | 5-6 |
| Objectives | 4 | Provide an explicit statement of question(s) being addressed in terms of participants, index test(s), and target condition(s). | 6 |
| **METHODS** | | | |
| Protocol and registration | 5 | Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number. | 6 |
| Eligibility criteria | 6 | Specify study characteristics (participants, setting, index test(s), reference standard(s), target condition(s), and study design) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale. | 7 |
| Information sources | 7 | Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched. | 7 |
| Search | 8 | Present full search strategies for all electronic databases and other sources searched, including any limits used, such that they could be repeated. | 7 |
| Study selection | 9 | State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis). | 7 |
| Data collection process | 10 | Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators. | 7 |
| Definitions for data extraction | 11 | Provide definitions used in data extraction and classifications of target condition(s), index test(s), reference standard(s) and other characteristics (e.g. study design, clinical setting). | 7 |
| Risk of bias and applicability | 12 | Describe methods used for assessing risk of bias in individual studies and concerns regarding the applicability to the review question. | 8 |
| Diagnostic accuracy measures | 13 | State the principal diagnostic accuracy measure(s) reported (e.g. sensitivity, specificity) and state the unit of assessment (e.g. per-patient, per-lesion). | 8 |
| Synthesis of results | 14 | Describe methods of handling data, combining results of studies and describing variability between studies. This could include, but is not limited to: a) handling of multiple definitions of target condition. b) handling of multiple thresholds of test positivity, c) handling multiple index test readers, d) handling of indeterminate test results, e) grouping and comparing tests, f) handling of different reference standards | 8-9 |

Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

| | | | |
|---|---|---|---|
| Meta-analysis | D2 | Report the statistical methods used for meta-analyses, if performed. | 8-9 |
| Additional analyses | 16 | Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified. | 9 |
| **RESULTS** | | | |
| Study selection | 17 | Provide numbers of studies screened, assessed for eligibility, included in the review (and included in meta-analysis, if applicable) with reasons for exclusions at each stage, ideally with a flow diagram. | 9 |
| Study characteristics | 18 | For each included study provide citations and present key characteristics including: a) participant characteristics (presentation, prior testing), b) clinical setting, c) study design, d) target condition definition, e) index test, f) reference standard, g) sample size, h) funding sources | 9-10, Table 1 |
| Risk of bias and applicability | 19 | Present evaluation of risk of bias and concerns regarding applicability for each study. | 11-12 |
| Results of individual studies | 20 | For each analysis in each study (e.g. unique combination of index test, reference standard, and positivity threshold) report 2x2 data (TP, FP, FN, TN) with estimates of diagnostic accuracy and confidence intervals, ideally with a forest or receiver operator characteristic (ROC) plot. | 10-11, Table 2 |
| Synthesis of results | 21 | Describe test accuracy, including variability; if meta-analysis was done, include results and confidence intervals. | 10-11, Figure 2, Table 3 |
| Additional analysis | 23 | Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression; analysis of index test: failure rates, proportion of inconclusive results, adverse events). | 11 |
| **DISCUSSION** | | | |
| Summary of evidence | 24 | Summarize the main findings including the strength of evidence. | 12 |
| Limitations | 25 | Discuss limitations from included studies (e.g. risk of bias and concerns regarding applicability) and from the review process (e.g. incomplete retrieval of identified research). | 15 |
| Conclusions | 26 | Provide a general interpretation of the results in the context of other evidence. Discuss implications for future research and clinical practice (e.g. the intended use and clinical role of the index test). | 15 |
| **FUNDING** | | | |
| Funding | 27 | For the systematic review, describe the sources of funding and other support and the role of the funders. | 16 |

*Adapted From:* McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, The PRISMA-DTA Group (2018). Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA. 2018 Jan 23;319(4):388-396. doi: 10.1001/jama.2017.19163.

For more information, visit: **www.prisma-statement.org**.

**Supplementary Table 3: Summary estimates of the diagnostic performance of EU-TIRADS class 5 versus 2, 3 and 4 in selecting malignant nodules: results of the sensitivity analysis.**

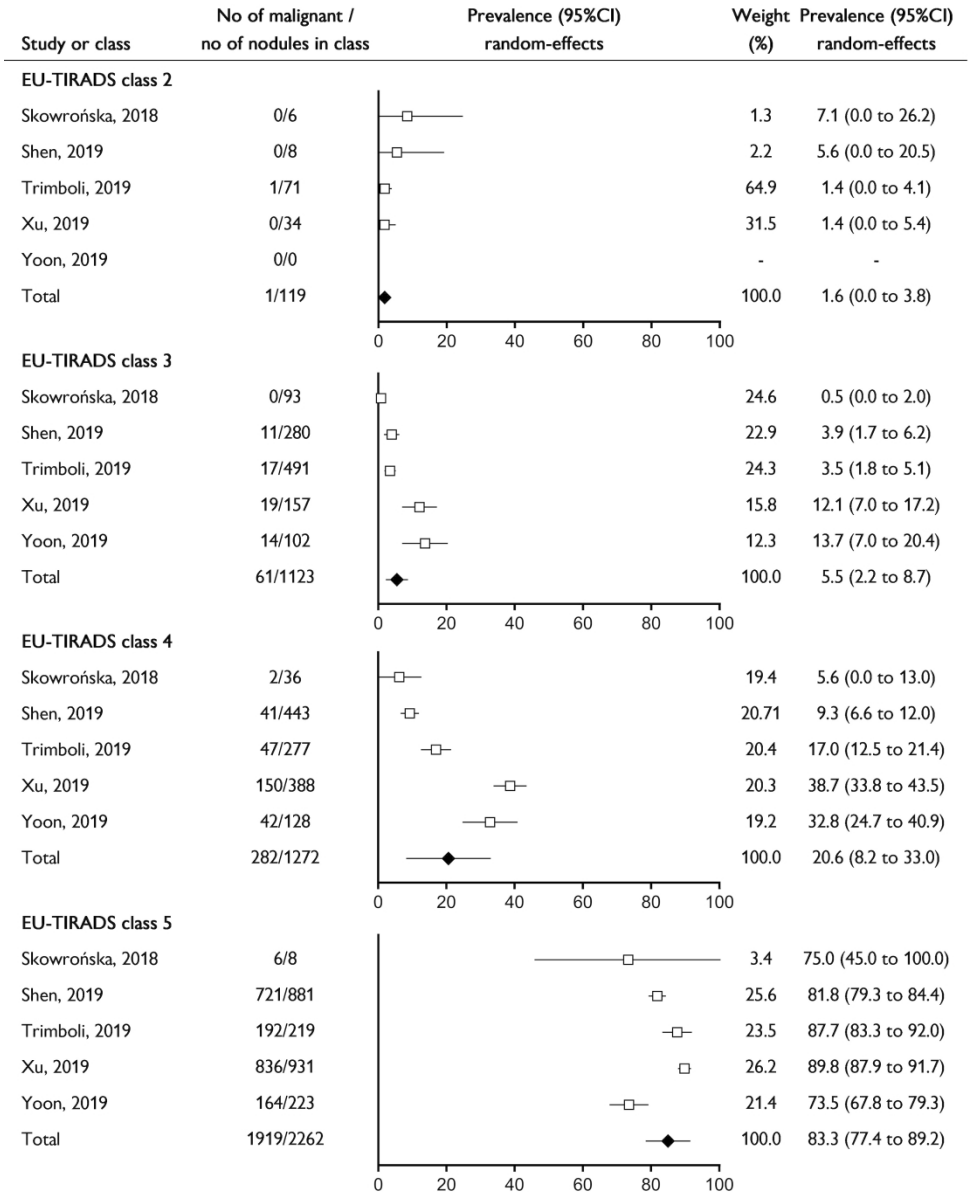
| Sensitivity (95% CI) | Specificity (95% CI) | PPV (95% CI) | NPV (95% CI) | LR+ (95% CI) | LR- (95% CI) | DOR (95% CI) |
|---|---|---|---|---|---|---|
| 81.9 (71.2-89.2) | 90.4 (77.0-96.4) | 83.3 (77.4-89.2) | 86.3 (78.6-94.0) | 7.2 (4.2-12.5) | 0.2 (0.1-0.4) | 36.9 (15.6-87.6) |

This sensitivity analysis was performed after excluding two studies, as reported in the manuscript (Dobruch-Sobczak, 2019; Grani, 2019) (20,21). Five studies were included, corresponding to 4,776 thyroid nodules. DOR, diagnostic odds ratio; LR+, likelihood ratio for positive results; LR-, likelihood ratio for negative results; NPV, negative predictive value; PPV, positive predictive value.

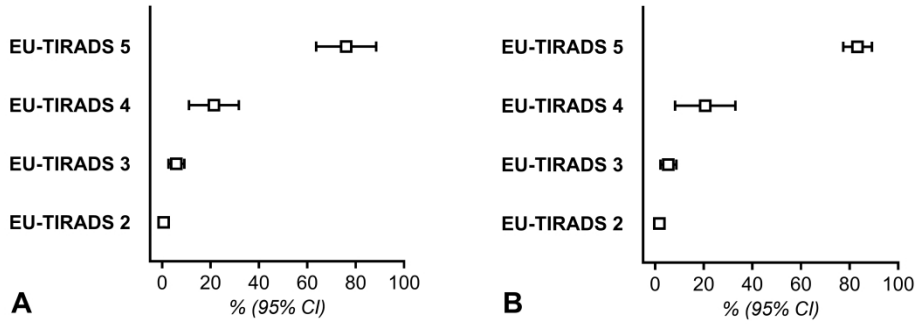
 Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020

**Supplementary Table 4: Risk of bias and applicability concerns summary: review authors' judgements about each domain for each included study**

| | Risk of Bias | | | | Applicability Concerns | | |
|---|---|---|---|---|---|---|---|
| | Patient Selection | Index Test | Reference Standard | Flow and Timing | Patient Selection | Index Test | Reference Standard |
| Skowrońska, 2018 (19) | Low | Low | Low | Low | Low | Low | Low |
| Dobruch-Sobczak, 2019 (20) | Unclear | Low | Low | Low | High | High | Low |
| Grani, 2019 (21) | Low | Low | Low | Low | High | Low | Low |
| Shen, 2019 (22) | Low | Low | Low | Low | High | Low | Low |
| Trimboli, 2019 (23) | Low | Low | Low | Low | High | Low | Low |
| Xu, 2019 (24) | Low | Low | Low | Low | Low | Low | Low |
| Yoon, 2019 (25) | Low | Low | Low | Low | High | Low | Low |



Accepted Manuscript published as EJE-20-0204.R3. Accepted for publication: 16-Jun-2020




196x74mm (600 x 600 DPI)