HEAD AND NECK

Comparison of the diagnostic performances of US-guided fine needle aspiration cytology and thyroglobulin measurement for lymph node metastases in patients with differentiated thyroid carcinoma: a meta-analysis

Rong-Bin Liu [1,2] · Da-Lei Zhou [3] · Bo-Heng Xu [3] · Xin-Hua Yang [3] · Qing Liu [3] · Xiao Zhang [3] · Tao Tang [3] · Zu-Lu Ye [3] · Yue Li [3]

Abstract

Objectives Ultrasound (US)-guided fine needle aspiration cytology (FNAC) and thyroglobulin measurement (FNA-Tg) are two common methods for confirming lymph node metastases (LNM) in patients with differentiated thyroid carcinoma (DTC). This study aimed at comparing the diagnostic performance of FNAC, FNA-Tg alone, and in combination by means of a meta-analysis.
Methods Eligible articles were selected according to predefined criteria, and their quality was evaluated as per the QUADAS-2 checklist. We calculated pooled sensitivity (Se), specificity (Sp), positive/negative likelihood ratio, and diagnostic odds ratio (DOR), and plotted the summary receiver operating characteristic (SROC) curve using the Meta-DiSc1.4 software.
Results Twenty-one studies pooling 1662 malignant and 1279 benign LNs from 2712 patients with DTC were included. The results showed that FNAC was more specific (pooled Sp, 0.98) while FNA-Tg was more sensitive (pooled Se, 0.94). FNAC and FNAC+FNA-Tg performed better postoperatively than FNA-Tg, while FNA-Tg performed better preoperatively. The combination of FNAC and FNA-Tg could achieve a better diagnostic performance than each alone (DOR 446.00, area under the curve [AUC] 0.9862), no matter preoperatively (DOR 378.14, AUC 0.9879) or postoperatively (DOR 788.72, AUC 0.9930). Besides, the combination of FNAC and FNA-Tg/serum-Tg ratio obtained a higher Sp (0.98) than the combination of FNAC and FNA-Tg.
Conclusion The addition of FNA-Tg, especially the FNA-Tg/serum-Tg ratio, to FNAC could increase the diagnostic performance of LNM in both preoperative and postoperative patients with DTC. Since one test or test combinations could perform differently according to the clinical situation, the best-fitting test should be chosen accordingly.

Key Points
• *FNAC is more specific than FNA-Tg while FNA-Tg is more sensitive than FNAC.*
• *The combination of FNAC and FNA-Tg could achieve a better diagnostic performance than either alone, no matter preoperatively or postoperatively.*
• *The combination of FNAC and FNA-Tg/serum-Tg ratio could reach a higher Sp than the combination of FNAC and FNA-Tg.*

Keywords Biopsy, Fine needle · Thyroglobulin · Lymph nodes · Differentiated thyroid carcinoma · Meta-analysis

---

Rong-Bin Liu and Da-Lei Zhou contributed equally to this work.

✉ Zu-Lu Ye
   yezl@sysucc.org.cn

✉ Yue Li
   liyue2@sysucc.org.cn

1   Department of Ultrasound, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510060, People's Republic of China

2   Guangdong Provincial Key Laboratory of Malignant Tumor Gene Regulation and Target Therapy of Guangdong Higher Education Institutes, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, People's Republic of China

3   Department of Molecular Diagnostics, Sun Yat-Sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, 651 Dongfeng Road East, Guangzhou 510060, Guangdong, People's Republic of China

Abbreviations

AUC	Area under the curve
CI	Confidence intervals
DOR	Diagnostic odds ratio
DTC	Differentiated thyroid carcinoma
ECLIA	Electrochemiluminescence immunoassay
FN	False negative
FNAC	Fine needle aspiration cytology
FNA-Tg	Fine needle aspiration-thyroglobulin
FP	False positive
ICMA	Immunochemiluminometric assay
IFMA	Immunofluorometric assay
IMA	Immunometric assay
IRMA	Immunoradiometric assay
LNM	Lymph node metastases
NLR	Negative likelihood ratio
PLR	Positive likelihood ratio
PTC	Papillary thyroid cancer
QUADAS-2	Quality Assessment of Diagnostic Accuracy Studies-2
Se	Sensitivity
Sp	Specificity
SROC	Summary receiver operating characteristic
TN	True negative
TP	True positivity

Introduction

Although disease etiology is still unclear [1], thyroid carcinoma, accounting for 95% of endocrine malignancy, is the sixth most common cancer in women in the USA, and an estimated over 52,000 new cases occurred in men and women in 2019 in this country [2]. Differentiated thyroid cancer (DTC) is the most frequent subtype of thyroid cancer, including papillary thyroid carcinoma (PTC) and follicular thyroid cancer. Like other primary malignancies of the head and neck, PTC presents a consistent pattern of metastases to regional lymph nodes (LNs), ranging from 20 to 50% of cases [3]. Although most patients with DTC have a favorable long-term survival, 5–20% of patients will develop recurrence in LNs during postoperative follow-up, and quite a few patients even present LN metastasis (LNM) at initial diagnosis [4].

A LN burden (ratio of positive LNs to total removed LNs) > 17% in the lateral neck is predictive for recurrence in patients of all ages, and patients with ≥ 2 metastatic LNs might benefit from total thyroidectomy and radioactive iodine therapy [5]. The 2015 guidelines of the American Thyroid Association for the diagnosis and management of DTC recommend that the central and lateral cervical LNs should be examined by routine US before surgery. LNs that are suspicious for DTC metastases on US should undergo US-guided fine needle aspiration cytology (FNAC) [6]. However, several factors may decrease the diagnostic efficacy of FNAC, like small LN size, lymphocyte infiltration, necrosis, or lack of epithelial component in cyst aspirates [7]. In that case, measurement of thyroglobulin (Tg) levels in FNA material has provided additional clues.

The diagnostic accuracy of FNA-Tg variates, due to relatively small sample sizes, different threshold value choices, or distinct clinical settings (before or after thyroidectomy) [8, 9]. In addition, those studies used different detection methods for FNA-Tg measurement, such as immunoradiometric assay (IRMA), immunochemiluminometric assay (ICMA), electrochemiluminescence immunoassay (ECLIA), immunometric assay (IMA), and immunofluorometric assay (IFMA), which might affect the accuracy of analysis since different assays have different analytical sensitivity (Se), functional Se, and inter-assay variability [10, 11]. Therefore, the diagnostic performances of FNA-Tg and its combination with FNAC in identifying LNM need determination under different conditions. This prompted assessment of diagnostic test accuracy by a meta-analysis.

In the era of evidence-based medicine, decision-makers need high-quality data to support decision to use a diagnostic test in a specific clinical situation and to choose the relevant one. Meta-analysis of diagnostic test accuracy studies is a useful method to increase the level of validity by combining data from multiple studies. Statistics dedicated to meta-analyses of diagnostic test accuracy provide either summary points of diagnostic accuracy (e.g., summary (Se), specificity (Sp), positive and negative likelihood ratios, and diagnostic odds ratio (DOR)) or summary lines (i.e., summary receiver operating characteristic [SROC] curves). The DOR takes advantage of accuracy as a single indicator, which is closely linked to both Se and Sp and expresses the strength of the association between test result and disease [12]. The current study had two aims: (1) to compare the diagnostic efficacy of FNAC, FNA-Tg, and their combination at preoperative and postoperative stages; (2) to explore the appropriate cut-off of FNA-Tg in diagnosing LNM of DTC.

Methods

Search strategy

Based on the recommendation in the *Cochrane handbook for systematic reviews of diagnostic test accuracy* (handbook. cochrane.org) [13], two independent investigators predefined the retrieval strategy and conducted a systematic search of bibliographic databases (PubMed, Web of Science, the Cochrane Library, Google Scholar, and CNKI) using different combinations of the search terms including "FNAC", "FNA-Tg/Tg", "lymph node/lymph node metastases", and "thyroid cancer/thyroid carcinoma", for articles

published up to October 2019. Furthermore, references cited in each identified literature were further searched manually for potential available studies. We deleted duplicated and unrelated articles by reading titles and abstracts, and then further excluded articles not meeting the inclusion criteria by reviewing the full text. We did not set any language restrictions to our search; any differences in the process of article searching and data extracting were resolved through team discussion. We contacted the author for specific raw data if the data provided in the article were not sufficient. To avoid the involvement of duplicated samples, authors of papers that appeared to be relevant were contacted via email if available. If overlapping data by the same first author were found, the article with the largest number of subjects was included.

Inclusion criteria

All the selected studies should meet the following criteria: (1) malignant LNs were confirmed through pathological examination after surgical resection (negative LNs not resected operatively could be confirmed if they stay negative after 2 years' follow-up); (2) detection results of FNAC+FNA-Tg should be presented; (3) data such as true positivity (TP), false positivity (FP), true negative (TN), and false negative (FN) should be provided or could be calculated with sensitivity (Se), specificity (Sp), accuracy, and so on.

Data extraction and quality assessment

Two reviewers independently extracted the following data: the first author's name, year of publication, study design, sample size of cases and controls, pathological type of disease, preoperative or postoperative diagnosis of LN metastases, cutoffs of FNA-Tg, Se, Sp, TP, FP, TN, and FN of FNAC, FNA-Tg, and FNAC+FNA-Tg. The included articles were evaluated item-by-item according to the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) checklist [14]. Risks of bias in regard to four key domains including patient selection, index test, reference standard, and the flow and timing were evaluated, and the first three domains were applied to assess applicability concerns. If a study is judged as "low risk" on all domains, then it should be considered as "low risk of bias" or "low concern regarding applicability." If a study is classified as "high risk" or "unclear" in one or more domains, then it should be judged as "at risk of bias" or having "concerns regarding applicability." Disagreements generated in this process were resolved by consensus.

Statistical analysis

The threshold effect was evaluated by assessing the shape of a ROC plane plot and the Spearman's correlation coefficient between the logit of Se and the logit of 1-Sp; p value for

Spearman's correlation coefficient (p^a) < 0.05 indicated the existence of a threshold effect. The inconsistency index I-square (I^2) and Cochran's Q were calculated to check heterogeneity caused by non-threshold effect factors, and we set I^2 > 50% as substantial heterogeneity. The positive likelihood ratio (PLR, Se/1-Sp) represents the ratio of TP and FP and the negative likelihood ratio (NLR, 1-Se/Sp) equals to the ratio of FN and TN; they act as relatively independent estimators of how much a test result will change the odds of having a disease [15]. The SROC curve with its corresponding area under the curve (AUC) and Q value is a suitable way to evaluate the stability and accuracy of a test. The Q value is the intersection point of the SROC curve with a diagonal from top left to bottom right, which corresponds to the highest value of Se and Sp for the test. p value for the difference between the b value and zero (p^b) < 0.05 indicated that an asymmetrical SROC should be plotted; otherwise, the symmetrical SROC was applied. The DOR was computed using the Moses-Shapiro-Littenberg model for symmetrical or asymmetrical SROC, which reflects the relationship between the result of the diagnostic test and the disease. All the statistical outcomes and result plots were produced using Meta-DiSc1.4 software.
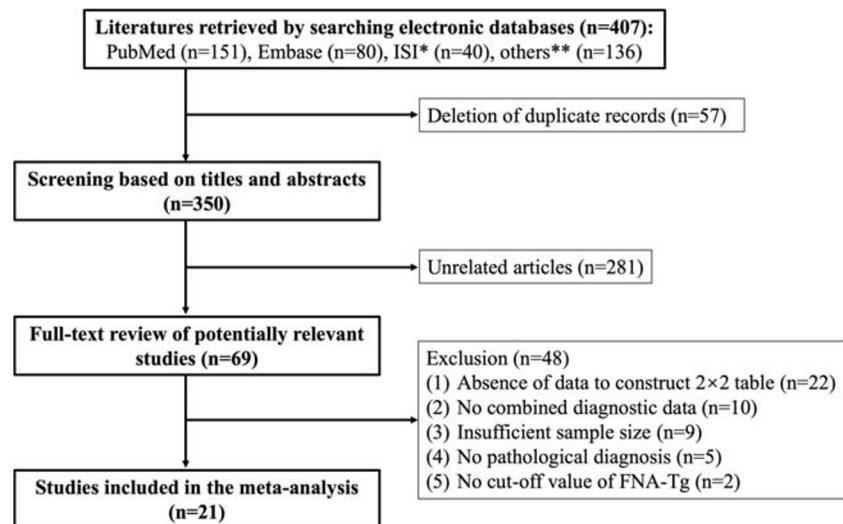
Results

Search results and characteristics of included studies

The flow chart summarizing the study screening process is shown in Fig. 1. The initial search provided a total of 407 records, of the 350 records left after removal of duplicate ones, and 281 unrelated ones were further excluded on the basis of titles and abstracts. Next, full texts of the remaining 69 articles were carefully read and screened according to our predefined inclusion criteria, and 48 articles were discarded due to the absence of data to calculate diagnostic parameters ($n = 22$), no FNAC+FNA-Tg combined diagnostic data ($n = 10$), insufficient sample size ($n = 9$), lack of pathological diagnosis ($n = 5$), and no mentioned FNA-Tg cut-off value ($n = 2$). Finally, 17 English [16–32] and 4 Chinese [33–36] articles including 1662 malignant and 1279 benign LNs from 2712 patients with DTC were involved into our meta-analysis.

Among them, 8 studies researched on LNs from preoperative patients, 7 studies included LNs from postoperative patients, 2 studies involved both preoperative and postoperative patients, and 4 studies did not clarify if the patients had received a thyroidectomy. Different detection methods for FNA-Tg were taken: 8, 5, 4, 1, and 1 studies took use of IRMA, ICMA, ECLIA, IMA, and IFMA, respectively, and the remaining 2 studies did not describe what methods they had used. Only about half of the included studies had reported the analytical Se and functional Se of the detection methods for FNA-Tg. Three studies took FNA-Tg/serum-Tg ratio

Fig. 1 Flow chart of study selection and inclusion in the meta-analysis. Excluded PubMed (single asterisk symbol); other databases searched: Google Scholar, Baidu Scholar, Global Health (excluded PubMed) (double asterisk symbols)

```
┌──────────────────────────────────────────────────────┐
│ Literatures retrieved by searching electronic         │
│ databases (n=407):                                    │
│ PubMed (n=151), Embase (n=80), ISI* (n=40),           │
│ others** (n=136)                                      │
└──────────────────────────────────────────────────────┘
                    │
                    │──────────► ┌────────────────────────────────┐
                    │            │ Deletion of duplicate records  │
                    │            │ (n=57)                         │
                    ▼            └────────────────────────────────┘
┌──────────────────────────────────────────┐
│ Screening based on titles and abstracts   │
│ (n=350)                                   │
└──────────────────────────────────────────┘
                    │
                    │──────────► ┌────────────────────────────────┐
                    │            │ Unrelated articles (n=281)     │
                    ▼            └────────────────────────────────┘
┌──────────────────────────────────────────┐
│ Full-text review of potentially relevant  │
│ studies (n=69)                            │
└──────────────────────────────────────────┘
                    │
                    │──────────► ┌──────────────────────────────────────────────┐
                    │            │ Exclusion (n=48)                             │
                    │            │ (1) Absence of data to construct 2×2 table   │
                    │            │     (n=22)                                   │
                    │            │ (2) No combined diagnostic data (n=10)       │
                    │            │ (3) Insufficient sample size (n=9)           │
                    │            │ (4) No pathological diagnosis (n=5)          │
                    ▼            │ (5) No cut-off value of FNA-Tg (n=2)         │
┌──────────────────────────────────────────┐ └──────────────────────────────────────────────┘
│ Studies included in the meta-analysis     │
│ (n=21)                                    │
└──────────────────────────────────────────┘
```

while the others took FNA-Tg levels as the positive standard for the FNA-Tg test. All the above characteristics of the included studies are summarized in Table 1.

Study quality evaluated by QUADAS-2

Based on the QUADAS-2 tool, the patient selection of all the studies matched corresponding signal questions, so it was of low risk of bias and applicability concerns. While, several studies might introduce high or unclear risk of bias in terms of index test and flow and timing, because they did not interpret the diagnostic results of FNAC in a blind manner or did not pre-establish a diagnostic threshold for FNA-Tg interpretation. In addition, part of the involved studies employed different FNA-Tg cut-offs or set a cut-off value after summarizing all the detection data, which might bring about concerns regarding applicability of the index test (Fig. 2).

Heterogeneity evaluation

As it could be seen from the ROC plane plots of the total group, no typical "shoulder arm" pattern was presented for FNAC, FNA-Tg, or FNAC+FNA-Tg (Fig. 3). All the p^a values were > 0.05, no matter the preoperative subgroup or the postoperative subgroup, suggesting that there was no statistical threshold effect existing in all the included studies (Table 2). The p^b values in preoperative and postoperative subgroups were all > 0.05, so symmetrical SROC curves were plotted, while p^b values in the total group were all < 0.05 (Table 2), so asymmetrical SROC curves were fitted for FNAC, FNA-Tg, and FNAC+FNA-Tg, respectively.

For non-threshold effect, I^2 and Cochran's Q for DOR were evaluated. In the total group, statistical heterogeneities existed in the series of studies of FNAC ($I^2 = 62.4\%$, p for Cochran's $Q = 0.0001$) and FNA-Tg ($I^2 = 63.5\%$, p for Cochran's $Q =$

0.0001) but not in FNAC+FNA-Tg ($I^2 = 26.24\%$, p for Cochran's $Q = 0.1581$) (Table 3). To deal with the heterogeneity and evaluate the robustness of our results, a random effects model was employed to calculate the pooled results, the data was reanalyzed after removing some studies that visibly deviated from the overall, and subgroup analysis was performed to explore the sources of heterogeneity.

Diagnostic performances of FNAC, FNA-Tg, and FNAC+FNA-Tg in the total group

FNAC

The summary Se and Sp with 95% CI were 0.82 (0.80–0.84) and 0.98 (0.97–0.99), respectively, indicating that FNAC is quite specific for LNM but it might produce false negative results to a certain degree. DOR of FNAC in diagnosing LNM was 183.75, and the computed AUC of SROC curve was 0.9353 (Figs. 4, 5, and 6; Table 3). We noticed that two studies (Cunha 2007 and Holmes 2014) had an extremely high Se of FNAC (Fig. 4), and another two studies (Al-Hilli 2016 and Sigstad 2007) presented an obvious low Sp of FNAC (Fig. 5). To evaluate the stability of the results, we took each of the above four studies separately out from the whole, and found the pooled Se and Sp both remained as the same. When the Al-Hilli 2016 study was removed, the DOR and AUC increased to 217.26 and 0.9644; I^2 for DOR decreased a little but was still > 50% (from 62.4 to 58.0%). The DOR, AUC, and I^2 for DOR did not changed much when any of the other three studies was deleted.

FNA-Tg

The combined Se, Sp, and DOR with 95% CI of FNA-Tg in recognizing LNM were 0.93 (0.92–0.95), 0.92 (0.91–0.93),

Table 1 Characteristics of included studies

Author and publication year	Country	Patient number	Patient status	LN number (malignant)	LN number (benign)	Pathological types	FNA-Tg detection			
							Methods	Analytical Se	Functional Se	Cut-offs
Al-Hilli 2016	USA	105	Preop	101	4	PTC	IMA	-	-	1.0 ng/ml
Bournaud 2010	France	98	Unclarified	52	46	DTC	IRMA	0.2 ng/ml	0.7 ng/ml	0.93 ng/ml
Cunha 2007	Portugal	23	Both	20	8	DTC	ICMA	0.2 ng/ml	0.9 ng/ml	0.2 ng/ml
Holmes 2014	USA	19	Postop	11	10	PTC	ICMA	0.2 ng/ml	-	0.2 ng/ml
Jeon 2009	Korea	76	Unclarified	56	20	PTC	-	-	-	36.0 ng/ml
Jo 2015	Korea	273	Postop	261	109	DTC	IRMA	-	0.7 ng/ml	6.0 ng/ml
Jung 2013	Korea	161	Preop	77	100	DTC	IRMA	-	-	1.8 ng/ml
Kim 2009	Korea	168	Both	119	49	DTC	IRMA	0.2 ng/ml	0.7 ng/ml	1.0 ng/ml
Lee 2010	Korea	40	Postop	21	19	PTC	IRMA	0.5 ng/ml	< 0.1 ng/ml	4.1 ng/ml
Li 2016	China	112	Postop	77	51	PTC	ICMA	-	-	FNA-Tg/sTg > 1
Luca 2011	Switzerland	108	Preop	86	40	DTC	IRMA	0.08 ng/ml	0.2 ng/ml	1.0 ng/ml
Moon 2013	Korea	419	Unclarified	190	338	PTC	IRMA	0.08 ng/ml	0.2 ng/ml	1.0 ng/ml
Shi 2015	China	148	Preop	94	54	PTC	-	-	-	10.0 ng/ml
Shin 2015	Korea	201	Postop	106	133	PTC	IRMA	0.2 ng/ml	0.7 ng/ml	19.9 ng/ml
Sigstad 2007	Norway	46	Unclarified	32	14	DTC	IFMA	0.1 ng/ml	0.15 ng/ml	0.1 ng/ml
Tang 2016	USA	97	Preop	21	122	DTC	ICMA	0.1 ng/ml	-	1.0 ng/ml
Wu 2018	China	136	Preop	74	79	PTC	ECLIA	-	-	1.0 ng/ml
Xu 2014	China	148	Postop	94	54	PTC	ECLIA	-	-	FNA-Tg/sTg > 1
Yap 2014	Australia	22	Postop	23	2	DTC	ICMA	0.2 ng/ml	0.9 ng/ml	0.9 ng/ml
Zhang 2014	China	58	Preop	46	22	DTC	ECLIA	0.1 ng/ml	1.0 ng/ml	1.0 ng/ml
Zhou 2017	China	254	Preop	101	5	PTC	ECLIA	-	-	FNA-Tg/sTg > 1

DTC differentiated thyroid cancer, *ECLIA* electrochemiluminescence immunoassay, *FNAC* fine needle aspiration cytology, *FNA-Tg/sTg* fine needle aspiration-thyroglobulin/serum-Tg, *ICMA* immunochemiluminometric assay, *IFMA* immunofluorometric assay, *IMA* immunometric assay, *IRMA* immunoradiometric assay, *PTC* papillary thyroid cancer, *Se* sensitivity
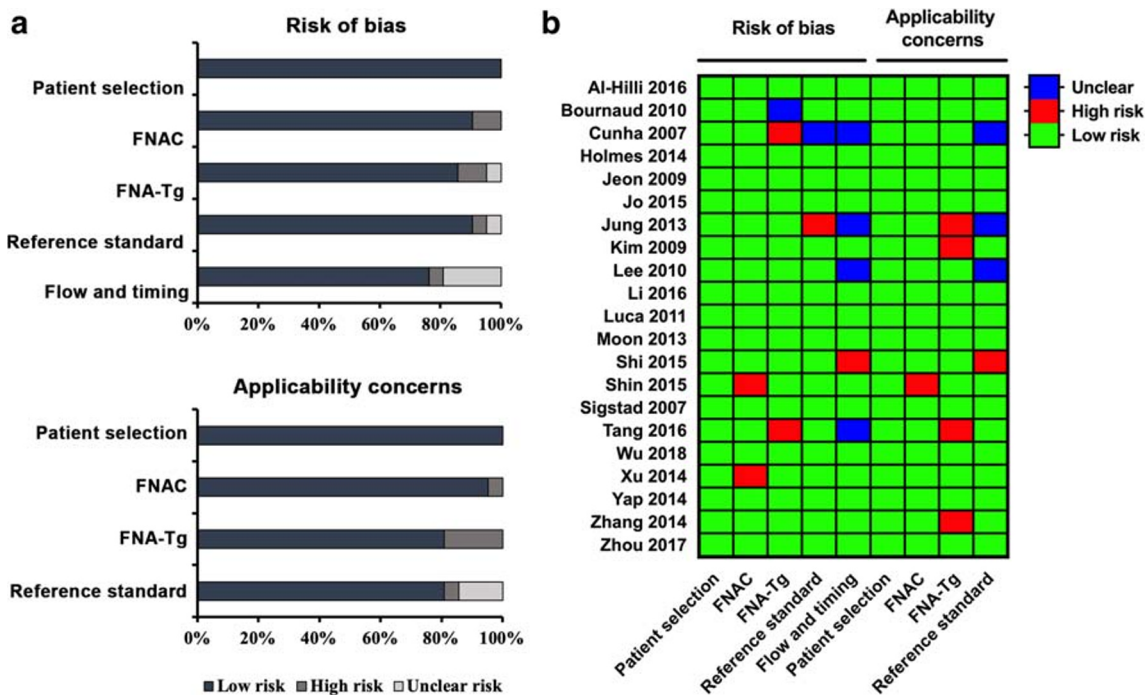

Fig. 2 Quality evaluation of the included studies according to QUADAS-2 guideline. **a** Proportion of studies with low, high, or unclear risk of bias and concerns regarding applicability. **b** Heatmap exhibiting the detailed assessments of involved studies in each key domain of risk of bias and applicability concerns. The included studies were referred as author name and publication year and listed in an alphabetical order of author names

Table 2 Statistical results of Spearman's correlation coefficient

Test	Group	Spearman's correlation coefficient	p^a	p^b
FNAC	Preop	0.139	0.701	0.3678
	Postop	0.576	0.104	0.2385
	Total	0.292	0.199	0.0173
FNA-Tg	Preop	− 0.406	0.244	0.0694
	Postop	− 0.067	0.865	0.0602
	Total	0.049	0.832	0.0276
FNAC+ FNA-Tg	Preop	0.431	0.286	0.2874
	Postop	0.515	0.192	0.4331
	Total	0.391	0.080	0.0263

FNAC fine needle aspiration cytology, *FNA-Tg* fine needle aspiration-thyroglobulin, *Preop* preoperation, *Postop* postoperation

p^a values for analysis of diagnostic threshold effect inter-study, and $p^a < 0.05$ indicates a threshold effect

p^b values for the comparison between the *b* value and zero, and $p^b < 0.05$ indicates that an asymmetrical SROC should be plotted; otherwise, a symmetrical SROC was selected

and 155.17 (86.39–278.74), respectively, and the corresponding AUC of SROC curve was 0.9674 (Figs. 4, 5, and 6; Table 3), indicating that FNA-Tg was more sensitive but less specific than FNAC. As shown in Fig. 5, four studies (Al-Hilli 2016, Kim 2009, Shi 2015, and Yap 2014) reported a much lower Sp of FNA-Tg than the others. We removed the above studies one by one and found the Sp, DOR, and I^2 for DOR only fluctuated minutely, suggesting the meta-analysis results of FNA-Tg were stable.

FNAC+FNA-Tg

The integrated Se and Sp with 95% CI of FNAC+FNA-Tg were 0.97 (0.96–0.98) and 0.94 (0.93–0.95), respectively. The DOR and the AUC of SROC curve of FNAC+FNA-Tg to identify LNM were 446.00 and 0.9862 (Figs. 4, 5, and 6; Table 3), higher than those of either FNAC or FNA-Tg alone. There were four studies (Al-Hilli 2016, Kim 2009, Sigstad 2007, and Zhang 2014) obviously deviating from

the whole in Sp (Fig. 5). When removing the Kim 2009 study out of the overall, we noticed an increase of DOR (from 446.00 to 513.61) and AUC (from 0.9862 to 0.9913) and a decrease of I^2 for DOR (from 23.8 to 0%). The one-by-one deletion of the other three studies did not alter the Sp or the DOR greatly (Sp fluctuated from 0.94 to 0.95, DOR ranged from 459.49 to 501.01).

Diagnostic performances in preoperative and postoperative subgroups

One test could perform differently in different clinical applied situations. As it could be seen from Table 4, the diagnostic efficacy of FNAC in the postoperative subset (DOR 141.82, AUC 0.9560) was higher than that in the preoperative subset (DOR 86.34, AUC 0.9488), while FNA-Tg performed a bit better in the preoperative subgroup (DOR 94.10, AUC 0.9727) than in the postoperative subgroup (DOR 91.37, AUC 0.9510). Noticeably, no matter in prior to or after operation, the combination of FNAC
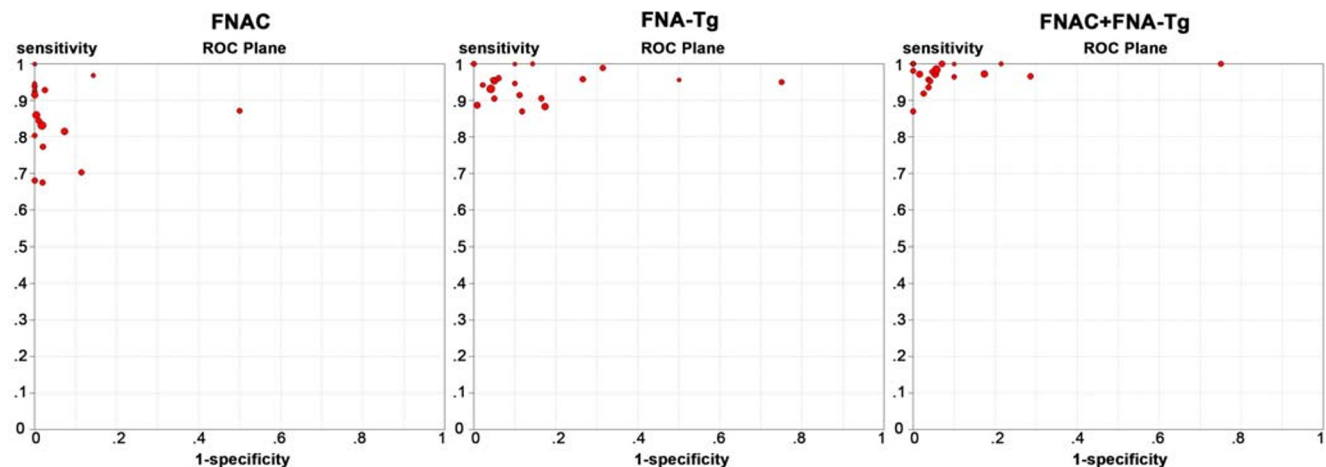

Fig. 3 ROC plane plots for FNAC, FNA-Tg, and FNAC+FNA-Tg

Table 3 The diagnostic performance of FNAC, FNA-Tg, and FNAC+FNA-Tg

Test	Integrated outcomes						AUC	Q
	Se (95% CI) $I^2\%$	Sp (95% CI) $I^2\%$	PLR (95% CI) $I^2\%$	NLR (95% CI) $I^2\%$	DOR (95% CI) $I^2\%$	Cochran's Q p		
FNAC	0.82 (0.80–0.84) 76.4	0.98 (0.97–0.99) 66.0	24.49 (12.30–48.74) 73.3	0.18 (0.14–0.23) 74.0	183.75 (87.00–388.11) 62.4	53.20 0.0001	0.9353*	0.8943
FNA-Tg	0.94 (0.92–0.95) 61.0	0.92 (0.91–0.93) 82.1	9.07 (5.59–14.71) 86.3	0.08 (0.06–0.11) 40.1	155.17 (86.39–278.74) 63.5	54.83 0.0001	0.9674*	0.9322
FNAC+ FNA-Tg	0.97 (0.96–0.98) 57.9	0.94 (0.93–0.95) 76.2	13.65 (7.91–23.55) 84.8	0.04 (0.03–0.06) 41.0	446.00 (267.17–744.53) 23.8	26.24 0.1581	0.9862*	0.9607

*An asymmetrical SROC was plotted. *AUC* area under the curve, *CI* confidence intervals, *DOR* diagnostic odds ratio, *FNAC* fine needle aspiration cytology, *FNA-Tg* fine needle aspiration-thyroglobulin, *NLR* negative likelihood ratio, *PLR* positive likelihood ratio, *Se* sensitivity, *Sp* specificity

and FNA-Tg performed more excellent than either test alone, and the best diagnostic performance was achieved by FNAC+FNA-Tg when applied postoperatively to monitor recurrence (DOR 788.72, AUC 0.9930). The heterogeneity for FNAC+FNA-Tg was much lower than that for FNAC and FNA-Tg alone, no matter preoperatively or postoperatively. The heterogeneity was obviously high in the preoperative subgroup for FNAC and FNA-Tg ($I^2 > 50\%$ and p for Cochran's $Q < 0.05$), indicating non-threshold effect might affect the result.

Subgroup analysis based on different cut-offs of FNA-Tg

The included studies had employed different numerical values of FNA-Tg or the combination of FNA-Tg and serum-Tg as the cut-offs in Tg measurement, which might influence the Se and Sp of each individual study,

thus accumulating obvious heterogeneity and decreasing the diagnostic performance of the pooled analysis. To explore the influence of different positive standards on the diagnostic performance of FNA-Tg and FNAC+FNA-Tg, we divided these studies into four subgroups: 0 < cut-off < 1.0 ng/ml, cut-off = 1.0 ng/ml, cut-off > 1.0 ng/ml, and FNA-Tg/serum-Tg ratio > 1. The number of studies that fell into the above subgroups was 7, 6, 5, and 3, respectively. Interestingly, no matter which cut-off was used, it was for sure that the combination of FNAC and FNA-Tg performed better than either alone (Table 5). The combination of FNAC and FNA-Tg/ serum-Tg ratio achieved a much higher pooled Se, Sp, and DOR (Se, 0.91; Sp, 0.98; and DOR, 448.05) than the combination of FNAC and FNA-Tg. The best Se (0.99) and AUC (0.9949) of FNAC+FNA-Tg was
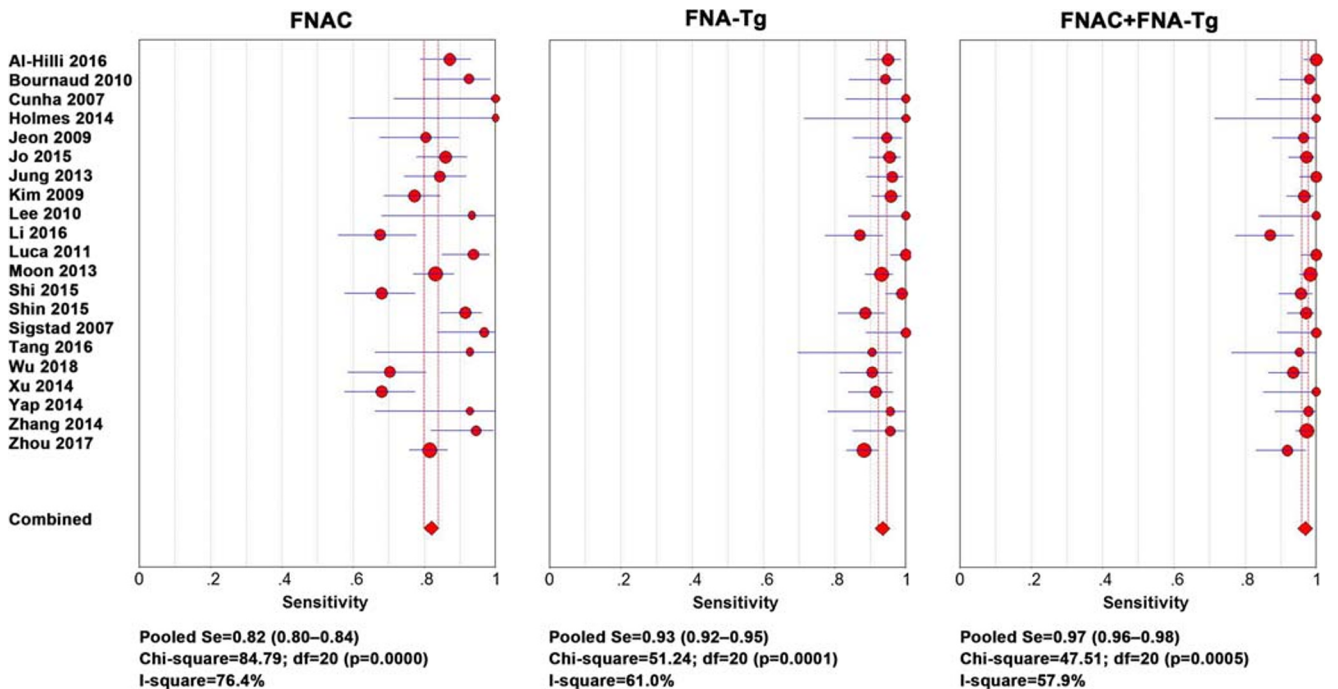

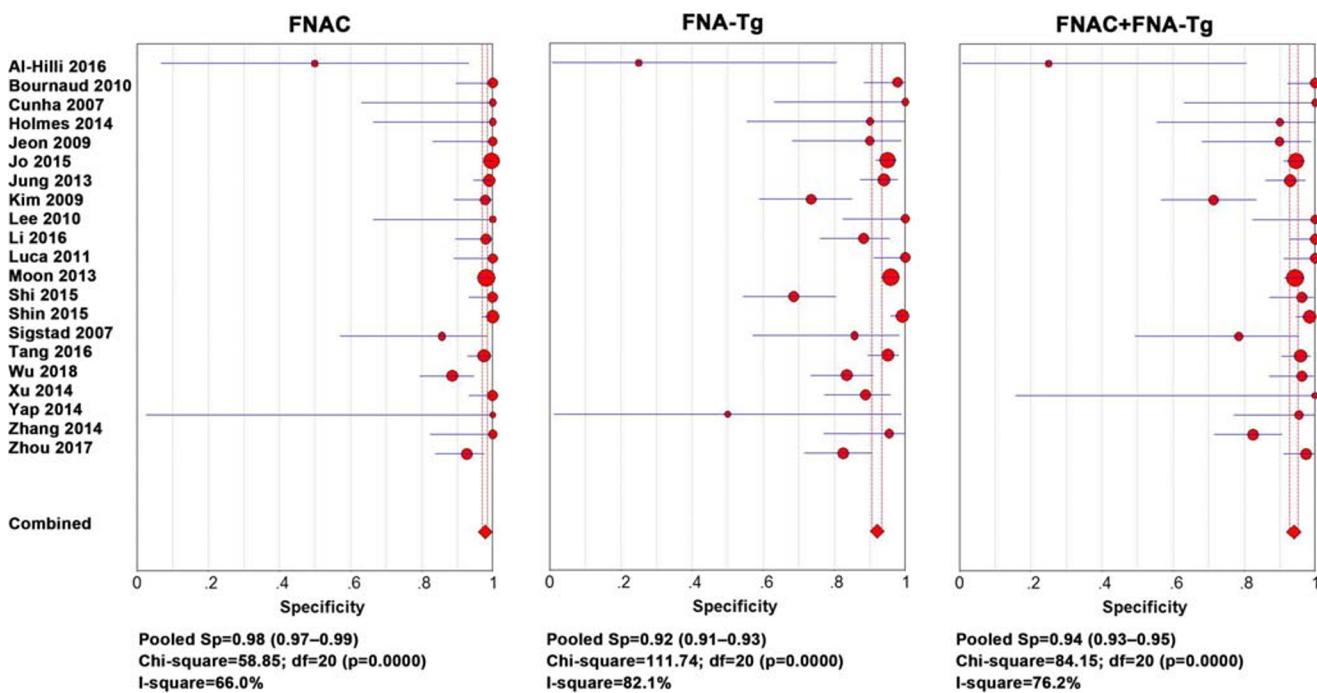**Fig. 4** Combined Se for FNAC, FNA-Tg, and FNAC+FNA-Tg

Fig. 5 Pooled Sp for FNAC, FNA-Tg, and FNAC+FNA-Tg

reached when using cut-offs < 1.0 ng/ml, while the highest Sp of FNAC+FNA-Tg was obtained when using FNA-Tg/serum-Tg ratio > 1 as the cut-off (Table 5).

Discussion

This study summarized the diagnostic performance of US-guided FNAC, FNA-Tg, and the combination of both for nodal metastasis assessment in preoperative and postoperative patients with DTC. Integrating results of the included studies, we found that FNAC had a better Sp while FNA-Tg got a higher Se, and that they could achieve a more excellent diagnostic performance when analyzed jointly than independently. FNAC and its combination with FNA-Tg performed better

when applied to recognize recurrence postoperatively, while FNA-Tg performed better when used to confirm staging preoperatively. On the whole, the combination of FNAC and FNA-Tg/serum-Tg ratio achieved a better diagnostic performance than the combination of FNAC and FNA-Tg. However, we should be aware that different categories of cut-off values may be suitable for different clinical situations.

A complete LN dissection is the key step of initial treatment for DTC. To reduce the number of preventative bilateral neck LN dissection and to build a personalized therapeutic strategy, the surgeons should be well informed of the metastatic status of neck LNs of individual patient [37]. After the initial treatment consisting in surgery with or without radioactive iodine, a timely recognition of recurrence in LNs could guarantee a timely intervention. FNAC is the most direct way
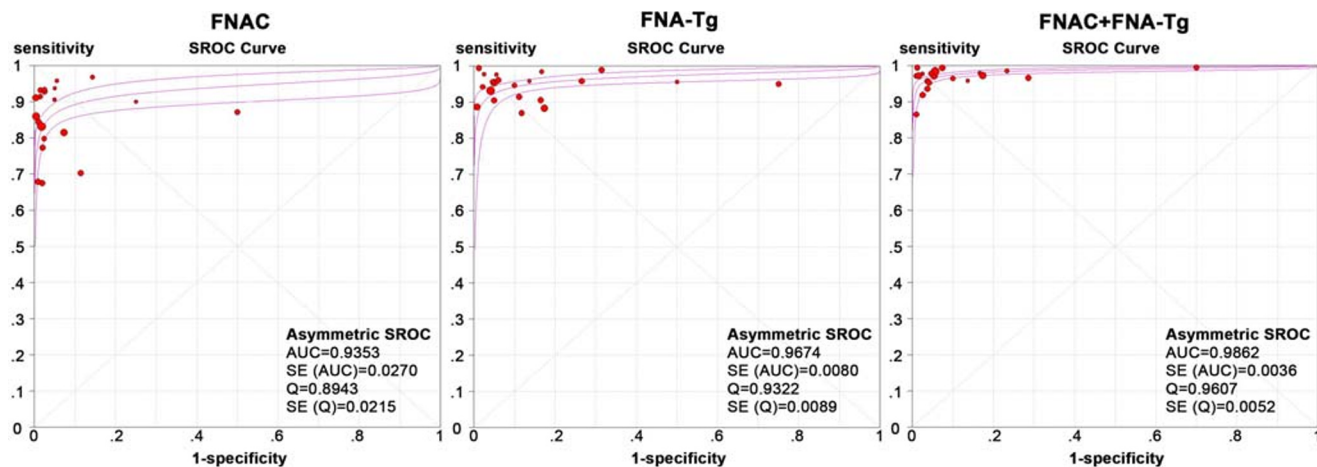

Fig. 6 SROC curves with AUC and Q value for FNAC, FNA-Tg, and FNAC+FNA-Tg

Table 4 The diagnostic performance in preoperative and postoperative subgroups

Test	Subgroup	Study number	Integrated outcomes						AUC	Q
			Se (95% CI) I²%	Sp (95% CI) I²%	PLR (95% CI) I²%	NLR (95% CI) I²%	DOR (95% CI) I²%	Cochran's Q p		
FNAC	Preop	10	0.79 (0.76–0.81) 88.5	0.95 (0.93–0.97) 68.3	12.61 (5.75–27.66) 72.5	0.20 (0.13–0.33) 87.6	86.34 (31.62–235.74) 64.3	25.17 0.0028	0.9488	0.8888
	Postop	9	0.66 (0.62–0.69) 93.0	0.98 (0.96–0.99) 83.4	19.10 (2.19–166.91) 91.4	0.20 (0.10–0.41) 94.2	141.82 (50.82–395.74) 17.8	9.74 0.2840	0.9560	0.8988
FNA-Tg	Preop	10	0.93 (0.91–0.95) 62.8	0.89 (0.86–0.91) 80.1	7.22 (3.74–13.93) 85.9	0.09 (0.06–0.14) 43.7	94.10 (41.84–211.60) 63.5	24.69 0.0033	0.9727	0.9243
	Postop	9	0.75 (0.72–0.79) 96.7	0.83 (0.79–0.86) 68.4	6.31 (3.83–10.40) 48.1	0.07 (0.01–0.49) 97.8	91.37 (37.46–222.87) 36.7	12.64 0.1249	0.9510	0.8918
FNAC+ FNA-Tg	Preop	9*	0.97 (0.96–0.98) 71.1	0.94 (0.91–0.96) 78.5	13.72 (4.99–37.77) 90.1	0.05 (0.03–0.07) 22.7	378.14 (185.46–771.02) 13.5	8.1 0.3242	0.9879	0.9534
	Postop	8*	0.96 (0.93–0.98) 59.1	0.97 (0.95–0.98) 39.4	21.27 (11.50–39.32) 20.3	0.04 (0.02–0.10) 57.8	788.72 (344.00–1808.37) 0	2.92 0.8919	0.9930	0.9664

*One involved study did not provide separate diagnostic data of FNAC+FNA-Tg based on operation status. *AUC* area under the curve, *CI* confidence intervals, *DOR* diagnostic odds ratio, *FNAC* fine needle aspiration cytology, *FNA-Tg/sTg* fine needle aspiration-thyroglobulin/serum-Tg, *NLR* negative likelihood ratio, *PLR* positive likelihood ratio, *Preop* preoperation, *Postop* postoperation, *Se* sensitivity, *Sp* specificity

to examine LNM before or after operation; it is highly specific because pathologists would not make a positive diagnosis until they observe definite tumor cells under the microscope. However, the Se of FNAC was only 0.82, as suggested by the pooled analysis in the current study (Table 3). Noticeably, the quality of aspiration material is largely dependent on the US guidance avoiding targets lacking cell components such as cystic part or necrotic areas [36]. Although the interpretation of conventional cytology is somewhat dependent on subjective experience of pathologists, it could achieve a more robust diagnostic efficacy when combined with other tests.

Since first reported in 1921 [38], the Se of the Tg level in the needle washout specimens for diagnosis of LNM has been well documented. Tg is only produced by the thyroid follicular cells and involved in thyroid hormone synthesis and iodine transport [7]. Like other auxiliary diagnostic examinations, such as BRAF mutation in cytology samples [39], the FNA-Tg test would not add patients' extra discomfort as it is an adjunct to FNA biopsy. Its interpretation is less dependent on personal experience than cytology. FNA-Tg could enhance FNAC diagnostic accuracy by decreasing false negative results at a modest cost [16]. Our analysis also indicated that the combination of FNAC and FNA-Tg could increase the pooled Se by 0.15 compared with FNAC along (Table 3). The DOR values of FNAC, FNA-Tg, and FNAC+FNA-Tg were 183.75, 155.17, and 446.00 in the total group, respectively, indicating the combination of FNAC and FNA-Tg has a stronger diagnostic power in discriminating LNM of DTC.

A quantitative measurement of Tg in LN aspirate samples is an objective indicator of DTC metastasis prior or after thyroidectomy. However, no definite consensus has been reached for the diagnostic threshold value of FNA-Tg worldwide [26]. Some researchers also mentioned that Tg values close to the cut-offs should be interpreted with caution. They may represent a minute metastatic focus, and conversely, Tg values before surgery might represent a false positive due to contamination with blood [30]. To avoid the interference of blood Tg, some studies suggested to use FNA-Tg/serum-Tg ratio as the positive standard instead [33, 36]. Our subgroup analysis reflected that the FNA-Tg performed less accurately using FNA-Tg/serum-Tg ratio than using Tg level as cut-off as a single test, but it could perform quite excellently when combined with FNAC using FNA-Tg/serum-Tg ratio as cut-off, especially in Sp (AUC 0.9860 and Sp 0.98). Obviously, no matter which FNA-Tg cut-off was employed, the diagnostic index such as Se, PLR, NLR, DOR, AUC, and Q value of the combination of FNA-Tg and FNAC was generally higher than that of FNA-Tg alone, again stressing the importance of combining FNA-Tg with FNAC in LNM detection in patients with DTC.

In a meta-analysis, intra-study variation and inter-study variation are the main causes for heterogeneity. According to our results, no threshold effect was detected in any included

Table 5 The subgroup analysis of FNA-Tg based on different cut-offs

Subgroup	Study number	Test	Integrated outcomes					AUC	Q
			Se (95% CI) I^2%	Sp (95% CI) I^2%	PLR (95% CI) I^2%	NLR (95% CI) I^2%	DOR (95% CI) I^2%		
Cut-off > 1.0 ng/ml	5	FNA-Tg	0.96 (0.94–0.97) 72.6	0.94 (0.92–0.96) 88.3	17.90 (5.88–54.46) 89.8	0.05 (0.03–0.09) 50.8	416.88 (215.01–808.28) 0	0.9886	0.9551
		FNAC+FNA-Tg	0.98 (0.96–0.99) 42.2	0.96 (0.94–0.97) 49.6	19.51 (12.24–31.09) 22.4	0.03 (0.02–0.05) 0	849.83 (399.32–1808.58) 0	0.9937	0.9684
Cut-off = 1.0 ng/ml	6	FNA-Tg	0.94 (0.91–0.96) 0	0.92 (0.89–0.94) 88.2	7.20 (2.64–19.67) 93.6	0.08 (0.06–0.11) 71.4	97.97 (35.21–272.59)	0.9502*	0.9278
		FNAC+FNA-Tg	0.97 (0.96–0.98) 52.8	0.91 (0.89–0.93) 88.2	7.58 (3.01–19.14) 93.2	0.04 (0.02–0.06) 10.3	254.26 (99.80–647.79) 55.5	0.9884	0.9546
0 < cut-off < 1.0 ng/ml	7	FNA-Tg	0.97 (0.93–0.99) 19.8	0.94 (0.860.98) 43	7.56 (2.49–22.98) 57.9	0.05 (0.02–0.12) 0	267.06 (67.78–1052.27) 0	0.9890	0.9560
		FNAC+FNA-Tg	0.99 (0.96–1.00) 0	0.95 (0.89–0.98) 62	11.51 (3.52–37.60) 0	0.03 (0.01–0.07) 0	583.42 (138.54–2456.95) 0	0.9949	0.9722
FNA-Tg/sTg ratio > 1	3	FNA-Tg	0.89 (0.85–0.92) 0	0.86 (0.80–0.91) 0	6.25 (4.31–9.06) 0	0.13 (0.10–0.18) 0	47.72 (27.88–81.67) 0	0.9435	0.8819
		FNAC+FNA-Tg	0.91 (0.85–0.95) 65.1	0.98 (0.95–1.00) 3.6	28.64 (7.75–105.85) 18.8	0.11 (0.07–0.18) 8.5	448.05 (117.15–1713.54) 0	0.9860	0.9493

*An asymmetrical SROC was plotted. *AUC* area under the curve, *CI* confidence intervals, *DOR* diagnostic odds ratio, *FNAC* fine needle aspiration cytology, *FNA-Tg/sTg* fine needle aspiration-thyroglobulin/serum-Tg, *NLR* negative likelihood ratio, *PLR* positive likelihood ratio, *Se* sensitivity, *Sp* specificity

study; however, the non-threshold effect evaluated by I^2 and Cochran's *Q* statistics was significant. Therefore, we adopted some measures to reduce the harmful influence of heterogeneity on overall conclusion: using a random effects model, reanalyzing data after deleting deviated studies, and conducting subgroup analyses. It should be noted that different detective combination should be chosen according to different detective purposes. For example, test combination with high Se, like FNAC+FNA-Tg with 0 < cut-off < 1.0 as cut-off, is suitable for screening in large-scale population; combination with high Sp, such as FNAC+FNA-Tg with FNA-Tg/serum-Tg > 1 as cut-off, is appropriate to assist in making a surgery plan. To avoid the risk of false negative results, close follow-up should be given when a less sensitive test, such as FNAC, was applied alone preoperatively or postoperatively.

Our study has several strengths and limitations. We included a large number of LNs and patients to improve the robustness of our analysis, and a search of the gray literature was performed and full texts available in other languages than English were also referred, which has increased the general applicability of our conclusions. Moreover, subgroup analyses were carefully conducted to explore the influence of patient status or FNA-Tg cut-offs on diagnostic performance. Admittedly, the number of involved studies, especially the studies using FNA-Tg/serum-Tg ratio as cut-off in Tg detection, was relatively limited. Additionally, significant non-threshold effect existed in the incorporated data. Further subgroup analysis suggested that different FNA-Tg cut-offs could add on the inter-study heterogeneity, but it did not impact the stability of the integrated outcomes very much, and the data still supported the conclusions on the whole.

Conclusion

In summary, FNA-Tg was more sensitive but less specific than FNAC in identifying LNM of DTC. The addition of FNA-Tg, especially the FNA-Tg/serum-Tg ratio, to FNAC could improve the diagnostic accuracy of LNM in patients with DTC and provide more comprehensive clues for preoperative staging and postoperative follow-up.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Prof. Cai-yun He.

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 Statistical analyses were performed using Meta-DiSc1.4 software and supervised by statistical expert prof. Cai-yun He.

Informed consent Written informed consent was not required for this study because only published data were used.

Ethical approval Institutional Review Board approval was not required because only published data were used and no animal subjects were involved in this study.

Study subjects or cohorts overlap Studies with duplicate data were excluded from the meta-analysis.

Methodology
This study is a meta-analysis.
• prospective
• diagnostic study
• multicenter study

References

1. Marotta V, Malandrino P, Russo M et al (2020) Fathoming the link between anthropogenic chemical contamination and thyroid cancer. Crit Rev Oncol Hematol 150:102950
2. Siegel RL, Miller KD, Jemal A (2019) Cancer statistics, 2019. CA Cancer J Clin 69:7–34
3. Lee J, Song Y, Soh EY (2014) Prognostic significance of the number of metastatic lymph nodes to stratify the risk of recurrence. World J Surg 38:858–862
4. Schlumberger MJ (1998) Papillary and follicular thyroid carcinoma. N Engl J Med 338:297–306
5. Wang LY, Palmer FL, Nixon IJ et al (2015) Lateral neck lymph node characteristics prognostic of outcome in patients with clinically evident N1b papillary thyroid cancer. Ann Surg Oncol 22:3530–3536
6. 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
7. Jiang HJ, Hsiao PJ (2020) Clinical application of the ultrasound-guided fine needle aspiration for thyroglobulin measurement to diagnose lymph node metastasis from differentiated thyroid carcinoma-literature review. Kaohsiung J Med Sci 36:236-243
8. Sohn YM, Kim MJ, Kim EK, Kwak JY (2012) Diagnostic performance of thyroglobulin value in indeterminate range in fine needle aspiration washout fluid from lymph nodes of thyroid cancer. Yonsei Med J 53:126–131
9. Salmaslioglu A, Erbil Y, Citlak G et al (2011) Diagnostic value of thyroglobulin measurement in fine-needle aspiration biopsy for detecting metastatic lymph nodes in patients with papillary thyroid carcinoma. Langenbecks Arch Surg 396:77–81
10. Kim M, Jeon MJ, Kim WG et al (2016) Comparison of thyroglobulin measurements using three different immunoassay kits: a BRAMHS Tg-Plus RIA kit, a BRAMHS hTg sensitive Kryptor kit, and a Beckman Coulter ACCESS immunoassay kit. Endocrinol Metab (Seoul) 31:462–468
11. Cho YY, Chun S, Lee SY, Chung JH, Park HD, Kim SW (2016) Performance evaluation of the serum thyroglobulin assays with immunochemiluminometric assay and immunoradiometric assay for differentiated thyroid cancer. Ann Lab Med 36:413–419
12. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56:1129–1135
13. Beynon R, Leeflang MM, McDonald S, et al (2013) Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.MR000022.pub3
14. Penny F, Whiting P, Anne WS et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155:529–536
15. Bolboacă S, Jäntschi L (2005) Binomial distribution sample confidence interval estimation for positive and negative likelihood ratio medical key parameters. AMIA Annu Symp Proc 2005:66–70
16. Al-Hilli Z, Strajina V, McKenzie TJ et al (2017) Thyroglobulin measurement in fine-needle aspiration improves the diagnosis of cervical lymph node metastases in papillary thyroid carcinoma. Ann Surg Oncol 24:739–744
17. Bournaud C, Charrie A, Nozieres C et al (2010) Thyroglobulin measurement in fine-needle aspirates of lymph nodes in patients with differentiated thyroid cancer: a simple definition of the threshold value, with emphasis on potential pitfalls of the method. Clin Chem Lab Med 48:1171–1177
18. Cunha N, Rodrigues F, Curado F et al (2007) Thyroglobulin detection in fine-needle aspirates of cervical lymph nodes: a technique for the diagnosis of metastatic differentiated thyroid cancer. Eur J Endocrinol 157:101–107
19. Holmes BJ, Sokoll LJ, Li QK (2014) Measurement of fine-needle aspiration thyroglobulin levels increases the detection of metastatic papillary thyroid carcinoma in cystic neck lesions. Cancer Cytopathol 122:521–526
20. Jeon SJ, Kim E, Park JS et al (2009) Diagnostic benefit of thyroglobulin measurement in fine-needle aspiration for diagnosing metastatic cervical lymph nodes from papillary thyroid cancer: correlations with US features. Korean J Radiol 10:106–111
21. Jo K, Kim MH, Lim Y et al (2015) Lowered cutoff of lymph node fine-needle aspiration thyroglobulin in thyroid cancer patients with serum anti-thyroglobulin antibody. Eur J Endocrinol 173:489–497
22. Jung JY, Shin JH, Han BK, Ko EY (2013) Optimized cutoff value and indication for washout thyroglobulin level according to ultrasound findings in patients with well-differentiated thyroid cancer. AJNR Am J Neuroradiol 34:2349–2353
23. Kim MJ, Kim EK, Kim BM et al (2009) Thyroglobulin measurement in fine-needle aspirate washouts: the criteria for neck node dissection for patients with thyroid cancer. Clin Endocrinol (Oxf) 70:145–151
24. Lee YH, Seo HS, Suh SI et al (2010) Cut-off value for needle washout thyroglobulin in athyrotropic patients. Laryngoscope 120:1120–1124
25. Giovanella L, Ceriani L, Suriano S (2011) Lymph node thyroglobulin measurement in diagnosis of neck metastases of differentiated thyroid carcinoma. J Thyroid Res 2011:621839
26. Moon JH, Kim YI, Lim JA et al (2013) Thyroglobulin in washout fluid from lymph node fine-needle aspiration biopsy in papillary thyroid cancer: large-scale validation of the cutoff value to determine malignancy and evaluation of discrepant results. J Clin Endocrinol Metab 98:1061–1068
27. Shi JH, Xu YY, Pan QZ, Sui GQ, Zhou JP, Wang H (2015) The value of combined application of ultrasound-guided fine needle aspiration cytology and thyroglobulin measurement for the diagnosis of cervical lymph node metastases from thyroid cancer. Pak J Med Sci 31:1152–1155
28. Shin HJ, Lee HS, Kim EK, Moon HJ, Lee JH, Kwak JY (2015) A study on serum antithyroglobulin antibodies interference in thyroglobulin measurement in fine-needle aspiration for diagnosing lymph node metastasis in postoperative patients. PLoS One 10:e0131096

29.	Sigstad E, Heilo A, Paus E et al (2007) The usefulness of detecting thyroglobulin in fine-needle aspirates from patients with neck lesions using a sensitive thyroglobulin assay. Diagn Cytopathol 35: 761–767

30.	Tang S, Buck A, Jones C, Sara Jiang X (2016) The utility of thyroglobulin washout studies in predicting cervical lymph node metastases: one academic medical center's experience. Diagn Cytopathol 44:964–968

31.	Yap NS, Maher R, Learoyd DL (2014) Any detectable thyroglobulin in lymph node biopsy washouts suggests local recurrence in differentiated thyroid cancer. Endocr Connect 3:150–155

32.	Zhang HS, Wang RJ, Fu QF et al (2014) Diagnostic value of thyroglobulin measurement with fine-needle aspiration biopsy for lymph node metastases in patients with a history of differentiated thyroid cancer. Asian Pac J Cancer Prev 15:10905–10909

33.	Li Q, Zhao BW, Lv JH et al (2016) Diagnostic value of fine-needle aspiration and thyroglobulin measurement in fine-needle aspiration for node metastasis in thyroid carcinoma before thyroidectomy. Chin J Ultrasound Med 32:400–403

34.	Wu Y, Yu XQ, Hu HY, Zhu GX, Guo SH, Du FY (2018) Value of fine needle aspiration cytology combined with fine needle aspiration thyroglobulin measurements in evaluation of metastatic lymph nodes in patients with papillary thyroid carcinoma. J Clin Ultrasound in Med 20:615–628

35.	Xu YY (2014) The value of joint application of ultrasonic imaging guided fine needle aspiration and thyroglobulin determination in diagnosis of postoperative lymph node metastasis of papillary thyroid cancer. Master Degree Thesis, Jilin University, Changchun, China

36.	Zhou L, Zhang G, Zhang DQ et al (2017) Application and evaluation of fine needle aspiration biopsy and eluate detection for cervical lymph node dissection decision in patients with thyroid carcinoma. Chin J Pract Surg 37:1002–1006

37.	Wang LY, Ganly I (2016) Nodal metastases in thyroid cancer: prognostic implications and management. Future Oncol 12:981–994

38.	Pacini F, Fugazzola L, Lippi F et al (1992) Detection of thyroglobulin in fine needle aspirates of nonthyroidal neck masses: a clue to the diagnosis of metastatic differentiated thyroid cancer. J Clin Endocrinol Metab 74:1401–1404

39.	Marotta V, Sapio MR, Guerra A, Vitale M (2011) BRAF mutation in cytology samples as a diagnostic tool for papillary thyroid carcinoma. Expert Opin Med Diagn 5:277–290