Deep learning to predict cervical lymph node metastasis from intraoperative frozen section of tumour in papillary thyroid carcinoma: a multicentre diagnostic study

Yihao Liu,[a,b,i] Fenghua Lai,[a,i] Bo Lin,[c,i] Yunquan Gu,[b,i] Lili Chen,[d,i] Gang Chen,[e,f,i] Han Xiao,[g] Shuli Luo,[a] Yuyan Pang,[e,f] Dandan Xiong,[e,f] Bin Li,[b] Sui Peng,[b] Weiming Lv,[c,***] Erik K. Alexander,[h,**] and Haipeng Xiao[a,*]

[a]Department of Endocrinology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[b]Clinical Trials Unit, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[c]Department of Breast and Thyroid Surgery, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[d]Department of Pathology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[e]Department of Pathology, The First Affiliated Hospital of Guangxi Medical University, Nanning, China
[f]Guangxi Zhuang Autonomous Region Engineering Research Center for Artificial Intelligence Analysis of Multimodal Tumour Images, The First Affiliated Hospital of Guangxi Medical University, Nanning, China
[g]Division of Interventional Ultrasound, Department of Medical Ultrasonics, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[h]Thyroid Section, Brigham & Women's Hospital, Harvard Medical School, Boston, USA

Summary

Background Lymph node metastasis (LNM) assessment in patients with papillary thyroid carcinoma (PTC) is of great value. This study aimed to develop a deep learning model applied to intraoperative frozen section for prediction of LNM in PTC patients.

Methods We established a deep-learning model (ThyNet-LNM) with the multiple-instance learning framework to predict LNM using whole slide images (WSIs) from intraoperative frozen sections of PTC. Data for the development and validation of ThyNet-LNM were retrospectively derived from four hospitals from January 2018 to December 2021. The ThyNet-LNM was trained using 1987 WSIs from 1120 patients obtained at the First Affiliated Hospital of Sun Yat-sen University. The ThyNet-LNM was then validated in the independent internal test set (479 WSIs from 280 patients) as well as three external test sets (1335 WSIs from 692 patients). The performance of ThyNet-LNM was further compared with preoperative ultrasound and computed tomography (CT).

Findings The area under the receiver operating characteristic curves (AUCs) of ThyNet-LNM were 0.80 (95% CI 0.74–0.84), 0.81 (95% CI 0.77–0.86), 0.76 (95% CI 0.68–0.83), and 0.81 (95% CI 0.75–0.85) in internal test set and three external test sets, respectively. The AUCs of ThyNet-LNM were significantly higher than those of ultrasound and CT or their combination in all four test sets (all $P < 0.01$). Of 397 clinically node-negative (cN0) patients, the rate of unnecessary lymph node dissection decreased from 56.4% to 14.9% by ThyNet-LNM.

Interpretation The ThyNet-LNM showed promising efficacy as a potential novel method in evaluating intraoperative LNM status, providing real-time guidance for decision. Furthermore, this led to a reduction of unnecessary lymph node dissection in cN0 patients.

Funding National Natural Science Foundation of China, Guangzhou Science and Technology Project, and Guangxi Medical High-level Key Talents Training "139" Program.

*Corresponding author. Department of Endocrinology, The First Affiliated Hospital of Sun Yat-sen University, No. 58, ZhongShan Second Road, Guangzhou, 510080, China.
**Corresponding author. Thyroid Section, Brigham & Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA, 02115, USA.
***Corresponding author. Department of Breast and Thyroid surgery, The First Affiliated Hospital of Sun Yat-sen University, No. 58, ZhongShan Second Road, Guangzhou, 510080, China.
    E-mail addresses: xiaohp@mail.sysu.edu.cn (H. Xiao), ekalexander@bwh.harvard.edu (E.K. Alexander), lvwm@mail.sysu.edu.cn (W. Lv).
[i]These authors contributed equally to this work.

Research in context

Evidence before this study

We searched PubMed on February 23, 2023, for research articles that contained the terms ["deep learning" OR "machine learning" OR "artificial intelligence" OR "convolutional neural network"] AND ["papillary thyroid carcinoma" OR "papillary thyroid cancer"] AND ["lymph node metastasis" OR "lymphatic metastasis"], without date or language restrictions. We identified twelve studies on the development and validation of artificial intelligence (AI) models for prediction of lymph node metastasis in papillary thyroid carcinoma. However, the development of these AI models were based on ultrasound images, or computed tomography images, or clinical characteristics. There was no study using histopathology images to build the AI model to predict lymph node metastasis in papillary thyroid carcinoma.

Added value of this study

To the best of our knowledge, this is the first study to develop a deep-learning model (ThyNet-LNM) for intraoperative prediction of cervical lymph node metastasis based on frozen section images of papillary thyroid carcinoma. ThyNet-LNM demonstrated better predictive performance compared with either pre-operative cervical ultrasound and computed tomography examination in multicentre cohorts. Furthermore, ThyNet-LNM reduced the rate of unnecessary lymph node dissection in clinically node-negative patients with T1-T2 stage.

Implications of all the available evidence

ThyNet-LNM showed promising efficacy as a potential novel method with practical clinical applicability in evaluating intraoperative lymph node metastasis status, providing real-time guidance for decision and a reduction of unnecessary lymph node dissection in clinically node-negative patients.

Introduction

Approximately 90% of patients with thyroid cancer harbor papillary thyroid carcinoma (PTC), the most common thyroid malignancy.[1] The problem of over-diagnosis in PTC is quite grave.[2] It is necessary to develop methods to identify aggressive PTC and reduce the overtreatment. Lymph node metastasis (LNM) is an important indicator of PTC prognosis and is correlated with an increased risk of local recurrence, typically seen in 30–60% of PTC patients.[3,4] Therefore, accurate identification of cervical LNM occurrence is of high clinical importance.

Cervical ultrasound and neck computed tomography (CT) examinations are the preferred methods for pre-operative detection of LNM in PTC. However, the accuracy of ultrasound and even more neck CT for cervical LNM detection are limited. Indeed, recent studies demonstrated that 60–70% of central LNM were missed diagnosis by cervical ultrasound or CT.[5,6] Therefore, central lymph node dissection (LND) is often considered for PTC patients throughout China and some other Asia-Pacific countries due to the unreliability of preoperative imaging examinations for operative guidance.[7–9] Worldwide, however, the role of prophylactic central LND for clinically node-negative (cN0) patients with PTC remains debated. A major argument against prophylactic LND was that it possibly increased the risk of operative complications and medical cost.[10] Hence, a reliable method to accurately predict LNM at the time of surgery itself and reduce unnecessary LND is desired.

The development of artificial intelligence (AI) applied to histopathology image analysis has provided a new means to predict LNM in PTC patients. Previous studies have reported that pathological features such as micro-calcification, lymphovascular invasion and extra-thyroidal extension were associated with cervical LNM in PTC.[11–13] However, the traditional pathological image analysis relies on the experience of pathologists, often days after the surgery is complete, and the rich biological information contained in pathological images cannot be fully recognized by the naked eyes. Deep learning, especially if applied to the real-time operative setting, could automatically extract pathological features from frozen section analysis which are invisible to the naked eyes and potentially predicted the biological behavior of tumours.[14] Recent studies indicated that deep learning models based on pathological images of tumour can accurately predict LNM in prostate cancer and colon cancer.[15,16]

Fine needle aspiration (FNA) is an important pre-operative tool for identifying benign and malignant nodules, which requires experienced radiologists and cytologists.[17,18] The availability of medical resources varies in Asia-Pacific countries and regions, making it difficult to generally implement FNA into clinical practice.[19] A recent multicentre prospective study demonstrated that only 37% of patients with differentiated thyroid cancer received preoperative FNA, and about 90% of patients underwent intraoperative frozen sections during initial surgical treatment in China.[20] Intraoperative frozen section is an important method for rapid pathology-based diagnosis and can guide further surgical decisions during thyroidectomy. Therefore, applicable to our medical environment, we

sought to create a deep learning model based upon real-time frozen section images and determine its ability to guide further intraoperative decision-making for PTC patients.

Materials and methods

Patients and datasets
Data for the development and validation of ThyNet-LNM were retrospectively derived from four hospitals (the First Affiliated Hospital of Sun Yat-sen University (FAHSYSU), the First Affiliated Hospital of Guangxi Medical University (FAHGXMU), Shunde Hospital of Southern Medical University (SDHSMU), and Sun Yat-Sen University Cancer Center (SYSUCC)) from January 2018 to December 2021 (Appendix Fig. S1). The inclusion criteria were: (1) aged ≥18 years old, (2) a pathologic diagnosis of PTC after thyroidectomy, (3) tissue was obtained for intraoperative frozen section anlaysis, (4) patients received LND and the total number of lymph nodes dissected ≥5, (5) patients were pathologically confirmed with or without LNM. The exclusion criteria were: (1) patients with disease recurrence at admission, (2) patients with distant metastases, (3) intraoperative frozen sections without tumour tissue, (4) low-quality intraoperative frozen sections (e.g., poor staining, fragmentation, twisty images, and a large amount of ice crystal).

Preoperative work-up, intraoperative frozen section and surgery
The routine preoperative work-up included physical examination, cervical image examination (including high-resolution ultrasound, with or without CT), and thyroid function test. Patients without preoperative clinical evidences of LNM by physical examination and cervical imaging evaluations were defined as cN0 patients, otherwise as clinically node-positive (cN1) patients. Frozen sections of the suspicious thyroid nodule were done intraoperatively and they must be diagnosed within 20 min after receipt. All cN0 patients received elective central LND. Therapeutic LND was performed in cN1 patients.

Data preparation
The whole slide images (WSIs) of intraoperative frozen sections were digitized at 40× magnification by a scanning machine (KFBIO, Ningbo, China) and were stored in TIFF file format. Two experienced pathologists reviewed the WSIs independently to exclude the duplicate images and control the images quality. Disagreements were resolved through discussion. To extract the information under different magnifications, WSIs meeting quality standards were cut into non-overlapping 512 × 512 pixels patches at magnification scales of 5×, 10×, 20×, and 40×, respectively. Patches with over 50% of background coverage were excluded.

Image sampling and magnification scale selection
To clarify the contribution of different tissue areas (tumour area, peri-tumour area, and the whole WSI) in the WSI to the prediction of LNM, the performances of models developed based on different tissue areas were compared. We randomly selected a subset with 300 WSIs from training set to develop a segmentation network (Appendix Methods).

In this study, each patient had one or more frozen slides. We compared the effects of independent slide score and different slide ensemble scores (mean score, minimum score, and maximum score of all slides) on the prediction models under different magnification scales. In addition, to further determine the optimal number of sampling patches for the prediction model, sensitivity analyses of the number of sampling patches under different magnifications were performed in this subset. In the preliminary experiment, the performance of model under 40× magnification scale was significantly lower than those of other magnifications, with the accuracy about 0.5. Therefore, 40× magnification was excluded from the final ensemble model. The ensemble model integrating magnification scales of 5×, 10×, and 20× was conducted.

Training of ThyNet-LNM
The ThyNet-LNM model was constructed based on a weakly supervised multiple-instance learning (MIL) framework.[21] The framework consisted of a convolutional neural network feature extraction layer, a MIL pooling layer and a fully connected layer. In the preliminary experiment, we used 300 WSIs (the same subset from training set used for developing segmentation network) to compare the performances of deep learning models constructed using different convolutional neural networks. By comparing the training time and accuracy of the networks, we found that Inception-v4 had the best comprehensive performance (Appendix Fig. S2). Therefore, a pre-trained Inception-v4 network was used as the backbone to extract the features of patches (Appendix Methods). Patch bags and their corresponding pathologic labels were input to train the prediction network. In the MIL pooling layer, the attention mechanism was introduced to aggregate the patch features through the attention score, and finally output the predicted value of the slide level through the fully connected layer (Appendix Fig. S3). The maximum predicted value of all slides in a patient was selected as the optimal prediction score. The average of optimal prediction scores integrating magnification scales of 5×, 10×, and 20× formed the final prediction score of ThyNet-LNM (Fig. 1A). Gradient weighted class activation mapping (Grad-CAM) was applied to provide an insight into regions of WSI that ThyNet-LNM used to generate predictions.
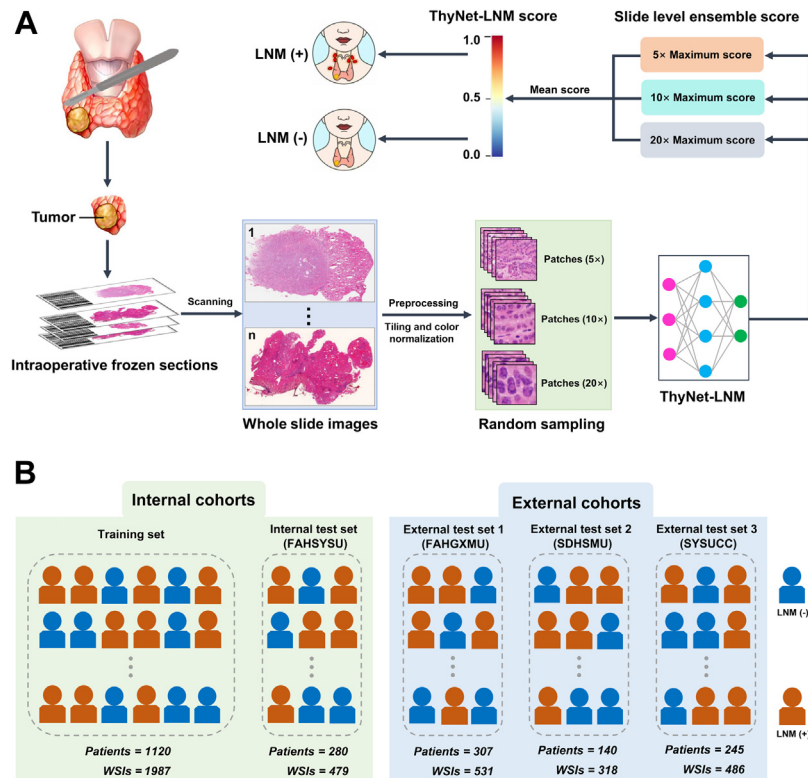
Fig. 1: Workflow of ThyNet-LNM and datasets allocation. (A) All WSIs were obtained from the intraoperative frozen sections of PTC. Patches sampled randomly from WSIs under different magnification scales were input to the prediction network. The predicted value of the slide level was output. The maximum predicted value of all slides in a patient was selected as the optimal prediction score. The average of optimal prediction scores integrating magnifications of 5×, 10×, and 20× formed the patient level ensemble score (ThyNet-LNM score). When the ThyNet-LNM score exceeded a certain threshold, the patient was predicted to be LNM (+). (B) PTC patients from the FAHSYSU were randomly divided into a training set and an independent internal test set. Patients from other three hospitals were enrolled as external test sets. FAHGXMU, First Affiliated Hospital of Guangxi Medical University; FAHSYSU, First Affiliated Hospital of Sun Yat-sen University; LNM, Lymph node metastasis; SDHSMU, Shunde Hospital of Southern Medical University; SYSUCC, Sun Yat-sen University Cancer Center; WSIs, Whole slide images.

Validation of ThyNet-LNM

The predictive performance of ThyNet-LNM was evaluated in the internal test cohort (FAHSYSU test set) and then three separate external test cohorts (FAHGXMU, SDHSMU, and SYSUCC). We further compared the performance of ThyNet-LNM with preoperative cervical imaging examinations. In this study, all PTC patients underwent cervical ultrasound, but only some underwent cervical CT. The LNM diagnoses of cervical imaging examinations were obtained based on the diagnostic reports of cervical ultrasound and CT from patients' medical records. If the diagnostic reports indicated that the lymph nodes were "malignant" or "metastatic" or "abnormal" or "suspicious", we defined them as positive LNM of cervical ultrasound or CT. Firstly, we calculated the performance of ultrasound alone in all patients. Then we evaluated the performance of combination of cervical ultrasound and CT in patients underwent both cervical ultrasound and CT. Negative

LNM imaging of both ultrasound and CT was defined as negative, and positive LNM was defined when at least one of the ultrasound and CT imaging was positive. Subgroup analysis was performed in cN0 patients and cN1 patients. In addition, we evaluated whether ThyNet-LNM could reduce the unnecessary LND rate in cN0 patients with T1-T2 stage.

Statistical analysis

Receiver operating characteristic (ROC) curves were generated to evaluate the performances of different models. The area under the ROC curve (AUC) values were calculated accordingly. An optimal cut-off of prediction score of ThyNet-LNM was determined by the ROC curve to reach the best accuracy in internal test set, which was 0.376 in this study. Logistic regression coefficients for the prediction probability calculations were applied to combination models of ThyNet-LNM with cervical imaging examinations.[22] A two by two confusion

matrix with the number of true positive, false positive, false negative, and true negative values was generated for each diagnosis. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated according to the confusion matrix. The 95% confidence interval (CI) for each value was calculated by the exact Clopper-Pearson method.[23] A two-sided DeLong test was used to compare the AUCs.[24] The number of unnecessary LND was compared by Chi-square test between the prophylactic central LND and ThyNet-LNM recommended. For baseline characteristics, continuous variables were described as median (interquartile range, IQR), and the categorical variables were described as the number of cases (percentage). All statistical analyses were performed using SAS software (version 9.4). Comparisons with two-sided P values less than 0.05 were considered to be statistically significant.

Ethics statement
This study complied with the Declaration of Helsinki and was approved by the Research Ethics Committee of FAHSYSU (Ethics number: [2023]189). Informed consent was waived for the retrospectively nature.

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results
Patient characteristics
A total of 2466 WSIs from 1400 PTC patients obtained at FAHSYSU were randomly divided into a training set (1987 WSIs of 1120 patients) and separately an independent internal test set (479 WSIs from 280 patients). For external test sets, 531 eligible WSIs from 307 PTC patients were enrolled from FAHGXMU, 318 WSIs from 140 PTC patients at SDHSMU, and 486 WSIs from 245 PTC patients at SYSUCC, respectively (Fig. 1B). Baseline characteristics of the patients in five datasets were summarized in Table 1. The proportion of patients with pathologically node-positive (pN1) in five datasets ranged from 55.5% to 58.5%.

Development of ThyNet-LNM
The performance of segmentation network was test, achieving AUCs of 0.98 (95% CI 0.96–0.99), 0.96 (95% CI 0.94–0.99), 0.99 (95% CI 0.96–1.00), and 0.99 (95% CI 0.97–1.00) under 5×, 10×, 20×, and 40× magnification scales, respectively (Appendix Fig. S4). We then evaluated the contribution of different tissue areas (tumour area, peri-tumour area, and the whole WSI) in the WSI

for the prediction of LNM status. The performances between different tissue categories and different magnification scales were compared. The models using patches from the WSI as inputs were significantly higher than those using patches from the peri-tumour area or tumour area at every magnification scale (all $P < 0.05$, Appendix Fig. S5A).

Next, the effects of independent slide score and different slide ensemble scores (mean score, minimum score, and maximum score) on the prediction model were compared. The models reached the best AUCs when maximum score of all slides was used as the slide ensemble score under 5×, 10×, and 20× magnification scales (Appendix Fig. S5B). Then, sensitivity analyses of the number of sampling patches from the WSIs under different magnification scales were performed to determine the optimal number of sampling patches for the prediction model. The models reached the best AUCs when 64 patches were randomly sampled under 5×, 10×, and 20× magnification scales (Appendix Fig. S5C). The ensemble model integrating 5×, 10×, and 20× magnification scales with 64 patches from the WSI as inputs reached an AUC of 0.86 (95% CI 0.79–0.91). The AUC of ensemble model was significantly higher than that under a single magnification scale (all $P < 0.05$, Appendix Fig. S5D). Therefore, we took the ensemble model with maximum score of all slides and integrating 5×, 10×, and 20× magnification scales and the sampling strategy with 64 patches from the WSI as the backbone model for the following training and validation of ThyNet-LNM. Grad-CAM showed the regions localized by ThyNet-LNM for LNM prediction in true positive instances and true negative instances, respectively (Appendix Fig. S6).

Validation of ThyNet-LNM
The AUC of ThyNet-LNM in the training set reached 0.83 (95% CI 0.80–0.85) (Appendix Fig. S7). Then, the performance of ThyNet-LNM was validated in the internal and external test sets. The AUCs of ThyNet-LNM were 0.80 (95% CI 0.74–0.84), 0.81 (95% CI 0.77–0.86), 0.76 (95% CI 0.68–0.83), and 0.81 (95% CI 0.75–0.85) in the internal test set (FAHSYSU), external test set 1 (FAHGXMU), external test set 2 (SDHSMU) and external test set 3 (SYSUCC), respectively (Fig. 2A–D).

The performance of ThyNet-LNM as an intra-operative tool to guide therapy was further compared with traditional pre-operative cervical imaging examination performed for such purposes. For all patients, the AUCs of cervical ultrasound were 0.69 (95% CI 0.63–0.74), 0.59 (95% CI 0.53–0.64), 0.58 (95% CI 0.50–0.67), and 0.65 (95% CI 0.59–0.71) in the internal test set (FAHSYSU), external test set 1 (FAHGXMU), external test set 2 (SDHSMU) and external test set 3 (SYSUCC), respectively (Fig. 2A–D). The AUCs from ThyNet-LNM was significantly higher than that of cervical ultrasound in all four test sets (all $P < 0.001$).

Characteristics	Training set (N = 1120)	Internal test set (FAHSYSU, N = 280)	External test set 1 (FAHGXMU, N = 307)	External test set 2 (SDHSMU, N = 140)	External test set 3 (SYSUCC, N = 245)	P value
Age, median (IQR)	39.0 (32.0, 47.0)	39.0 (32.0, 48.0)	38.0 (31.0, 46.0)	39.0 (32.0, 48.0)	38.0 (32.0, 48.0)	0.385
Sex, N (%)						0.282
Male	272 (24.3)	79 (28.2)	68 (22.1)	42 (30.0)	62 (25.3)	
Female	848 (75.7)	201 (71.8)	239 (77.9)	98 (70.0)	183 (74.7)	
LNM status confirmed by pathology, N (%)						0.946
pN0	465 (41.5)	118 (42.1)	128 (41.7)	59 (42.1)	109 (44.5)	
pN1	655 (58.5)	162 (57.9)	179 (58.3)	81 (57.9)	136 (55.5)	
LNM status assessed by ultrasound, N (%)						<0.001
Negative	586 (52.3)	163 (58.2)	147 (47.9)	110 (78.6)	125 (51.0)	
Positive	534 (47.7)	117 (41.8)	160 (52.1)	30 (21.4)	120 (49.0)	
Lymph node assessed by CT, N (%)						<0.001
No assessment	682 (60.9)	170 (60.7)	74 (24.1)	76 (54.3)	11 (4.5)	
Assessment	438 (39.1)	110 (39.3)	233 (75.9)	64 (45.7)	234 (95.5)	
LNM status assessed by CT[a], N (%)						<0.001
Negative	144 (32.9)	37 (33.6)	153 (65.7)	29 (45.3)	95 (40.6)	
Positive	294 (67.1)	73 (66.4)	80 (34.3)	35 (54.7)	139 (59.4)	
LNM status assessed by ultrasound[a], N (%)						<0.001
Negative	132 (30.1)	31 (28.2)	94 (40.3)	25 (39.1)	93 (39.7)	
Positive	306 (69.9)	79 (71.8)	139 (59.7)	39 (60.9)	141 (60.3)	
Clinically LNM status						<0.001
cN0	531 (47.4)	144 (51.4)	119 (38.8)	90 (64.3)	112 (45.7)	
cN1	589 (52.6)	136 (48.6)	188 (61.2)	50 (35.7)	133 (54.3)	
N grade[b] by pathology, N (%)						0.017
0	465 (41.5)	118 (42.1)	127 (41.4)	59 (42.1)	109 (44.5)	
1a	400 (35.7)	103 (36.8)	118 (38.4)	67 (47.9)	122 (49.8)	
1b	255 (22.8)	59 (21.1)	62 (20.2)	14 (10.0)	14 (5.7)	
Total number of lymph nodes dissected, median (IQR)	12.0 (7.0, 21.0)	13.0 (8.0, 20.0)	10.0 (7.0, 15.0)	13.0 (8.0, 19.0)	9.0 (6.0, 13.0)	0.106
ACR-TIRADS of tumor, N (%)						<0.001
3	13 (1.2)	1 (0.4)	1 (0.3)	11 (7.9)	5 (2.0)	
4	131 (11.7)	24 (8.6)	88 (28.7)	56 (40.0)	17 (6.9)	
5	976 (87.1)	255 (91.1)	218 (71.0)	73 (52.1)	223 (91.0)	
Tumour number[c], N (%)						0.097
Single	714 (63.8)	191 (68.2)	215 (70.0)	92 (65.7)	144 (58.8)	
Multiple	406 (36.3)	89 (31.8)	92 (30.0)	48 (34.3)	101 (41.2)	
Tumour diameter (mm), median (IQR)	10.0 (7.0, 15.0)	10.0 (7.0, 14.0)	10.0 (6.0, 15.0)	9.0 (7.0, 14.0)	8.0 (6.0, 13.0)	0.210
Microscopic extrathyroidal extension, N (%)						0.002
Yes	129 (11.5)	35 (12.5)	36 (11.7)	29 (20.7)	17 (6.9)	
No	991 (88.5)	245 (87.5)	271 (88.3)	111 (79.3)	228 (93.1)	
T stage[b] by pathology, N (%)						0.013
T1-2	981 (87.6)	243 (86.8)	267 (87.0)	111 (79.3)	225 (91.8)	
T3-4	139 (12.4)	37 (13.2)	40 (13.0)	29 (20.7)	20 (8.2)	
Preoperative FNA						0.786
Yes	420 (37.5)	102 (36.4)	112 (36.5)	50 (35.7)	95 (38.8)	
No	700 (62.5)	178 (63.6)	195 (63.5)	90 (64.3)	150 (61.2)	
Early postoperative hypocalcemia						0.542
Yes	583 (52.1)	153 (54.6)	170 (55.4)	79 (56.4)	123 (50.2)	
No	537 (47.9)	127 (45.4)	137 (44.6)	61 (43.6)	122 (49.8)	

Abbreviations: ACR-TIRADS, American College of Radiology-thyroid imaging reporting and data system; cN0, clinically node-negative; cN1, clinically node-positive; CT, Computed tomography; FAHGXMU, First Affiliated Hospital of Guangxi Medical University; FAHSYSU, First Affiliated Hospital of Sun Yat-sen University; FNA, Fine needle aspiration; IQR, Inter quartile range; LNM, Lymph node metastasis; pN0, pathologically node-negative; pN1, pathologically node-positive; SDHSMU, Shunde Hospital of Southern Medical University; SYSUCC, Sun Yat-sen University Cancer Center. [a]The LNM status accessed by ultrasound or CT were conducted in patients receiving both ultrasound and CT. [b]T stage and N grade were defined according to the American Joint Committee on Cancer thyroid cancer staging system (8th edition). [c]The Tumour number was confirmed by postoperative paraffin pathology.

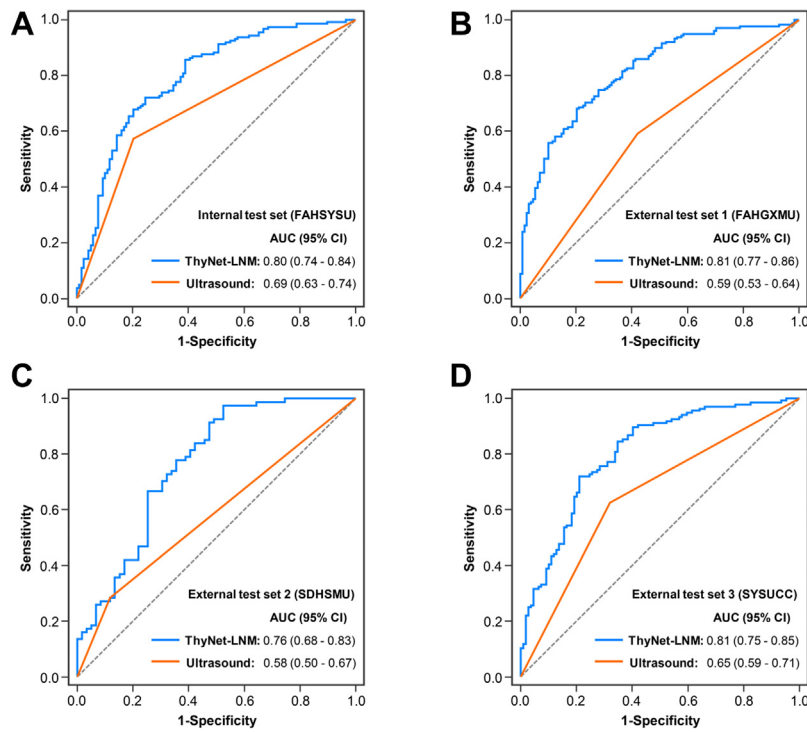*Table 1:* **Baseline characteristics of the patients in five datasets.**

Fig. 2: Performances of ThyNet-LNM and cervical ultrasound for prediction of cervical lymph node metastasis. Predictive performance of ThyNet-LNM and cervical ultrasound in the internal test set (FAHSYSU) (A), external test set 1 (FAHGXMU) (B), external test set 2 (SDHSMU) (C), and external test set 3 (SYSUCC) (D). AUC, Area under the receiver operating characteristic curve; CI, Confidence interval; FAHGXMU, First Affiliated Hospital of Guangxi Medical University; FAHSYSU, First Affiliated Hospital of Sun Yat-sen University; SDHSMU, Shunde Hospital of Southern Medical University; SYSUCC, Sun Yat-sen University Cancer Center.

Details of performances of ThyNet-LNM and cervical ultrasound for all patients in test sets were summarized in Appendix Table S1.

Among patients who received both cervical ultrasound and CT, the AUCs of ThyNet-LNM were also significantly higher than those of cervical ultrasound, CT, and the combination of both (all $P < 0.01$, Fig. 3). However, the integrated combination of ThyNet-LNM with cervical ultrasound and CT did not improve further predictive performance (all $P > 0.05$). Details of performances of ThyNet-LNM and cervical imaging examinations in patients with cervical ultrasound and CT are listed in Appendix Table S2.

We also conducted subgroup analysis in cN0 and cN1 patients on the performance of ThyNet-LNM. As shown in Appendix Fig. S8, the AUCs of ThyNet-LNM in cN0 patients were relatively higher than those in cN1 patients (0.80 [95% CI 0.76–0.85] vs. 0.72 [95% CI 0.67–0.75] in internal test set, 0.85 [95% CI 0.80–0.89] vs. 0.77 [95% CI 0.73–0.80] in external test set 1, 0.78 [95% CI 0.74–0.84] vs. 0.73 [95% CI 0.67–0.75] in external test set 2, and 0.81 [95% CI 0.77–0.87] vs. 0.78 [95% CI 0.74–0.83] in external test set 3, respectively, all $P < 0.05$. The sensitivity, specificity, PPV, NPV of ThyNet-LNM in cN0 and cN1 patients are summarized in Appendix Table S3. In four test sets, a total of 397 cN0 patients with T1–T2 stage received prophylactic central LND. If LND is performed according to the results of ThyNet-LNM among these patients, the rate of unnecessary LND will be reduced from 56.4% to 14.9% ($P < 0.001$) (Fig. 4).

Discussion

To the best of our knowledge, this is the first study to develop a deep-learning predictive model utilizing intraoperative frozen section images to guide operative strategy for removal of localized disease among PTC patients. The ThyNet-LNM was generated, and then independently validated in both internal and external prospective cohorts. We found that application of ThyNet-LNM during the procedure demonstrated better predictive performance compared with either preoperative cervical ultrasound and CT examination. Furthermore, ThyNet-LNM reduced the rate of unnecessary LND in cN0 patients with T1–T2 stage.

The phenotypic information in pathological tissue reflected the overall effect of the tumour microenvironment on the behavior of cancer cells. Recent studies demonstrated that deep learning models can identify
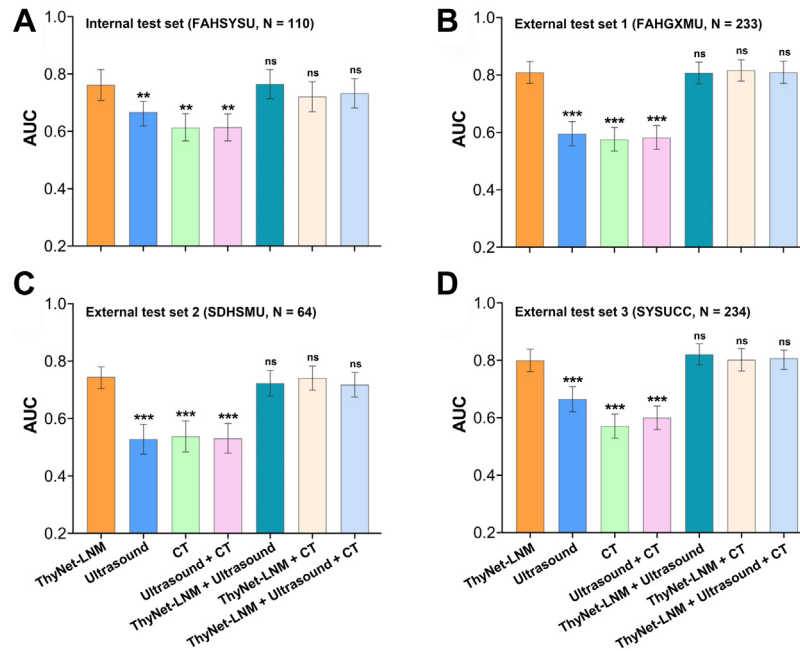
Fig. 3: **Performances of ThyNet-LNM and preoperative imaging examinations for patients with cervical ultrasound and CT.** Comparison of AUCs of ThyNet-LNM and preoperative imaging examinations for patients with cervical ultrasound and CT the internal test set (FAHSYSU) (A), external test set 1 (FAHGXMU) (B), external test set 2 (SDHSMU) (C) and external test set 3 (SYSUCC) (D). ** 0.01 > P > 0.001, ***P < 0.001, [ns]P > 0.05 for comparison of AUC with ThyNet-LNM. AUC, Area under the receiver operating characteristic curve; CT, Computed tomography; FAHGXMU, First Affiliated Hospital of Guangxi Medical University; FAHSYSU, First Affiliated Hospital of Sun Yat-sen University; LNM, Lymph node metastasis; SDHSMU, Shunde Hospital of Southern Medical University; SYSUCC, Sun Yat-sen University Cancer Center.
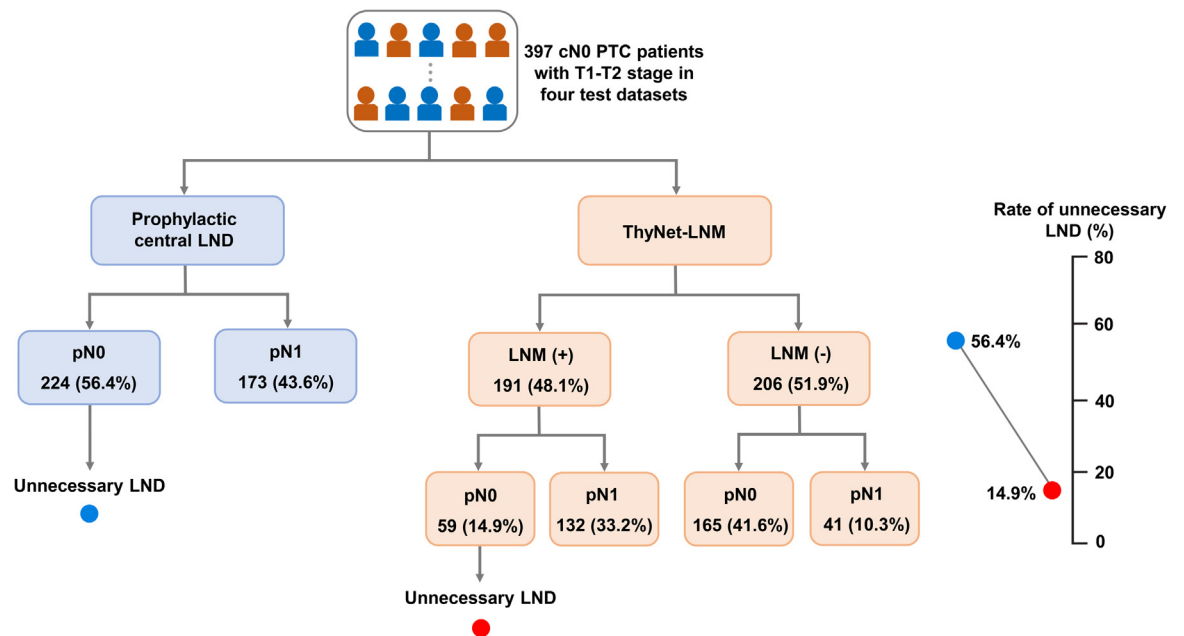
Fig. 4: **Recommendation for LND according to ThyNet-LNM in cN0 PTC patients with T1-T2 stage.** The T stage was defined according to the American Joint Committee on Cancer thyroid cancer staging system (8th edition). cN0, clinically node-negative; LND, Lymph node dissection; LNM, Lymph node metastasis; pN0, pathologically node-negative; pN1, pathologically node-positive.

protein expression alterations and genetic mutations based on histological images across multiple cancer types.[25] Deep learning therefore has the potential to identify oncological relevant patterns by analyzing histological images and using these patterns to predict invasive biological behavior (such as risk of metastasis or recurrence).[26] Previous deep learning models based on thyroid pathological images mainly focused on differential diagnosing of malignancy from benign nodules.[27,28] In the present study, we constructed a deep learning model to predict LNM of thyroid cancer based on pathological images. Our results showed that this histology-based deep learning model was well-validated in the independent internal and external cohorts, indicating the model had considerable generalization ability. However, the specific histological characteristics that the deep learning model used to detect LNM remained unknown due to the "black-box" nature of convolutional neural network.

To evaluate the contribution of different tissue categories for LNM prediction, the patches from the tumour area, peri-tumour area, and WSI were used as input. We found that the model based on WSI achieved the best predictive performance, indicating that pathological features from both tumour area and peri-tumour area were associated with LNM. It was concordant with previous study that the biological information on oncological outcome was not limited to the tumour area.[29] The tumour-infiltrating lymphocytes and the immune cells in peri-tumour were also related to LNM in PTC patients.[30,31] Interestingly, we also found that the performance of WSI-based model was diverse under different magnification scales. The reason could be that images under different magnification scales revealed different types of histological features.[32] Images under high magnification scale better reflected the details of cell morphology and internal cell structure, while images under lower magnification scales revealed the overall cell morphology, the distribution pattern of tumour cells and their surroundings. The histological features under different magnification scales were important for LNM prediction. By assembling models under different magnification scales, the final model may better combined cell morphology, tumour microenvironment and intercell relationships, and eventually achieved better predictive performance. Together these support the novelty and unique aspects that AI brings to the clinical environment.

In the present study, the preoperative FNA rate among training and validation sets was only 36.4–38.8%. In Asia-Pacific countries, frozen section is an important alternative diagnostic tool in lieu of widely available FNA.[33] The information of intraoperative frozen section may provide guidance for surgeons to make further surgical decisions. In such a medical circumstance, we proposed an intraoperative process to assess cervical LNM based on frozen sections. In this scenario, the pathologists will make the classification diagnosis of tumour and at the same time, the frozen section slides will be scanned and inputted to ThyNet-LNM without increasing operative time. The prediction results of ThyNet-LNM can provide guidance for surgeons to make further surgical decisions for LND. ThyNet-LNM based on intraoperative frozen section images can compensate for the limitations of preoperative cervical ultrasound or CT assessment of lymph nodes and allow PTC patients to receive the most reasonable treatment.

Therapeutic LND is generally recommended for cN1 patients with PTC by guidelines, while the role of prophylactic central LND for cN0 patients is still controversial.[34] It was reported that 60–70% of central LNM failed to be detected by preoperative cervical imaging,[5,6] and the undetected LNM increased the risk of local recurrence.[35] In Asia-Pacific countries, to minimize local recurrence and reoperation, surgeons may perform routine prophylactic central LND for cN0 patients with PTC.[7–9] However, prophylactic LND may bring the increased risk of parathyroid gland injury, recurrent laryngeal nerve paralysis, bleeding and other surgical complications, instead of survival benefit.[36] The incidence of early postoperative hypocalcemia in the present study was 50.2–56.4%, which was higher than those of other studies (most patients included undergoing thyroidectomy without LND).[37,38] A personalized method is necessary for reducing unnecessary prophylactic LND. In the present study, ThyNet-LNM reduced the rate of unnecessary LND in cN0 patients with T1-T2 stage from 56.4% to 14.9%. Therefore, ThyNet-LNM may be a potential method for individualized LND in PTC patients.

The present study has several limitations. First, the main limitation of this study is that ThyNet-LNM is a binary classifier, meaning that the model does not specifically predict LNM in central and lateral districts. Second, the reproducibility of this study may be limited in European countries and the United States due to the small use of intraoperative frozen section during thyroidectomy in these countries. Third, there may be similarities among these datasets that the training and test datasets were obtained from South China and three of them were from the same province. Further validation in hospitals from different countries and regions is necessary to address this concern. Lastly, this study is a retrospective study, and further prospective studies are needed to improve the diagnostic effectiveness of ThyNet-LNM and apply it into clinical practice.

In conclusion, the ThyNet-LNM showed promising efficacy as a potential novel method with practical clinical applicability in evaluating intraoperative LNM status, providing real-time guidance for decision and a reduction of unnecessary LND in cN0 patients. As FNA remains the standard for initial care in guidelines, moving toward greater use of FNA is still a laudable goal. Encouragingly, the application rate of FNA in China and some other

developing countries have kept increasing. Further efforts might be paid to use AI approach on FNA cytology–to see if it predicts LNM as well.

References
1 Miranda-Filho A, Lortet-Tieulent J, Bray F, et al. Thyroid cancer incidence trends by histology in 25 countries: a population-based study. *Lancet Diabetes Endocrinol*. 2021;9:225–234.
2 Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. *N Engl J Med*. 2016;375:614–617.
3 Lundgren CI, Hall P, Dickman PW, Zedenius J. Clinically significant prognostic factors for differentiated thyroid carcinoma: a population-based, nested case-control study. *Cancer*. 2006;106:524–531.
4 Ito Y, Kudo T, Kobayashi K, Miya A, Ichihara K, Miyauchi A. Prognostic factors for recurrence of papillary thyroid carcinoma in the lymph nodes, lung, and bone: analysis of 5,768 patients with average 10-year follow-up. *World J Surg*. 2012;36:1274–1278.
5 Zhao H, Li H. Meta-analysis of ultrasound for cervical lymph nodes in papillary thyroid cancer: diagnosis of central and lateral compartment nodal metastases. *Eur J Radiol*. 2019;112:14–21.
6 Xing Z, Qiu Y, Yang Q, et al. Thyroid cancer neck lymph nodes metastasis: meta-analysis of US and CT diagnosis. *Eur J Radiol*. 2020;129:109103.
7 Guidelines Working Committee of Chinese Society of Clinical Oncology. Guidelines of Chinese society of clinical oncology (CSCO) differentiated thyroid cancer. *J Cancer Control Treat*. 2021;34:1164–1201 [In Chinese].
8 Yi KH, Lee EK, Kang H, et al. 2016 revised Korean thyroid association management guidelines for patients with thyroid nodules and thyroid cancer. *Int J Thyroidol*. 2016;9:59–126 [In Korean].
9 The Japan Association of Endocrine Surgeons and the Japanese Society of Thyroid Surgery. The JAES/JSTS task force on the guidelines for thyroid tumors. Japanese clinical practice guidelines for thyroid tumors. *J JAES JSTS*. 2018;35 [In Japanese].
10 Carling T, Long WR, Udelsman R. Controversy surrounding the role for routine central lymph node dissection for differentiated thyroid cancer. *Curr Opin Oncol*. 2010;22:30–34.
11 Qu H, Sun GR, Liu Y, He QS. Clinical risk factors for central lymph node metastasis in papillary thyroid carcinoma: a systematic review and meta-analysis. *Clin Endocrinol*. 2015;83:124–132.
12 Roh JL, Kim JM, Park CI. Central lymph node metastasis of unilateral papillary thyroid carcinoma: patterns and factors predictive of nodal metastasis, morbidity, and recurrence. *Ann Surg Oncol*. 2011;18:2245–2250.
13 Oh HS, Park S, Kim M, et al. Young age and male sex are predictors of large-volume central neck lymph node metastasis in clinical no papillary thyroid microcarcinomas. *Thyroid*. 2017;27:1285–1290.
14 Niazi M, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol*. 2019;20:e253–e261.
15 Wessels F, Schmitt M, Krieghoff-Henning E, et al. Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer. *BJU Int*. 2021;128:352–360.
16 Kwak MS, Lee HH, Yang JM, et al. Deep convolutional neural network-based lymph node metastasis prediction for colon cancer using histopathological images. *Front Oncol*. 2020;10:619803.
17 Scappaticcio L, Trimboli P, Iorio S, et al. Repeat thyroid FNAC: inter-observer agreement among high- and low-volume centers in naples metropolitan area and correlation with the EU-TIRADS. *Front Endocrinol*. 2022;13:1001728.
18 Choi SH, Han KH, Yoon JH, et al. Factors affecting inadequate sampling of ultrasound-guided fine-needle aspiration biopsy of thyroid nodules. *Clin Endocrinol*. 2011;74:776–782.
19 Yang SP, Ying LS, Saw S, Tuttle RM, Venkataraman K, Su-Ynn C. Practical barriers to implementation of thyroid cancer guidelines in the Asia-Pacific region. *Endocr Pract*. 2015;21:1255–1268.
20 Ming J, Zhu JQ, Zhang H, et al. A multicenter, prospective study to observe the initial management of patients with differentiated thyroid cancer in China (DTCC study). *BMC Endocr Disord*. 2021;21:208.
21 Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. 2019;25:1301–1309.
22 Hosmer D, Lemeshow S, Sturdivant RX. *Applied logistic regression*. New York, NY: A Wiley-Interscience Publication; 2001.
23 Clopper C, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26:404–413.
24 Delong ER, Delong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
25 Fu Y, Jung AW, Torne RV, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Cancer*. 2020;1:800–810.
26 van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. *Nat Med*. 2021;27:775–784.
27 Lin YJ, Chao TK, Khalil MA, et al. Deep learning fast screening approach on cytological whole slides for thyroid cancer diagnosis. *Cancers*. 2021;13.
28 Dov D, Kovalsky SZ, Assaad S, et al. Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. *Med Image Anal*. 2021;67:101814.
29 Yamamoto Y, Tsuzuki T, Akatsuka J, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. *Nat Commun*. 2019;10:5642.
30 Modi J, Patel A, Terrell R, Tuttle RM, Francis GL. Papillary thyroid carcinomas from young adults and children contain a mixture of lymphocytes. *J Clin Endocrinol Metab*. 2003;88:4418–4425.
31 Jeong JS, Kim HK, Lee CR, et al. Coexistence of chronic lymphocytic thyroiditis with papillary thyroid carcinoma: clinical manifestation and prognostic outcome. *J Korean Med Sci*. 2012;27:883–889.
32 Saco A, Ramirez J, Rakislova N, Mira A, Ordi J. Validation of whole-slide imaging for histolopathical diagnosis: current state. *Pathobiology*. 2016;83:89–98.
33 Liu Z, Liu D, Ma B, et al. History and practice of thyroid fine-needle aspiration in China, based on retrospective study of the practice in Shandong University Qilu Hospital. *J Pathol Transl Med*. 2017;51:528–532.
34 Canu GL, Medas F, Conzo G, et al. Is prophylactic central neck dissection justified in patients with cN0 differentiated thyroid carcinoma? An overview of the most recent literature and latest guidelines. *Ann Ital Chir*. 2020;91:451–457.

35 Maksimovic S, Jakovljevic B, Gojkovic Z. Lymph node metastases papillary thyroid carcinoma and their importance in recurrence of disease. *Med Arch*. 2018;72:108–111.

36 Dismukes J, Fazendin J, Obiarinze R, et al. Prophylactic central neck dissection in papillary thyroid carcinoma: all risks, no reward. *J Surg Res*. 2021;264:230–235.

37 Li X, Yidiresi A, Tian Y, Zhang L, Luo J. Clinical analysis of risk factors of early hypocalcemia after total thyroidectomy. *Chin J Oper Proc Gen Surg*. 2021;15:77–79 [In Chinese].

38 Wang W, Li X, Xia F, Bai N, Zhang Z. Risk factors for hypoparathyroidism after thyroidectomy. *J Cen South Univ (Med Sci)*. 2019;44:315–321 [In Chinese].