Remodeling ^{99m}Tc-Pertechnetate Thyroid Uptake: Statistical, Machine Learning, and Deep Learning Approaches

Geoffrey M. Currie[1] and Basit Iqbal[2]

[1]*Charles Sturt University, Wagga Wagga, Australia, and Baylor College of Medicine, Houston, Texas; and* [2]*Gujranwala Institute of Nuclear Medicine and Radiotherapy, Gujranwala, Pakistan*

Although reference ranges for ^{99m}Tc thyroid percentage uptake vary, the seemingly intuitive evaluation of thyroid function does not reflect the complexity of thyroid pathology and biochemical status. The emergence of artificial intelligence in nuclear medicine has driven problem solving associated with logic and reasoning, warranting reexamination of established benchmarks in thyroid functional assessment. **Methods:** This retrospective study of 123 patients compared scintigraphic findings with grounded truth established through biochemistry status. Conventional statistical approaches were used in conjunction with an artificial neural network to determine predictors of thyroid function from data features. A convolutional neural network was also used to extract features from the input tensor (images). **Results:** Analysis was confounded by subclinical hyperthyroidism, primary hypothyroidism, subclinical hypothyroidism, and triiodothyronine toxicosis. Binary accuracy for identifying hyperthyroidism was highest for thyroid uptake classification using a threshold of 4.5% (82.6%), followed by pooled physician interpretation with the aid of uptake values (82.3%). Visual evaluation without quantitative values reduced accuracy to 61.0% for pooled physician determinations and 61.4% classifying on the basis of thyroid gland intensity relative to salivary glands. The machine learning (ML) algorithm produced 84.6% accuracy; however, this included biochemistry features not available to the semantic analysis. The deep learning (DL) algorithm had an accuracy of 80.5% based on image inputs alone. **Conclusion:** Thyroid scintigraphy is useful in identifying hyperthyroid patients suitable for radioiodine therapy when using an appropriately validated cutoff for the patient population (4.5% in this population). ML artificial neural network algorithms can be developed to improve accuracy as second-reader systems when biochemistry results are available. DL convolutional neural network algorithms can be developed to improve accuracy in the absence of biochemistry results. ML and DL do not displace the role of the physician in thyroid scintigraphy but can be used as second-reader systems to minimize errors and increase confidence.

Key Words: thyroid uptake; hyperthyroidism; machine learning; deep learning; artificial intelligence

J Nucl Med Technol 2022; 50:143–152
DOI: 10.2967/jnmt.121.263081

In 1967, Atkins and Richards (*1*) evaluated the potential role of ^{99m}Tc-pertechnetate in evaluating thyroid function as an alternative to sodium iodide with ^{131}I on the basis that ^{99m}Tc uptake in the thyroid reflects the gland's trapping function. This landmark work used a probe detector rather than γ-camera imaging for the uptake calculation. A small number of hypothyroid patients were included, and all had percentage uptakes below 0.5%. Only 2 of 15 hyperthyroid patients fell below 4%, whereas 4 of 133 euthyroid patients had uptake above 4%. Thus, a cutoff for normality was set at 0.4%–4.0% to provide 87% accuracy in hyperthyroidism, 97% accuracy in euthyroidism, and 100% accuracy in hypothyroidism.

Later work, in 1973, by Maisey et al. (*2*) used a γ-camera, pinhole collimation, and interfaced computer to generate regions of interest for calculation of ^{99m}Tc-pertechnetate uptake in the thyroid. Uptake was 0.2%–3.6% in euthyroid patients, 0.3%–6.2% in the presence of a goiter, 2.8%–8.8% in hyperthyroidism, and 0.1%–0.3% in hypothyroidism, leading to establishment of a reference range of 0.3%–3.4%. More recently, ^{99m}Tc-pertechnetate uptake in euthyroidism was characterized in the range of 0.4%–1.7% in 47 clinically normal patients (*3*). It is widely acknowledged that reference values change with geography and time, particularly in relation to iodine deficiency (*4*). Although widespread use of international standards is common (0.5%–4.5% for example), these values may not reflect either the technique used (probe vs. γ-camera) or population characteristics (e.g., iodine deficiency). In Namibia, investigators found the reference range to be 0.15%–2.14% (*4*) although the study included only 76 patients and all were euthyroid. A U.K. study (*5*) used 60 euthyroid patients to estimate the local reference range as 0.2%–2.0%.

Although reference ranges for percentage uptake vary, the method for calculation of thyroid function on ^{99m}Tc scintigraphy also varies (*6*). A seemingly intuitive evaluation of thyroid function has also been used as a visual evaluation of thyroid activity relative to salivary gland activity (Fig. 1). Such an evaluation does not reflect the complexity of thyroid pathology and biochemical status. When the bulk of patients are euthyroid or hyperthyroid, this simplification is intuitive, but it fails to accommodate subclinical hyperthyroidism, which can produce a low thyroid accumulation of ^{99m}Tc;

Received Aug. 19, 2021; revision accepted Nov. 10, 2021.
For correspondence or reprints, contact Geoffrey M. Currie (gcurrie@csu.edu.au).
Published online Dec. 7, 2021.

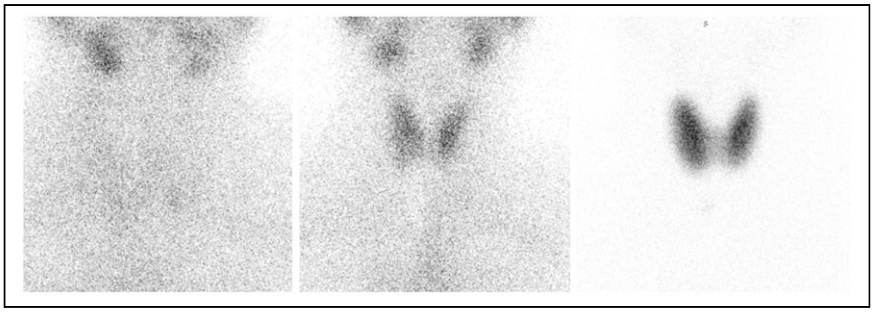

FIGURE 1. Intuitive, but sometimes inaccurate, visual evaluation of thyroid status relative to salivary gland activity. (Left) Salivary gland activity exceeding thyroid gland activity suggests hypothyroidism. (Middle) Salivary gland activity and thyroid gland activity being similar (within same scale) suggests euthyroidism. (Right) Salivary gland activity not being apparent relative to thyroid activity suggests hyperthyroidism. All images were obtained with ^{99m}Tc-pertechnetate using high-resolution, parallel-hole imaging.

triiodothyronine (T3) toxicosis, which can have high or low ^{99m}Tc uptake; subclinical hypothyroidism, which can have elevated or normal ^{99m}Tc accumulation; and primary hypothyroidism, which can have normal or elevated ^{99m}Tc accumulation. Thus, the accuracy of ^{99m}Tc uptake may be more dependent on the pathologic cross section of patients than on the technique itself.

The emergence of artificial intelligence in nuclear medicine has driven problem solving associated with logic and reasoning (*7,8*). Developments in machine learning (ML) and deep learning (DL) provide valuable research tools, particularly for image segmentation and interpretation (*9*). The artificial neural network (ANN) provides the backbone for both ML and DL algorithms. The ANN relies on input of specific data (features) and generally refers to ML. More complex ANNs can produce deep architectures (a high number of layers and nodes) and refers to DL. Deep ANNs are generally associated, in medical imaging, with convolutional neural networks (CNN) that use convolution and pooling layers to extract features from input tensors (images) (*9,10*). Although there have been historical uses of neural networks to classify thyroid-based ophthalmologic conditions and evaluate in vitro laboratory tests, it is only recently that DL approaches have been applied to thyroid scintigraphy. Using SPECT thyroid scintigraphy, 3 DL models based on AlexNet, VGGNet, and ResNet architectures trained on 1,430 clinical studies were modeled and compared with residents in nuclear medicine (*11*). Although the investigators concluded that DL approaches performed well in thyroid scintigraphy, the role of DL might be limited to assisting the physician in training rather than having any specific clinical utility. The algorithms marginally outperformed first-year residents but did not perform as well as second-year residents, let alone experienced physicians. Concurrent use of the DL approaches improved the performance of residents on the order of 5% and reduced reporting time. Nonetheless, there is a need to explore potential clinical and research applications, and the less complex nature of planar thyroid scintigraphy may be better suited to DL approaches. The performance of these algorithms was enhanced by a sanitized dataset with a case population comprising healthy individuals (175), patients with Graves disease (834), and patients with subacute thyroiditis (421). The 3 DL architectures reported a high degree of recall for subacute thyroiditis, poor accuracy for normality, and moderate accuracy for Graves disease (*11*).

The aim of this investigation was to correlate each of the following with biochemical status and compare performance: percentage uptake of ^{99m}Tc, visual correlation of thyroid activity in the thyroid, ML algorithms using an ANN, and DL approaches using a CNN.

MATERIALS AND METHODS

The study retrospectively analyzed 123 patients (90.2% female), with a mean age of 35 y (range, 10–70 y). The mean intravenous dose of ^{99m}Tc was 153.4 MBq. ^{99m}Tc-based thyroid uptake was determined using background-corrected thyroid regions of interest and a measured standard. All calculations were decay-corrected and accounted for residual dose in the syringe after injection. The extracted image features included both background-corrected and non–background-corrected total thyroid, left-side and right-side area (cm^2), counts, and counts per pixel. The ratio of the right lobe to the left lobe for area (cm^2), counts, and counts per pixel was also determined with and without background correction. Additionally, the ratio of thyroid count to background count for total thyroid, right lobe, and left lobe was determined (trapping index). The dose relative to the total count was also calculated, and visual classification of thyroid activity relative to the salivary glands was recorded. Biochemical features included the levels of free thyroxine (T4) (pmol/L), free T3 (pmol/L), and thyroid-stimulating hormone (μIU/mL). The biochemical status of the patient was determined (Table 1) and was further stratified as ternary (hypothyroid, euthyroid, or hyperthyroid) or binary (hyperthyroid or not hyperthyroid) (*1–6,12,13*). Other imaging features were also recorded (e.g., hot or cold nodule and multinodular goiter). Only 96 patients had both imaging features and biochemical status available. The investigation was approved by the institutional ethics committee.

Conventional statistical analysis was undertaken using JMP software (version 15.2.1; SAS Institute). The statistical significance was calculated using χ^2 analysis for nominal data and the Student t test for continuous data. The Pearson χ^2 test was used for categoric data with a normal distribution, and the likelihood ratio χ^2 test was used for categoric data without a normal distribution. F test ANOVA was used to determine statistically significant differences within grouped data. A P value of less than 0.05 was considered significant. Interobserver correlation was evaluated with χ^2 analysis, and interobserver reliability was measured using the Cohen κ-coefficient.

The data were also evaluated using an ANN (Neural Analyser, version 2.9.5; Artificial Intelligence Techniques, Ltd.). There were

TABLE 1
Biochemical Stratification of Patient Studies and Findings (*1–6,12,13*)

Free T3 (2–7 pmol/L*)	Free T4 (12–30 pmol/L*)	Thyroid-stimulating hormone (0.45–4.5 μIU/mL*)	Biochemical status	^{99m}Tc uptake (%)	Comment on uptake reference range
High	High	Low	Hyperthyroidism	>4.5	0% FN rate
Normal	Normal	Low	Subclinical hyperthyroidism	<4.5 including <0.45 or absent	0% TP, comprised FN or FP hypothyroidism
High	Normal	Low	T3 toxicosis	>4.5 or <0.45	FP hypothyroidism
Normal	High	Low	Thyroiditis		No cases
Low	Low	Low	Secondary hypothyroidism		No cases
Normal	Normal	High	Subclinical hypothyroidism	>0.45 but <4.5	100% FN
Low or normal	Low	High	Primary hypothyroidism	>0.45 and in over 50% of cases >4.5	100% FN
Normal	Normal	Normal	Euthyroid	<4.5%	9% FP rate (6% hyperthyroid, 3% hypothyroid)

*Reference range.

42 input variables in 123 patients (instances) using a binary classification of hyperthyroid or euthyroid. A 50:25:25 split of 96 valid instances (excluded missing biochemistry data) was used for training, selection, and testing. The initial network architecture included 16 scaling layer inputs and 3 hidden layers of 6, 4, and 3 nodes, using a logistic activation function (defines the output of each node based on its input) for a single probabilistic layer (binary). The weighted squared error method was used to determine the loss index, and the neural parameter norm was used for the regularization method. A quasi-Newtonian training method was applied using gradient information to estimate the inverse Hessian matrix for each iteration of the algorithm (no second derivatives). The loss function associated with the training phase estimates the error associated with the data that the neural network observes.

A single anterior neck image for the 96 patients was evaluated by 3 independent expert physicians masked to other image and biochemical features. On the basis of the visual appearance, each scan was recorded as euthyroid, hypothyroid, or hyperthyroid. On completion of the stratification, each physician reevaluated the ternary status, with the visual inspection supplemented by the calculated thyroid uptake (%). The physician rating was determined by majority group consensus.

Individual, nonannotated, anterior neck images representative of each patient were evaluated using a CNN classifier (Deep Learning Toolkit Deep Network Designer App in MATLAB, version R2020b; MathWorks). Given the lack of discriminatory power of either visual evaluation or thyroid uptake quantitation using various cutoffs to identify hypothyroidism, the CNN classifier was designed to identify hyperthyroidism or no hyperthyroidism (euthyroid and hypothyroid). Given the lack of complexity in the image data, the architecture used for the CNN was initially modeled on a binary version of AlexNet with 25 layers but optimized using a model that resembled the VGG-19 CNN architecture with a binary output and 30 layers (Table 2; Fig. 2). All patient files were trained and validated 3 times (70:30 random data split) for each of 3 image types; white on black gray scale, black on white gray scale, and the magnitude spectrum of the Fourier transformation of each image (Fig. 3). Specific parameters included an ADAM (adaptive movement estimation) stochastic gradient descent optimizer algorithm, an initial learn rate of 0.001, a maximum of 50 epochs (1 epoch = 1 iteration), and randomization with each epoch.

Situation analysis was undertaken using the confusion matrix for classifier prediction, including true-positives (TPs), false-positives (FPs), true-negatives (TNs), and false-negatives (FNs). Several performance indicators can be gleaned from the confusion matrix, including precision (TPs/[TPs + FPs]), recall (TPs/[TPs + FNs]), accuracy ([TPs + TNs]/[TPs + TNs + FPs + TNs]), and F1 score ($2 \times$ TPs/[$2 \times$ {TPs + FPs + FNs}]).

RESULTS

Statistical Analysis

For the 123 patients, the mean thyroid uptake was 4.4% (95% CI, 3.3%–5.5%), with a median of 2.2% (Table 3). Among the visual findings, 9 patients had increased uptake associated with primary hypothyroidism, 22 had increased uptake because of Graves disease, 9 had multinodular goiters, 2 had nodular thyroids, 28 had a normal morphology, 3 had goiters, 11 had reduced or absent uptake, 7 had autonomous glands with contralateral suppression (6 on the right), 24 had cold nodules (16 on the right), and 8 had hot nodules (4 on the right). Table 4 summarizes other key demographic data.

The mean age of hypothyroid patients (48.0 y) was statistically higher than that for biochemically euthyroid patients (33.7 y) ($P = 0.041$) but not for hyperthyroid patients (36.7 y). There was also a weak positive correlation between age and thyroid size ($P < 0.001$; $R^2 = 0.117$). No other statistically significant relationships were noted for patient age. Men demonstrated a statistically higher mean thyroid area (48.5 cm^2) than women (32.2 cm^2) ($P = 0.003$). There was also a statistically significant difference in the biochemical status ($P = 0.019$), with a disproportionately high

TABLE 2
CNN Architecture, Activations, and Parameters

Layer	Name	Activations	Parameters
1	Tensor input layer	[725,725,3]	
2	2D convolution layer	[239,239,64]	Weights [11,11,3,64], bias [1,1,64]
3	Batch normalization	[239,239,64]	Offset and scale [1,1,64]
4	ReLU layer	[239,239,64]	
5	Max pooling layer	[119,119,64]	Size [3,3], stride [2,2], padding [0,0,0,0]
6	2D convolution layer	[40,40,128]	Weights 5,5,64,128], bias [1,1,128]
7	Batch normalization	[40,40,128]	Offset and scale [1,1,128]
8	ReLU layer	[40,40,128]	
9	Max pooling layer	[19,19,128]	Size [3,3], stride [2,2], padding [0,0,0,0]
10	2D convolution layer	[19,19,256]	Weights [3,3,128,256], bias [1,1,256]
11	Batch normalization	[19,19,256]	Offset and scale [1,1,256]
12	ReLU layer	[19,19,256]	
13	Max pooling layer	[9,9,256]	Size [3,3], stride [2,2], padding [0,0,0,0]
14	2D convolution layer	[9,9,192]	Weights [3,3,256,192], Bias [1,1,192]
15	Batch normalization	[9,9,192]	Offset and scale [1,1,192]
16	ReLU layer	[9,9,192]	
17	Max pooling layer	[4,4,192]	Size [3,3], stride [2,2], padding [0,0,0,0]
18	2D convolution layer	[4,4,192]	Weights [3,3,256,192], bias [1,1,192]
19	Batch normalization	[4,4,192]	Offset and scale [1,1,192]
20	ReLU layer	[4,4,192]	
21	Max pooling layer	[1,1,192]	Size [3,3], stride [2,2], padding [0,0,0,0]
22	Fully connected layer	[1,1,192]	Weights [192,192], bias [192,1]
23	ReLU layer	[1,1,192]	
24	Dropout layer	[1,1,192]	0.5
25	Fully connected layer	[1,1,86]	Weights [86,192], bias [86,1]
26	ReLU layer	[1,1,86]	
27	Dropout layer	[1,1,86]	0.5
28	Fully connected layer	[1,1,2]	Weights [2,86], bias [2,1]
29	Softmax layer	[1,1,2]	
30	Classification layer		Cross entropy loss function

2D = 2-dimensional; ReLU = rectified linear unit.

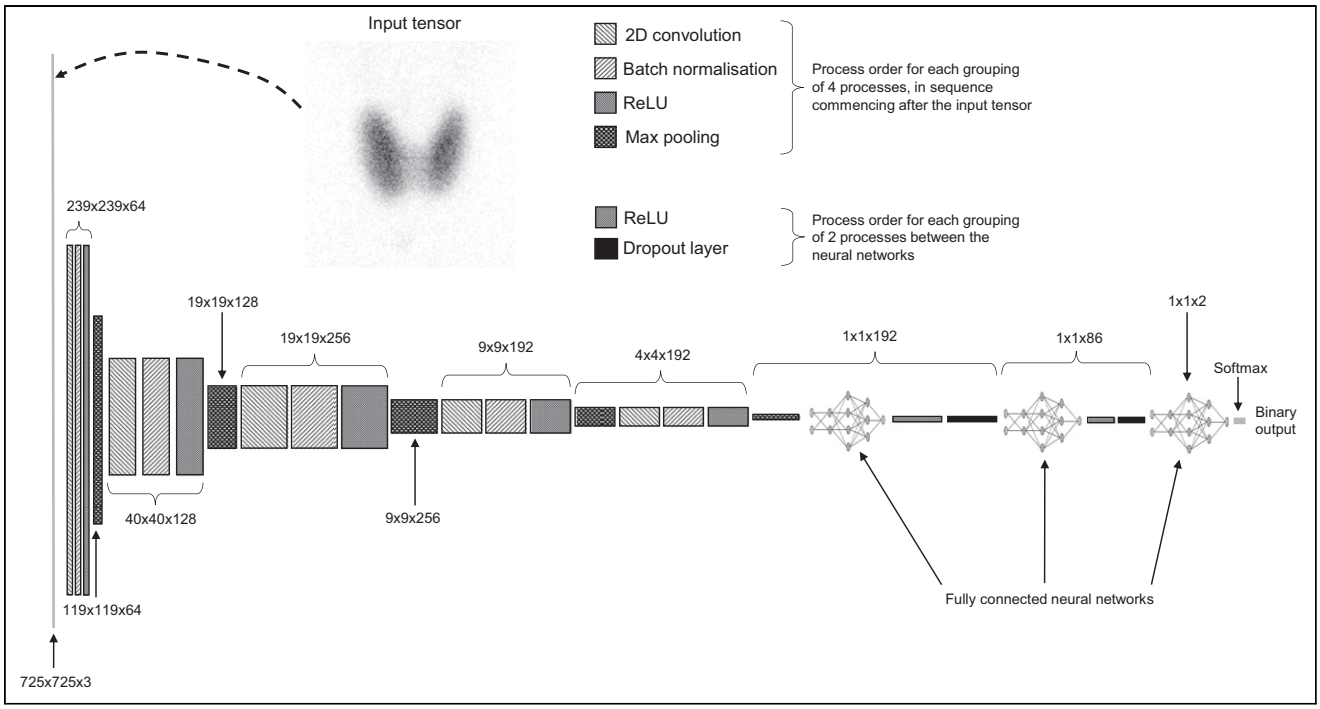

FIGURE 2. CNN architecture. 2D = 2-dimensional; ReLU = rectified linear unit.

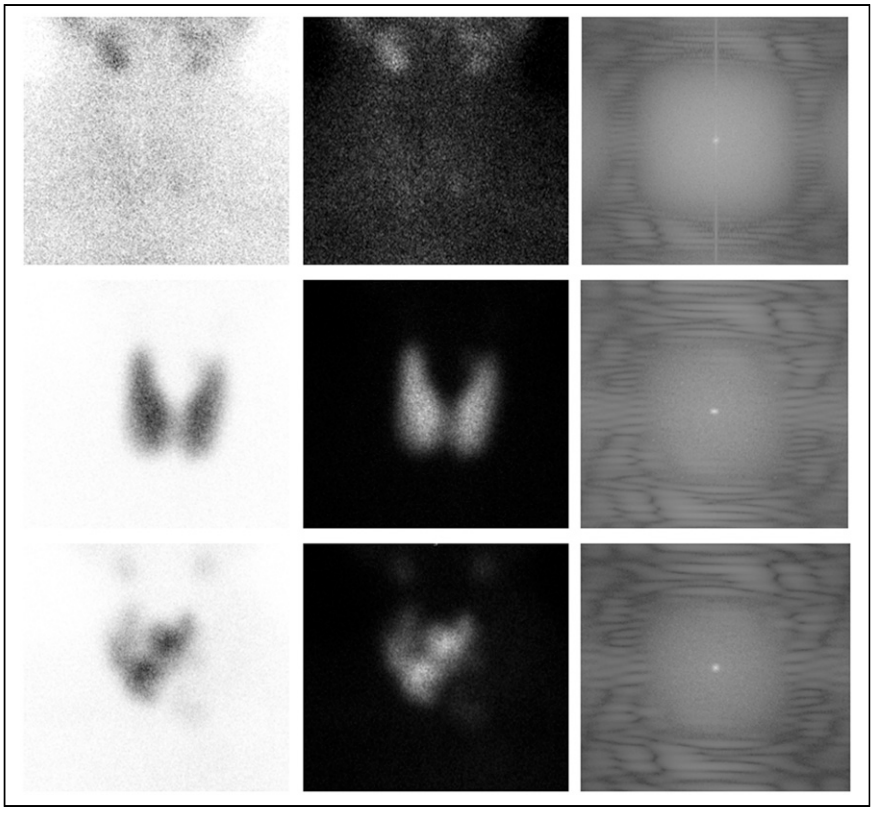

FIGURE 3. Three example patients (top, middle, and bottom) with black on white (left), white on black (center), and magnitude spectrum from Fourier transformation (right) used as inputs for CNN.

representation of hyperthyroidism for men and a lower euthyroid rate. Given the lower representation of men in the thyroid scan population, this observation may reflect lower presentation rates for men in the absence of markedly abnormal thyroid function driving more pressing symptoms. No other statistically significant relationships were noted for patient gender or patient dose (MBq).

There was no statistically significant correlation between thyroid uptake and right-lobe–to–left-lobe ratio ($P = 0.672$), thyroid area ($P = 0.166$), or background counts per pixel (CCP) ($P = 0.416$). The increase in thyroid uptake associated with increasing total counts ($P < 0.001$; $R^2 = 0.458$) and total CPP ($P < 0.001$; $R^2 = 0.356$) was expected. There were also statistically significant relationships between increasing thyroid uptake and increasing thyroid-to-background ratios ($P < 0.001$; $R^2 = 0.376$). The mean thyroid uptake was statistically higher ($P < 0.001$) when the scan showed—relative to appropriately thresholded thyroid activity—no salivary activity (9.1%) than when it showed—relative to faint thyroid activity (2.5%)—salivary activity less than thyroid activity (1.7%), salivary activity equal to thyroid activity (1.1%), or salivary activity greater than thyroid activity (0.4%). A positive correlation between thyroid uptake and both free T4 ($P < 0.001$; $R^2 = 0.351$) and free T3 ($P < 0.001$; $R^2 = 0.365$) was noted; however, no correlation was noted between thyroid uptake and thyroid-stimulating hormone ($P = 0.695$; $R^2 = 0.002$).

Biochemical status demonstrated a statistically significant difference ($P < 0.001$) in mean thyroid uptake stratified as hyperthyroid (9.5%; 95% CI, 7.1%–12.0%), hypothyroid (4.0%; 95% CI, 1.3%–6.7%), and euthyroid (2.5%; 95% CI, 0.9%–4.2%). Hypothyroid studies had a higher mean thyroid uptake than euthyroid studies because of the primary hypothyroidism cases. Excluding primary hypothyroidism, there was no statistically significant difference in thyroid uptake between hypothyroidism and euthyroidism, or between hypothyroidism

TABLE 3
Ternary Classification of Thyroid Function Based on Various Published Reference Ranges

Reference range	Euthyroid	Hyperthyroid	Hypothyroid	Reference
0.45%–4.5%	67.5%	26.8%	7.7%	*6*
0.4%–1.7%	35.0%	61.0%	4.0%	*3*
0.4%–4.0%	65.0%	31.0%	4.0%	*4*
0.3%–3.4%	57.7%	38.2%	4.1%	*2*
0.2%–2.0%	43.1%	52.8%	4.1%	*5*
Biochemical status	53.1%	27.1%	19.8%*	*11*
Salivary classification	44.8%	50.0%	5.2%	—
Physician visual rating	51.0%	43.8%	5.2%	—
Physician rating with uptake value	64.6%	29.2%	6.3%	—

*15.6% were hypothyroid without suppression of uptake (2.1% autonomous, 2.1% secondary hypothyroidism, 11.5% primary hypothyroidism, and 4.2% subclinical hypothyroidism).

TABLE 4
Key Variables

Variable	Mean	95% CI
Total count ratio of right-lobe activity to left-lobe activity	1.5	1.03–2.02
CPP ratio of right-lobe activity to left-lobe activity	1.29	0.98–1.60
Area	33.8 cm^2	31.1–36.5
Size, right	3092 pixels	2,848–3,340
Size, left	2937 pixels	2,662–3,212
Ratio of thyroid to background	4.06	3.43–4.69
Right	4.01 CPP	3.49–4.52
Left	4.08 CPP	3.28–4.89
Ratio of dose to total counts	4.85	3.44–6.26
FT4	21.1 pmol/L	18.1–24.2
FT3	7.1 pmol/L	5.1–9.1
Thyroid-stimulating hormone	4.2 pmol/L	2.3–6.1

and subclinical hyperthyroidism or suppressed hyperthyroidism. Although 4.5% is the cutoff reflecting 100% sensitivity for standard hyperthyroidism, clinical hyperthyroidism with suppression and subclinical hyperthyroidism (both biochemically) are not identified by this reference range.

The optimized cutoff range for thyroid uptake against biochemical status was 0.45%–4.5%, although the lower end of this range is a poor discriminator for hypothyroidism against euthyroidism. For biochemical hyperthyroidism, 70.8% of cases had an uptake greater than 4.5% whereas 29.3% fell below 4.5%. Of those below 4.5%, 100% had biochemically subclinical hyperthyroidism or T3 toxicosis. Of patients with true hyperthyroidism biochemically, 100% had uptake above 4.5%. Conversely, 27.8% of hypothyroidism cases had uptake above 4.5%. There were no hypothyroidism cases that had uptake values below the 0.45% cutoff (all values below this were hyperthyroid or euthyroid biochemically). In the biochemically euthyroid range, only 6% had an uptake above 4.5%, and only 2% had an uptake below 0.45%.

Using the ternary classification, a thyroid uptake above 4.5% had a sensitivity of 70.8% and a specificity of 88.2% for detecting hyperthyroidism. A thyroid uptake below 0.45% had a sensitivity of 0% and specificity of 95.9% for hypothyroidism (Fig. 4, left). A broader biochemical classification of hyperthyroidism saw the sensitivity of the 4.5% cutoff reach 100%, with a specificity of 88.2% (Fig. 4, right).

On the basis of the ternary biochemical status, there was a statistically higher thyroid area for hyperthyroidism (40.7 cm^2) than for hypothyroidism (29.5 cm^2) or euthyroidism (33.0 cm^2) ($P = 0.049$). With reference to Figure 1, the scintigraphic appearance of thyroid activity relative to salivary gland activity correctly identified 70.3% of hyperthyroid studies, 0% of hypothyroid studies, and 62.7% of euthyroid studies (Table 5). Excluding subclinical hyperthyroidism and T3 toxicosis, 94.1% of hyperthyroidism studies were identified using the visual criteria. Table 5 also provides an outline of TP rate (recall) for each set of cutoffs against the biochemical status.

ML

There were 42 input variables in 96 patients (instances) using a binary classification of hyperthyroid or euthyroid. The heat-map correlation matrix identified several redundant

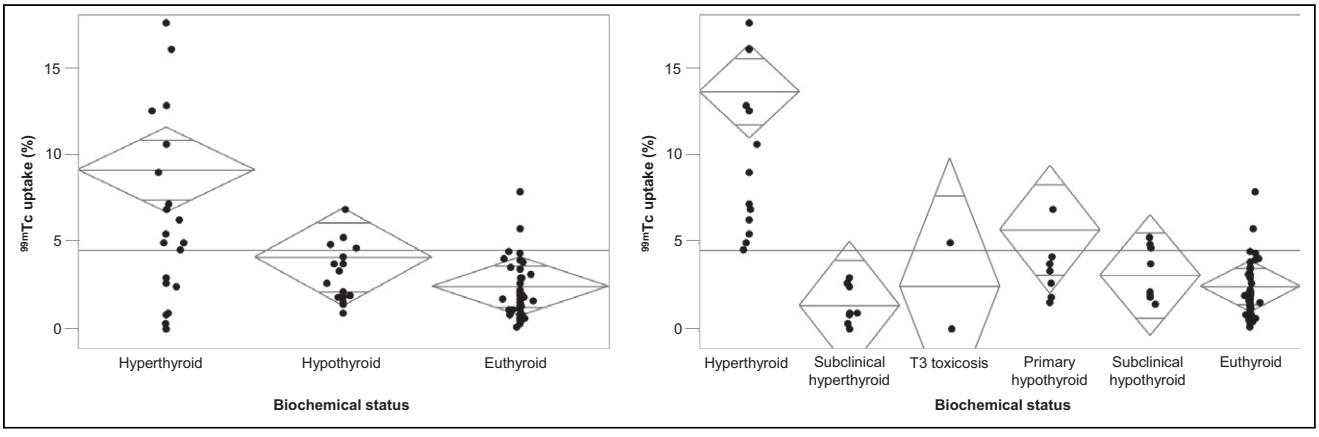

FIGURE 4. (Left) Ternary biochemical status classification against thyroid uptake. (Right) Broader biochemical status classification against thyroid uptake. Horizontal line represents overall mean, and diamonds represent class mean and 95% CIs.

TABLE 5
Ternary Classification of Thyroid Function Based on Recall Against Biochemical Status

Reference range	Euthyroid	Hyperthyroid*	Hypothyroid	Accuracy†
0.45%–4.5%	71.4%	66.6% (100%)	0%	82.6%
0.4%–1.7%	49.0%	74.1% (94.1%)	0%	51.0%
0.4%–4.0%	86.3%	63.0% (94.1%)	0%	77.1%
0.3%–3.4%	74.5%	63.0% (94.1%)	0%	68.8%
0.2%–2.0%	58.8%	74.1% (94.1%)	0%	59.4%
Salivary classification	62.7%	70.3% (94.1%)	0%	61.4%
Physician rating	72.5%	63.0% (89.5%)	0%	61.0%
Physician rating with uptake	88.2%	70.3% (100%)	0%	82.3%

*Data in parentheses exclude subclinical hyperthyroidism and T3 toxicosis.
†Binary accuracy for reference to Table 6.
Accuracy is also provided for binary classification.

variables, and the highest correlation scores were associated with thyroid-stimulating hormone (0.888), appearance of salivary glands on scans (0.627), free T4 (0.575), percentage uptake (0.501), and free T3 (0.491), consistent with the conventional statistical analysis. The network architecture included 16 scaling layer inputs and 3 hidden layers of 6, 4 and 3 nodes. The initial value of the training loss was 1.5473, and the final value after 105 iterations was 0.0172. The initial value of the selection loss was 1.5570, and the final value after 105 iterations was 1.1895.

A growing-inputs method was used to calculate the correlation for every input against each output in the dataset. Beginning with the most highly correlated inputs, progressively decreasing correlated inputs were added to the network until the selection loss increased. The final architecture of the neural network reflected the optimized subset of inputs with the lowest selection loss. In this case, the selection loss and the training loss identified the optimal number of inputs to be 4, with the training loss optimized at 0.0298 and the selection loss being less than 0.0001. The final architecture was 4 scaling-layer inputs; 3 hidden layers of 6, 4, and 1 nodes; an unscaling layer; and a single binary probabilistic layer (Fig. 5).

Several metrics were used to test the final architecture using a subset of the original patient data. Receiver-operator-characteristic analysis demonstrated an area under the curve of 0.933. This correlated with a sensitivity of 100%, a specificity of 80%, and a classification accuracy of 0.846. These results were consistent with scores of 0.60 for precision, 0.75 for F1 score (harmonic mean of sensitivity and precision), 0.693 for the Matthew correlation (correlation between targets and outputs), and 0.8 for the Youden index (probability of a correct decision as opposed to guessing). The cumulative gain analysis demonstrates the benefit of using the developed model over a random guess. On the graph in Figure 6, the positive cumulative gain shows the percentage of positive instances found (y-axis) against the percentage of population (x-axis). Similarly, the negative cumulative gain shows the percentage of negative instances found against the percentage of population. The straight line represents a random classifier. The broader the separation, the better the predictive model. Since the instance ratio provides maximum separation (maximized percentage of positive and negative instances), an instance ratio of 0.40 has a maximum gain score of 0.8. Specifically, but individually, hyperthyroidism is predicted by ^{99m}Tc uptake above 5.7%, free T4 below 20 pmol/L or above 34 pmol/L, free T3 above 9.8 pmol/L, and thyroid-stimulating hormone below 5.5 μIU/mL. In combination, these scaled and weighted input features of the neural network can be expressed mathematically, enhancing the collective predictive capability.

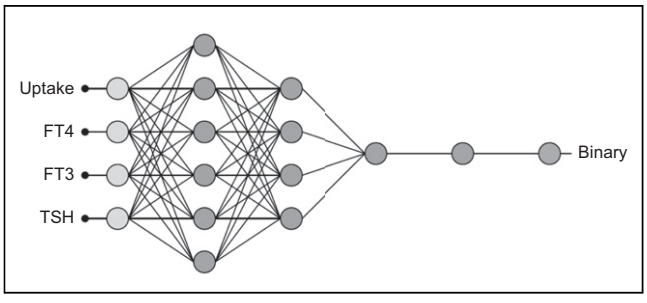

FIGURE 5. Final architecture of trained and validated neural network. TSH = thyroid-stimulating hormone.

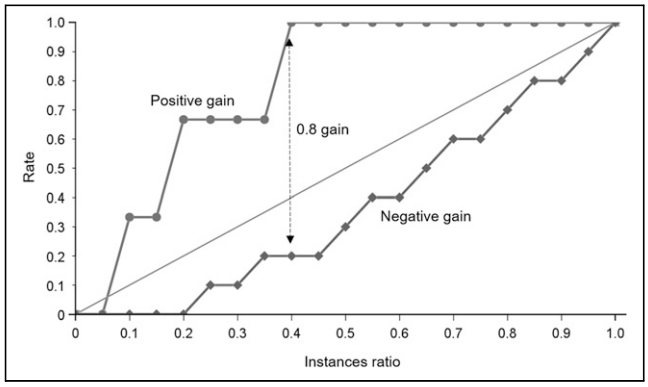

FIGURE 6. Cumulative gain chart demonstrating maximum separation of positive and negative curves to provide cumulative gain score of 0.8 and instances ratio of 0.4 (arrow).

DL

Preliminary network development demonstrated overfitting beyond 30 iterations (epochs); therefore, the maximum epoch number was reset to 30. The results of the triplicated training and validation passes are summarized in Table 6. The variations in validation accuracy reflect the smaller dataset and the random assignment of cases to training and validation. No statistically significant differences (grouped F test) were noted between training and validation accuracy for different types of input tensors ($P = 0.161$ for training accuracy and 0.531 for validation accuracy) despite the higher accuracy for white on black and the lower accuracy for the magnitude spectrum. A direct comparison of white on black against the magnitude spectrum showed P values of 0.068 for training accuracy and 0.280 for validation accuracy.

DISCUSSION

Although thyroid scintigraphy is a well-established technique for the assessment of thyroid function, opinions vary on the role in identifying low versus high thyroid uptake to guide radionuclide therapy. Thyroid scintigraphy is useful in the evaluation of hyperthyroidism to differentiate causes and guide therapy (*14*). Although the specific scintigraphic patterns associated with thyroid pathology do not easily differentiate the biochemical status of the patient (Fig. 7), scintigraphic imaging does provide information useful in identifying patients suitable for radioiodine therapy (*14*). Despite being in widespread use for this purpose internationally, ^{99m}Tc-pertechnetate–based thyroid uptake is not considered suitable in some circles for guiding the therapeutic dosage of radioiodine (*14*). Consistent with the observations of this study, scintigraphy has a limited role in hypothyroidism (*15*).

The challenges and limitations of thyroid scintigraphy are highlighted by poor agreement of physician interpretation. However, with the exclusion of patient history and biochemistry results, the physician interpretation is not done under normal conditions, but for the purpose of this study, the constrained interpretation provides a useful benchmark. Using a thyroid uptake cutoff of 0.45%–4.5%, agreement with physician interpretation was only 63.5%, and using salivary gland appearance, agreement was just 53.1%. Agreement between physicians was not strong, at 59.4%–86.5%, and agreement with biochemistry-grounded truth ranged from 42.7% to 68.8%. This, combined with the poor prediction utility of the salivary gland appearance, contradicts the simplicity of thyroid imaging depicted in Figure 1.

Using the ternary classification of euthyroid, hyperthyroid, and hypothyroid, a thyroid uptake above 4.5% had a sensitivity of 70.8% for detecting hyperthyroidism and a specificity of 88.2%. A thyroid uptake below 0.45% had a sensitivity for hypothyroidism of 0% and a specificity of 95.9%. Specific biochemical classification of hyperthyroidism that excluded T3 toxicosis and subclinical hyperthyroidism improved the sensitivity of the 4.5% cutoff to 100%, with a specificity of 88.2%. This finding highlights the value of thyroid uptake with a cutoff of 4.5% in identifying patients suitable for radioiodine therapy. Given that this goal is the primary one and that scintigraphy has a limited role in hypothyroidism in adults, a binary classification (hyperthyroidism or no hyperthyroidism) provides a more suitable evaluation. The value of an appropriate thyroid uptake cutoff is highlighted in Table 5, which shows that in this

TABLE 6
Triplicate Training and Validation Binary Results (Hyperthyroid or Not Hyperthyroid) for 30-Layer CNN Architecture

Input tensor	Training accuracy	Training loss	Validation accuracy	Validation loss	Mean validation accuracy	Binary accuracy
White on black	82.1%	0.420	75.9%	0.536	80.5%	
	94.0%	0.225	79.3%	0.602		
	91.0%	0.218	86.2%	0.414		
Black on white	83.6%	0.383	82.8%	0.405	78.2%	
	80.6%	0.452	72.4%	0.544		
	91.0%	0.232	79.3%	0.690		
Magnitude spectrum	76.1%	0.459	75.9%	0.530	75.9%	
	74.6%	0.508	72.4%	0.542		
	85.1%	0.306	79.3%	0.380		
Mean	84.2%	0.356	78.2%	0.516		
Initial 25-layer CNN					69.0%	
Conventional metrics						
Normal cutoff, 4.5%						82.6%
Normal cutoff, 4.0%						77.1%
Salivary classification						61.5%
Physician rating						61.0%
Physician rating with uptake						82.3%

Corresponding binary accuracies of best-performing thyroid uptake cutoffs, visual classification against salivary activity relative to thyroid activity, and physician rating are included for comparison.

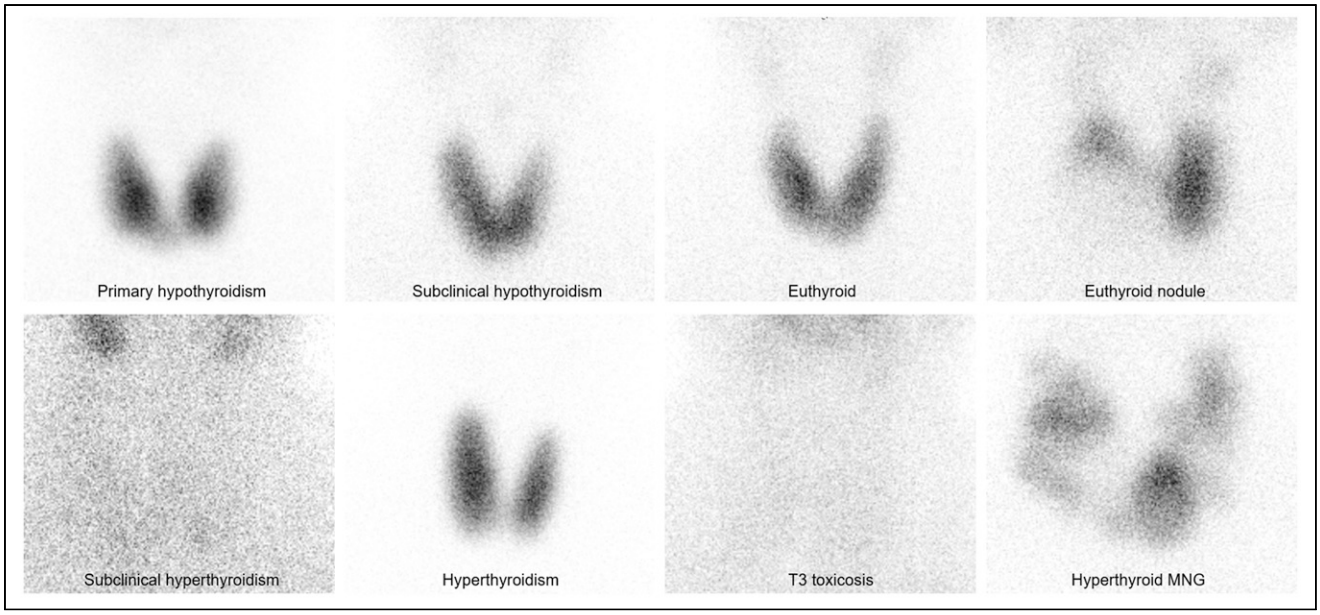

FIGURE 7. Various scintigraphic appearances of thyroid pathology using parallel-hole (high-resolution) collimation and ^{99m}Tc-pertechnetate. MNG = multinodular goiter.

population, binary accuracy was high for a 4.5% cutoff (82.6%) and for physician interpretation augmented by uptake value (82.3%) but was low for salivary gland appearance alone (59.4%) and for masked physician interpretation (61.0%). Indeed, the value and accuracy of 4.5% as the cutoff are reinforced by the similarity in physician interpretation with and without the uptake-augmented information.

Although ML was able to demonstrate improved accuracy to 100%, the algorithm relied on biochemistry not available for physician interpretation. Indeed, the grounded truth relied on the additional value of biochemistry insights to physician insights. In the absence of available biochemistry results, the ML algorithm relies on uptake alone. Conversely, the physician interpretation would improve substantially with the additional insights from biochemistry. In this study, regardless of the apparent performance results, ML augmentation outperformed physician interpretation only because the physician was masked to the biochemistry results available to the ML algorithm. Nonetheless, the role of ML is not and should not be to displace physician reporting but rather to improve accuracy by eliminating error. In this instance, the ML algorithm has been shown to be an accurate second-reader system that can be automated with minimal cost and resources to identify hyperthyroid patients suitable for radioiodine therapy.

In contrast to the success of ML algorithm development, the DL CNN performed more poorly than either the 4.5% cutoff discriminator or the uptake-augmented physician interpretation. The best result was achieved using the white-on-black images (80.5%). Although this result represents only a marginal decrease in performance compared with uptake alone (82.6%) and physician interpretation (82.3%), the CNN was trained on only a single anterior neck image and had no inputs for either the thyroid uptake percentage or the biochemistry results. As a result, the comparative performance should be considered the physician rating without uptake values. In this regard, the 80.5% binary accuracy of the CNN was superior to the physician interpretation (61.0%) and the visual classification against salivary gland appearance (61.5%). Although this result does not suggest displacement of physician interpretation, it does indicate that the accuracy of physician reporting might be improved using the CNN algorithm when biochemistry results are not available.

CONCLUSION

Thyroid scintigraphy is useful in identifying hyperthyroid patients suitable for radioiodine therapy. Physician interpretation relies on an accurate thyroid function assessment (uptake) and an appropriately validated cutoff for the patient population (4.5% in this population). An inappropriate cutoff significantly undermines accuracy. ML ANN algorithms can be developed to improve accuracy as second-reader systems when biochemistry results are available. DL CNN algorithms can be developed to improve accuracy in the absence of biochemistry results. ML and DL do not displace the role of the physician in thyroid scintigraphy but can be used as second-reader systems to minimize errors and increase confidence.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENT

We thank the 3 physicians who performed the visual analysis of the images. We also thank Hugo Currie from the College of Engineering and Computer Science, Australian

National University, Canberra, Australia, for producing the Fourier magnitude spectrum images for analysis.

KEY POINTS

QUESTION: Can ML and DL approaches improve semantic evaluation of thyroid scintigraphy and uptake in hyperthyroidism?

PERTINENT FINDINGS: ML algorithms can be developed to improve accuracy as second-reader systems when biochemistry results are available. DL CNN algorithms can be developed to improve accuracy in the absence of biochemistry results.

IMPLICATIONS FOR PATIENT CARE: ML and DL do not displace the role of the physician in thyroid scintigraphy but can be used as second-reader systems to minimize errors and increase confidence.

REFERENCES

1. Atkins HL, Richards P. Assessment of thyroid function and anatomy with technetium-99m as pertechnetate. *J Nucl Med.* 1968;9:7–15.
2. Maisey MN, Natarajan TK, Hurley PJ, Wagner HN Jr. Validation of a rapid computerized method of measuring ^{99m}Tc pertechnetate uptake for routine assessment of thyroid structure and function. *J Clin Endocrinol Metab.* 1973;36:317–322.
3. Ramos CD, Wittmann DEZ, de Camargo Etchebehere ECS, Tambascia MA, Silva CAM, Camargo EE. Thyroid uptake and scintigraphy using ^{99m}Tc pertechnetate: standardization in normal individuals. *Sao Paulo Med J.* 2002;120:45–48.
4. Hamunyela RH, Kotze T, Philotheou GM. Normal reference values for thyroid uptake of technetium-99m pertechnetate for the Namibian population. *J Endocrin Metab Diabetes S Afr.* 2013;18:142–147.
5. Macauley M, Shawgi M, Ali T, et al. Assessment of normal reference values for thyroid uptake of technetium-99m pertechnetate in a single centre UK population. *Nucl Med Commun.* 2018;39:834–838.
6. Currie G, Dixon C, Vu T. Validation of a normal range for trapping index in thyroid scintigraphy. *ANZ Nucl Med.* 2004;35:11–16.
7. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine learning and deep learning in medical imaging: intelligent imaging. *J Med Imaging Radiat Sci.* 2019;50:477–487.
8. Currie GM. Intelligent imaging: artificial intelligence augmented nuclear medicine. *J Nucl Med Technol.* 2019;47:217–222.
9. Currie G. Intelligent imaging: anatomy of machine learning and deep learning. *J Nucl Med Technol.* 2019;47:273–281.
10. Currie G, Rohren E. Intelligent imaging in nuclear medicine: the principles of artificial intelligence, machine learning and deep learning. *Semin Nucl Med.* 2021;51:102–111.
11. Qiao T, Liu S, Cui Z, et al. Deep learning for intelligent diagnosis in thyroid scintigraphy. *J Int Med Res.* 2021;49:300060520982842.
12. Alswat K, Assiri SA, Althaqafi RMM, et al. Scintigraphy evaluation of hyperthyroidism and its correlation with clinical and biochemical profiles. *BMC Res Notes.* 2020;13:324.
13. Wagieh S, Salman K, Bakhsh A, et al. Retrospective study of Tc-99m thyroid scan in patients with Graves' disease: is there significant difference in lobar activity? *Indian J Nucl Med.* 2020;35:122–129.
14. Mariani G, Tonacchera M, Grosso M, Orsolini F, Vitti P, Strauss HW. The role of nuclear medicine in the clinical management of benign thyroid disorders, part 1: hyperthyroidism. *J Nucl Med.* 2021;62:304–312.
15. Mariani G, Tonacchera M, Grosso M, et al. The role of nuclear medicine in the clinical management of benign thyroid disorders, part 2: nodular goiter, hypothyroidism, and subacute thyroiditis. *J Nucl Med.* 2021;62:886–895.