Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study

Xiangchun Li†, Sheng Zhang*†, Qiang Zhang*, Xi Wei*, Yi Pan, Jing Zhao, Xiaojie Xin, Chunxin Qin, Xiaoqing Wang, Jianxin Li, Fan Yang, Yanhui Zhao, Meng Yang, Qinghua Wang, Zhiming Zheng, Xiangqian Zheng, Xiangming Yang, Christopher T Whitlow, Metin Nafi Gurcan, Lun Zhang, Xudong Wang, Boris C Pasche, Ming Gao, Wei Zhang†, Kexin Chen†*

Summary

Background The incidence of thyroid cancer is rising steadily because of overdiagnosis and overtreatment conferred by widespread use of sensitive imaging techniques for screening. This overall incidence growth is especially driven by increased diagnosis of indolent and well-differentiated papillary subtype and early-stage thyroid cancer, whereas the incidence of advanced-stage thyroid cancer has increased marginally. Thyroid ultrasound is frequently used to diagnose thyroid cancer. The aim of this study was to use deep convolutional neural network (DCNN) models to improve the diagnostic accuracy of thyroid cancer by analysing sonographic imaging data from clinical ultrasounds.

Methods We did a retrospective, multicohort, diagnostic study using ultrasound images sets from three hospitals in China. We developed and trained the DCNN model on the training set, 131 731 ultrasound images from 17 627 patients with thyroid cancer and 180 668 images from 25 325 controls from the thyroid imaging database at Tianjin Cancer Hospital. Clinical diagnosis of the training set was made by 16 radiologists from Tianjin Cancer Hospital. Images from anatomical sites that were judged as not having cancer were excluded from the training set and only individuals with suspected thyroid cancer underwent pathological examination to confirm diagnosis. The model's diagnostic performance was validated in an internal validation set from Tianjin Cancer Hospital (8606 images from 1118 patients) and two external datasets in China (the Integrated Traditional Chinese and Western Medicine Hospital, Jilin, 741 images from 154 patients; and the Weihai Municipal Hospital, Shandong, 11 039 images from 1420 patients). All individuals with suspected thyroid cancer after clinical examination in the validation sets had pathological examination. We also compared the specificity and sensitivity of the DCNN model with the performance of six skilled thyroid ultrasound radiologists on the three validation sets.

Findings Between Jan 1, 2012, and March 28, 2018, ultrasound images for the four study cohorts were obtained. The model achieved high performance in identifying thyroid cancer patients in the validation sets tested, with area under the curve values of 0·947 (95% CI 0·935–0·959) for the Tianjin internal validation set, 0·912 (95% CI 0·865–0·958) for the Jilin external validation set, and 0·908 (95% CI 0·891–0·925) for the Weihai external validation set. The DCNN model also showed improved performance in identifying thyroid cancer patients versus skilled radiologists. For the Tianjin internal validation set, sensitivity was 93·4% (95% CI 89·6–96·1) versus 96·9% (93·9–98·6; p=0·003) and specificity was 86·1% (81·1–90·2) versus 59·4% (53·0–65·6; p<0·0001). For the Jilin external validation set, sensitivity was 84·3% (95% CI 73·6–91·9) versus 92·9% (84·1–97·6; p=0·048) and specificity was 86·9% (95% CI 77·8–93·3) versus 57·1% (45·9–67·9; p<0·0001). For the Weihai external validation set, sensitivity was 84·7% (95% CI 77·0–90·7) versus 89·0% (81·9–94·0; p=0·25) and specificity was 87·8% (95% CI 81·6–92·5) versus 68·6% (60·7–75·8; p<0·0001).

Interpretation The DCNN model showed similar sensitivity and improved specificity in identifying patients with thyroid cancer compared with a group of skilled radiologists. The improved technical performance of the DCNN model warrants further investigation as part of randomised clinical trials.

Funding The Program for Changjiang Scholars and Innovative Research Team in University in China, and National Natural Science Foundation of China.

Introduction

The incidence of thyroid cancer has been increasing worldwide over the past two decades, including in the USA, where a decrease in the incidence of many other cancer types has been reported.[1] Thyroid cancer is three times more prevalent in women than in men[1] and is the most frequently diagnosed type of cancer in women younger than 30 years of age in China.[2] Patients who are suspected of thyroid disease undergo ultrasound imaging, the results of which are interpreted by a radiologist for clinical diagnosis. A key aspect of a radiologist's interpretation of thyroid cancer is recognition of the malignant

Lancet Oncol 2018

Published Online December 21, 2018 http://dx.doi.org/10.1016/S1470-2045(18)30762-9

See Online/Comment http://dx.doi.org/10.1016/S1470-2045(18)30835-0

*Contributed equally and are joint first authors

†Contributed equally and are joint senior authors

Tianjin Cancer Institute (Prof X Li PhD), **Department of Diagnostic and Therapeutic Ultrasonography** (Prof S Zhang MD, X Wei MD, J Zhao MD, X Xin MD, Xi Wang MD, F Yang MD), **Department of Maxillofacial and Otorhinolaryngology Oncology** (Q Zhang MD, Prof L Zhang MD, Xu Wang MD), **Department of Pathology** (Y Pan MD), **Department of Epidemiology and Biostatistics** (M Yang PhD, Q Wang MS, Prof K Chen MD), **and Department of Thyroid and Neck Cancer** (Prof X Zheng MD, Prof M Gao MD), **National Clinical Research Center for Cancer, Key Laboratory of Cancer Prevention and Therapy of Tianjin, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China; Department of Thyroid and Breast Surgery** (C Qin MD) **and Department of Ultrasonography** (J Li MD), **Weihai Municipal Hospital, Shandong, China; Department of Ultrasonography, Affiliated Hospital of Chifeng University, Inner Mongolia, China** (Y Zhao MD)**; Department of Ultrasonography, Integrated Traditional Chinese and Western Medicine Hospital, Jilin, China** (Z Zheng MD)**; Department of Ultrasonography, Dezhou Municiple Hospital, Shandong, China** (X Yang MD)**; and Departments of Radiology and**

Biomedical Engineering (C T Whitlow MD), **Center for Biomedical Informatics Department of Internal Medicine** (Prof M N Gurcan PhD), **and Wake Forest Baptist Comprehensive Cancer Center, Wake Forest Baptist Medical Center, Department of Cancer Biology** (Prof B C Pasche MD, Prof W Zhang PhD), **Wake Forest School of Medicine, Winston-Salem, NC, USA**

Correspondence to: Prof Kexin Chen, Department of Epidemiology and Biostatistics, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300060, China
chenkexin@tjmuch.com

Research in context

Evidence before this study

We searched PubMed on Aug 26, 2018, for research articles that contained the terms "deep learning" OR "convolutional neural network" AND "large scale thyroid imaging data", without date or language restrictions. We found no studies that examined the use of deep learning to improve diagnostic accuracy of thyroid cancer by analysing large-scale sonographic imaging datasets. When we searched PubMed with the terms "deep learning" OR "convolutional neural network" AND "thyroid cancer", we found seven studies that either used deep learning or conventional feature extraction-based machine-learning algorithms to characterise malignancy of thyroid nodules from ultrasonographic images. However, these studies did not include large training datasets (<100 000 images) or external validation sets. The best diagnostic classification method obtained so far was trained with 15 000 images and was not externally validated. Speculatively, the heterogeneity of thyroid nodules was not fully characterised with a limited dataset, and its generalisability remains unknown.

Added value of this study

The high performance of the deep learning model we developed in this study was validated in several cohorts. The improvement in accuracy and specificity seen with this model could lead to a reduction in unnecessary invasive fine-needle aspiration biopsy procedures and overdiagnosis and overtreatment of thyroid cancer. Furthermore, it has the potential to reduce barriers and provide equal access to diagnostic tools for thyroid cancer in regions and countries where medical resources are scarce.

Implications of all the available evidence

The results of our study could improve accuracy, efficiency, and reproducibility of thyroid cancer diagnosis. The artificial intelligence approach proposed could be particularly valuable in community hospitals in which expertise in radiological imaging interpretation is insufficient. Construction of a website running this deep learning framework is ongoing and will be freely available online.

thyroid nodule, according to the Thyroid Imaging, Reporting and Data System (TI-RADS) guidelines. The American College of Radiology (ACR) TI-RADS,[3] European TI-RADS,[4] and American Thyroid Association guidelines[5] propose multiple criteria to interpret sonographic images. Among these criteria, solid aspect, hypoechogenicity, taller-than-wide shape, irregular margin, extrathyroidal extension, calcification, and punctate echogenic foci are clinically relevant features associated with suspicion of malignant disease.[3–8] Patients with suspected thyroid cancer undergo fine-needle aspiration biopsy or surgical resection, which is assessed by pathological examination (the gold standard for diagnosis). Therefore, diagnosis of thyroid cancer is a time-consuming and often subjective process requiring substantial experience and expertise of radiologists.

There are four main subtypes of thyroid cancer: papillary, follicular, medullary, and anaplastic.[7] The 5-year relative survival of patients with thyroid cancer is 99·7%,[1] but this value varies substantially for different subtypes when stratified by stages: near 100% for stage I and II papillary, follicular, and medullary carcinoma; 71% for stage III follicular carcinoma, 81% for stage III medullary carcinoma, and 93% for stage III papillary carcinoma; and 7% for anaplastic, 28% for medullary, 50% for follicular, and 51% for papillary carcinoma at stage IV.[9] All anaplastic thyroid cancers are considered stage IV.[9] In view of the good prognostic outcome of early-stage thyroid cancer, analysis of thyroid ultrasound imaging data by an artificial intelligence algorithm with high performance could help differentiate patients at different risk and avoid unnecessary fine-needle aspiration biopsy or thyroidectomy for those at lower risk, particularly for those patients with papillary carcinomas.

The widespread use of sensitive imaging methods for screening has led to a steady increase in incidence of thyroid cancer, causing overdiagnosis and overtreatment in this setting.[10,11] Indolent and well-differentiated papillary carcinomas and other early-stage thyroid cancers are the main reasons for the growth in incidence, since the incidence of advanced-stage thyroid cancer is rising only marginally. Mortality from thyroid cancer has decreased slightly during the past decade.[10] The frequency of estimated age-standardised thyroidectomy has risen annually by threefold to fourfold in both sexes over the same period.[10] Therefore, development of an artificial intelligence framework based on a precise algorithm with high sensitivity and specificity could maintain a high recall rate for patients with thyroid cancer and identify individuals at low risk for developing advanced disease, thus avoiding unnecessary fine-needle aspiration biopsy. Recently, deep convolutional neural network (DCNN) models have been shown to achieve dermatologist-level classification accuracy in skin cancer diagnosis.[12] Deep learning models have also shown improved performance compared with human experts in detection of diabetic retinopathy and eye-related diseases from raw input pixels of retinal fundus photographs.[13–15]

A traditional machine-learning algorithm for diagnosis of thyroid cancer has been previously developed,[16] but it used as inputs features that were identified explicitly by human experts. Unlike traditional machine learning, deep learning does not require engineered features designed by human experts. Rather, deep learning takes raw image pixels and corresponding class labels from medical imaging data as inputs and automatically learns feature representation with a general manner.[17] Learned representations can be used for classification and object

detection. In this study, we aimed to ascertain the capability of deep learning models for automated diagnosis of thyroid cancer using real-world sonographic data from clinical thyroid ultrasound examinations. We compared results with pathological examination reports (the diagnostic gold standard). This study encompassed model development with a cohort of more than 300 000 images, and validation of the model in three validation datasets.

Methods

Study design and participants

We did a retrospective, multicohort, diagnostic study using ultrasound images sets from three hospitals in China. We obtained ultrasound images for the training set (312 399 images from 42 952 patients) from the thyroid imaging database at Tianjin Cancer Hospital, Tianjin, China. We obtained images for validation sets from thyroid imaging databases at Tianjin Cancer Hospital (internal validation set, 8606 images from 1118 patients), the Integrated Traditional Chinese and Western Medicine Hospital, Jilin, China (Jilin external validation set, 741 images from 154 patients), and Weihai Municipal Hospital, Shandong, China (Weihai external validation set, 11 039 images from 1420 patients).

We included adult patients aged 18 years or older. Clinical diagnosis of the training set was made by 16 radiologists from Tianjin Cancer Hospital, according to TI-RADS guidelines.[3–5] All patients with thyroid cancer and 5651 negative control individuals in the training set, and all individuals in the three validation sets, underwent pathological examination. Pathological examination reports were provided by the pathology department at Tianjin Cancer Hospital. All ultrasound images and pathological examination reports were deidentified before they were transferred to investigators.

This study was approved by the institutional review board (IRB) of Tianjin Cancer Hospital and undertaken according to the Declaration of Helsinki. Informed consent from patients with thyroid cancer and controls was exempted by the IRB because of the retrospective nature of this study.

Procedures

All thyroid ultrasound images extracted from the thyroid imaging database at all three hospital sites were in jpeg format. Ultrasound equipment manufactured by Philips, Toshiba, and GE Healthcare (various models) was used to generate ultrasound images.

Image quality control was performed for the training set; we removed images from thyroid cancer patients if the anatomical sites did not have cancer as per the pathological review report, according to the location sign on the image. For example, if the image available was from the left lobes of the thyroid but pathology data were for the isthmus of the thyroid, the image was considered not suitable for training. For the validation sets, all images were included. Sonographic images with lymph nodes were also included in both training and validation sets.

	Training set* (n=42 952)	Tianjin internal validation set (n=1118)	Jilin external validation set (n=154)	Weihai external validation set (n=1420)
Inpatients with thyroid cancer	17 627 (41%)	563 (50%)	70 (45%)	542 (38%)
Images	131 731	4491	347	4818
Control inpatients	5651 (13%)	555 (50%)	84 (55%)	878 (62%)
Images	51 255	4115	394	6221
Control outpatients†	19 674 (46%)	0	0	0
Images	129 413	0	0	0
Male sex	10 832 (25%)	261 (23%)	34 (22%)	282 (20%)
Images	78 768	1785	154	1992
Female sex	32 032 (75%)	866 (77%)	120 (78%)	1138 (80%)
Images	233 268	6830	587	9047
Age (years)	44 (36–54)	47 (24–41)	51 (45–59)	50 (41–59)
Age ≤30 years male	2009 (5%)	112 (10%)	1 (<1%)	24 (2%)
Age >30 years male	8823 (21%)	146 (13%)	33 (21%)	258 (18%)
Age ≤30 years female	5830 (14%)	381 (34%)	5 (3%)	76 (5%)
Age >30 years female	26 202 (61%)	479 (43%)	115 (75%)	1062 (75%)

Data are n, n (%) or median (IQR). *No information on sex was available for 88 individuals in the training set (corresponding to 363 images). †These individuals did not have any malignant characteristics, decided by doctors at clinical examination.

Table 1: **Baseline characteristics**

A DCNN classification model, in which image input features (eg, image pixels) are mapped to the corresponding output label (eg, benignity or malignancy), was used to train the deep learning algorithm. The DCNN algorithm can learn hierarchical representations from the input imaging data. Such a trained model can make predictions on input data. We used the ResNet model[18] with 50 layers (ResNet-50) and the Darknet model[19] with 19 layers (Darknet-19) for image classification. Layers are functional units of neural network and can have different functions in that they learn and store abstract features of the input image. The ResNet-50 and Darknet-19 models were first trained iteratively for classification of patients with thyroid cancer (using 131731 images) and controls (using 180 668 images). We next combined these two deep learning models by weighting their performance (measured by area under the curve [AUC]) and assessed the ensemble DCNN model with the internal and external validation sets.

Darknet-19 was proposed as the backbone for the object detection algorithm[19] because it is more computationally efficient than ResNet-50 (in that Darknet-19 has fewer arithmetic operations compared with ResNet-50) and achieved performance metrics with ImageNet data[19] that were comparable with those obtained with ResNet-50 (appendix p 4). The weights of ResNet-50 and Darknet-19 were initialised from the same network that had been trained to classify 1000 objects in the ImageNet dataset,[20] except the last layer. The weights of last layer were

See Online for appendix

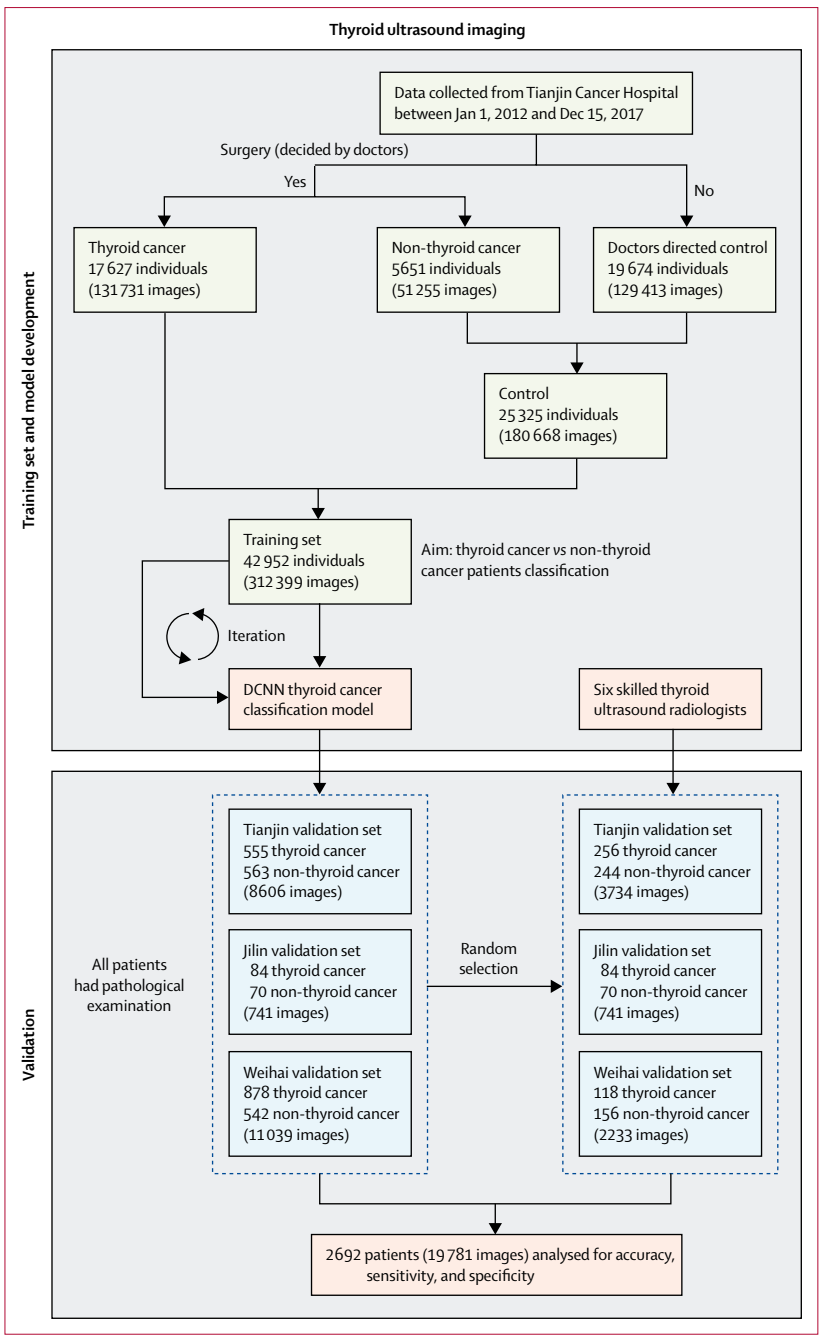

Figure 1: **Flowchart of the procedures in the development and evaluation of deep learning models for automated thyroid cancer diagnosis**
Controls were patients with thyroid diseases that were negative on pathological examination or that did not show any evidence associated with thyroid malignancy in clinical testing, as determined by doctors. DCNN=deep convolutional neural network.

For more on **R** see www.R-project.org

randomly initialised and the output unit was changed to two for matching the number of classes in our study (ie, thyroid cancer *vs* control). We trained the network with stochastic gradient descent running on an NVIDIA graphic processing unit (GPU) with a GTX 1080Ti graphics card (NVIDIA, Beijing, China). We also applied on-the-fly data augmentation[12,21] for each image during training to avoid overfitting. On-the-fly augmentation generates more training images through image processing such as random cropping, rotation, horizontal or vertical flipping, scaling, translations, and adjustment of the saturation and exposure, which mimic the data diversity observed in the real world, avoiding model overfitting. Image augmentation was not done for the validation sets. Additionally, a weight decay rate of 0·0005 was also set to additionally combat for overfitting. Weight decay can prevent the weights of neural network from growing too large.

To quantify the contribution of the pixels that most influence the DCNN model's prediction, we generated a class activation map[22] by using global average pooling in the ResNet model (appendix p 4).

To derive individual-level prediction scores, we denoted n as the total number of images available from that patient and let $P_{cancer}=[P_1, P_2, \ldots, P_n]$ denote the predicted probabilities for these n images that were classified as cancer. The score θ assigned to an individual was defined as the average value of log-transformed P_{cancer}.

$$\theta = -[\ln(1-P_1) + \ln(1-P_2) + \ldots + \ln(1-P_n)] / n$$

The prediction scores obtained from ResNet-50 and Darknet-19 were combined, which is weighted by their performance—ie, the area under the receiver operating characteristic (ROC) curve (AUC) value of ResNet-50 ($AUC_{ResNet\text{-}50}$) and AUC value of Darknet-19 ($AUC_{Darknet\text{-}19}$).

$$\theta_{combined} = w_1 \times \theta_{ResNet\text{-}50} + w_2 \times \theta_{Darknet\text{-}19}$$

Here, $w_1 = AUC_{ResNet\text{-}50} / (AUC_{ResNet\text{-}50} + AUC_{Darknet\text{-}19})$ and $w_2 = 1 \cdot 0 - w_1$

We compared the performance of the deep learning model predictions for thyroid cancer diagnosis with those of six skilled thyroid ultrasound radiologists (XiWe, XX, XiWa, FY, JZ, and SZ) with at least 6 years' experience each. We asked the radiologists to read and interpret subsets of thyroid ultrasound imaging data randomly selected from validation sets. We made the random selections using the random sampling function implemented in R software (*sample*). The entire image subset for selected patients was shown to the radiologists, who interpreted the images according to the guidelines of ACR TI-RADS. Each radiologist read image subsets from two validation sets. The performance of the radiologists was assessed by comparing their predictions with pathological reports (which are the diagnostic gold standard).

Prediction scores derived from DCNN models were compared with pathological examination reports of formalin-fixed and paraffin-embedded samples of suspected cancers removed surgically, which is the gold standard for diagnosis. Pathological examination was done to confirm diagnosis for all individuals in the training set

with thyroid nodules displaying malignant characteristics at clinical examination (17 627 [76%] of 23 278 individuals). The remaining 19 674 individuals were used as negative controls. All individuals in the three validation sets had pathological examination results. Pathological assessment was done by board-certified pathologists at individual sites according to WHO Classification of Tumors of Endocrine Organs. All pathological assessments were based on haematoxylin and eosin-stained whole-slide images.

Statistical analysis

For classification purposes, we used the ROC curve to show the diagnostic ability of the deep learning model in discriminating thyroid cancer patients from controls. The ROC curve was created by plotting the true positive rate (sensitivity) against the false positive rate (1–sensitivity), by varying the predicted probability threshold, and we calculated AUC values. We calculated 95% CIs for sensitivity and specificity with the Clopper-Pearson method.[23] Sensitivity was calculated as the fraction of patients with cancer who were correctly identified, and specificity was calculated as the fraction of patients without thyroid cancer who were correctly identified. We calculated AUC values, accuracy, sensitivity, and specificity using R software *caret* (version 6.0-78) and *GenBinomApps* (version 1.0-2). The ROC curve was plotted by R software *pROC* (version 1.10.0).

We also calculated likelihood ratios for positive and negative results. We calculated the likelihood ratio for positive results as sensitivity divided by 1–specificity and the likelihood ratio for negative results as 1–sensitivity divided by specificity. The confusion matrix in our study is a 2×2 contingency table that reports the number of true positives, false positives, false negatives, and true negatives. We used the average accuracy, sensitivity, and specificity of the radiologists when comparing performance between the deep learning model and the radiologists. The inter-radiologist agreement rate and Fleiss' kappa value[24] were calculated for each validation set using R software *irr* (version 0.84). We used the binomial test to statistically evaluate the difference in accuracy, sensitivity, and specificity between the deep learning model and the radiologists. Statistical analyses were done with R software (version 3.4.3).

Role of the funding source

The funder 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

Between Jan 1, 2012, and Dec 15, 2017, 396 998 ultrasound images were obtained for the training set from the Thyroid Imaging Database in Tianjin Cancer Hospital. After quality control evaluation, 84 599 (21%) images that did not match with pathological reports in terms of anatomical locations were removed from this set. The complete training set consisted of 312 399 images from 42 952 individuals: 17 627 patients with thyroid cancer (131 731 images) and 25 325 controls (180 668 images).

	Tianjin cohort (n=1118)	Jilin cohort (n=154)	Weihai cohort (n=1420)
Accuracy (95% CI)	0·889 (0·869–0·907)	0·857 (0·792–0·908)	0·863 (0·844–0·880)
Sensitivity (95% CI)	0·922 (0·897–0·943)	0·843 (0·736–0·919)	0·849 (0·816–0·878)
Specificity (95% CI)	0·856 (0·824–0·884)	0·869 (0·778–0·933)	0·871 (0·847–0·893)
Positive predictive value	0·866	0·843	0·803
Negative predictive value	0·915	0·869	0·903
Kappa*	0·778	0·712	0·712
F_1†	0·893	0·843	0·825

DCNN=deep convolutional neural network. *Measures the agreement between the DCNN model prediction and the pathological report. †Measures the accuracy of the DCNN model prediction against the pathological report.

Table 2: **Performance metrics for the ensemble DCNN model, assessed on the validation sets**

Between Jan 1, 2018, and Mar 28, 2018, 8606 images from 1118 individuals for the internal validation set were obtained from Tianjin Cancer Hospital. Between Apr 1, 2016, and Feb 28, 2018, 741 images from 154 individuals for the first external validation set were obtained from Integrated Traditional Chinese and Western Medicine Hospital (Jilin set). Between Jan 1, 2016, and Dec 29, 2017, 11 039 images from 1420 individuals for the second external validation set were obtained from Weihai Municipal Hospital (Weihai set). Baseline characteristics of the training set and three validation sets are shown in table 1. Clinicopathological information related to tumour subtype and tumour size are provided in the appendix (p 1).

A flowchart depicting processes during the study is shown in figure 1. The model achieved high performance in identifying thyroid cancer patients in the validation sets tested (table 2), with AUC values of 0·947 (95% CI 0·935–0·959) for the Tianjin internal validation set, 0·912 (0·865–0·958) for the Jilin external validation set, and 0·908 (0·891–0·925) for the Weihai external validation set (figure 2). Likelihood ratios for positive and negative diagnostic results were, respectively, 6·40 (95% CI 5·27–7·96) and 0·09 (0·07–0·12) for the internal validation set, 6·43 (3·92–12·77) and 0·18 (0·09–0·29) for the Jilin set, and 6·74 (5·68–8·14) and 0·18 (0·14–0·21) for the Weihai set. The appendix (p 1) shows exemplified class activation maps that identify the pixels on which the ResNet-50 model was fixating its attention for prediction. Confusion matrices reporting the number of true-positive, false-positive, false-negative, and true-negative results for ResNet-50, Darknet-19, and the ensemble DCNN model are shown in the appendix (p 2).

500 (45%) of 1118 individuals from the Tianjin internal validation set (3734 [43%] of 8606 images), 274 (19%) of 1420 individuals from the Weihai external validation set (2233 [16%] of 13 949 images), and all 154 (100%) individuals from the Jilin external validation set (all 741 images) were selected, and these images were

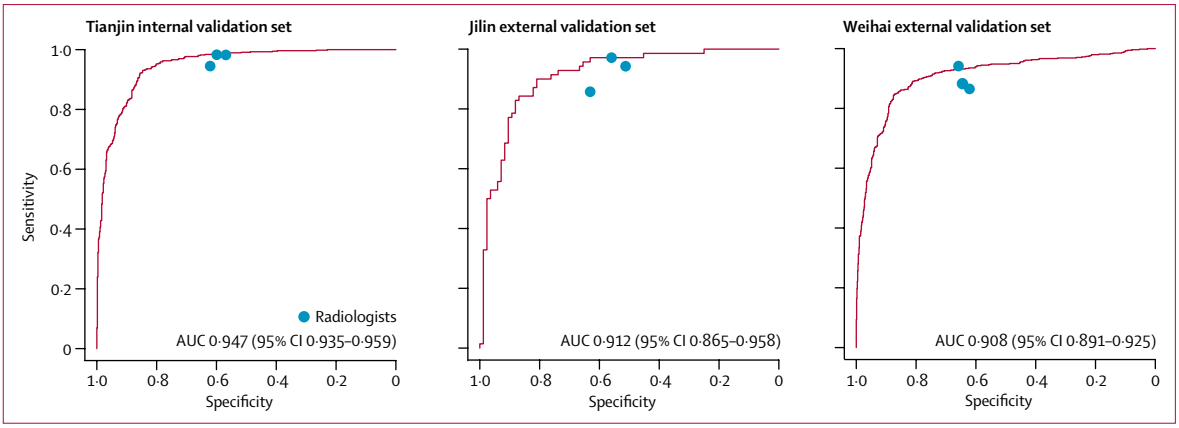

Figure 2: **Performance of the ensemble DCNN model in identifying patients with thyroid cancer on three validation sets**
The blue dots on each ROC curve indicate the performance of the radiologists. AUC=area under the curve. DCNN=deep convolutional neural network. ROC=receiver operating characteristics curve.

used to assess the performance of the ensemble DCNN model versus the group of six skilled thyroid ultrasound radiologists (table 3). Radiologist 1 read 4483 images (n=654 individuals), radiologist 2 read 5967 images (n=774), radiologist 3 read 3734 images (n=500), radiologist 4 read 2982 images (n=482), radiologist 5 read 741 images (n=154), and radiologist 6 read 2233 images (n=274). The entire image set for every selected patient was shown to and read by the radiologists. Radiologists' manual interpretation results were aggregated and the classification accuracy, sensitivity, and specificity were calculated and compared with that of deep learning models.

Among the radiologists, for the Tianjin internal validation set, accuracy ranged from 78·0% (95% CI 74·1–81·6; 390 of 500 individuals) to 79·6% (75·8–83·0; 398 of 500 individuals), sensitivity ranged from 94·1% (90·5–96·7; 241 of 256 individuals) to 98·4% (96·0–99·6; 252 of 256 individuals), and specificity from 57·0% (50·5–63·3; 139 of 244 individuals) to 62·3% (55·9–68·4; 152 of 244 individuals). For the Jilin external validation set, accuracy ranged from 70·8% (95% CI 62·9–77·8; 109 of 154 individuals) to 74·7% (67·0–81·3; 115 of 154 individuals), sensitivity from 85·7% (75·3–92·9; 60 of 70 individuals) to 97·1% (90·1–99·7; 68 of 70 individuals), and specificity from 51·2% (40·0–62·3; 43 of 84 individuals) to 63·1% (51·9–73·4; 53 of 84 individuals). For the Weihai external validation set, accuracy ranged from 72·6% (66·9–77·8; 199 of 274 individuals) to 81·8% (76·7–86·1; 223 of 274 individuals), sensitivity from 85·6% (77·9–91·4; 101 of 118 individuals) to 94·1% (88·2–97·6; 111 of 118 individuals), and specificity from 62·2% (54·1–69·8; 97 of 156 individuals) to 78·8% (71·6–85·0; 123 of 156 individuals). The inter-radiologist agreement rate was 86·4% (95% CI 83·1–89·3; 432 of 500 individuals; Fleiss' Kappa 0·79) in the Tianjin internal validation set, 76·6% (69·1–83·1; 118 of 154 individuals; Fleiss' Kappa 0·65) in the Jilin external validation set, and 69·7% (63·9–75·1; 191 of 274 individuals; Fleiss' Kappa 0·59) in the Weihai external validation set.

Compared with the skilled radiologists, the ensemble DCNN model achieved high performance in identifying thyroid cancer patients. For the Tianjin internal validation set, accuracy was 89·8% (95% CI 86·8–92·3; 994 of 1118 individuals) with the DCNN model versus 78·8% (75·0–82·3; 394 of 500 individuals; p<0·0001) with the radiologists, sensitivity was 93·4% (95% CI 89·6–96·1; 519 of 563 individuals) versus 96·9% (93·9–98·6; 248 of 256 individuals; p=0·003), and specificity was 86·1% (95% CI 81·1–90·2; 475 of 555 individuals) versus 59·4% (53·0–65·6; 145 of 244 individuals; p<0·0001). For the Jilin external validation set, accuracy was 85·7% (95% CI 79·2–90·8; 132 of 154 individuals) versus 72·7% (65·0–79·6%; 112 of 154 individuals; p<0·0001), sensitivity was 84·3% (95% CI 73·6–91·9%; 59 of 70 individuals) versus 92·9% (84·1–97·6; 65 of 70 individuals; p=0·048), and specificity was 86·9% (95% CI 77·8–93·3; 73 of 84 individuals) versus 57·1% (45·9–67·9%; 48 of 84 individuals; p<0·0001). For the Weihai external validation set, accuracy was 86·5% (95% CI 81·9–90·3; 1225 of 1420 individuals) versus 77·4% (72·0–82·2; 212 of 274 individuals; p<0·0001), sensitivity was 84·7% (95% CI 77·0–90·7; 460 of 542 individuals) versus 89·0% (81·9–94·0%; 105 of 118 individuals; p=0·25), and specificity was 87·8% (95% CI 81·6–92·5; 765 of 878 individuals) versus 68·6% (60·7–75·8; 107 of 156 individuals; p<0·0001). At the same specificity as the group of radiologists, the ensemble DCNN model had higher or at least comparable sensitivity and specificity across these three validation sets (figure 2, table 3). Additionally, the ensemble DCNN model had higher kappa coefficient, positive predictive value, and F_1 score compared with the performance of the radiologists (table 3). Classification confusion matrices reporting the number of true-positive, false-positive, false-negative, and true-negative results achieved by the group of skilled ultrasound radiologists, the ResNet-50 model, the

Darknet-19 model, and the ensemble DCNN model are provided in the appendix (pp 3, 4).

Discussion

The findings of our retrospective study show that our DCNN model tested in three validation sets can achieve high accuracy, sensitivity, and specificity in automated thyroid cancer diagnosis in a real-world setting. The developed artificial intelligence system had significantly higher accuracy and specificity in classifying thyroid cancer patients compared with a group of skilled radiologists. The thyroid ultrasound images used in our study were produced by several different types of ultrasound equipment, which contributed to increased data diversity to train the algorithm and test interpretation subjectivity from radiologists.

Thyroid cancer diagnosis requires accurate recognition of malignant thyroid nodules. However, thyroid nodules are characterised by heterogeneous appearances and vague boundaries, leading to difficulties in accurate recognition and consistent interpretation of malignant nodules by radiologists, as shown by varying agreement rates between radiologists in the validation sets. Deep learning has advantages in overcoming the problem of heterogeneity, because feature representation learned from thyroid ultrasound images is not limited by engineered features used by radiologists. Instead, the DCNN model learned feature representations with an automated procedure. Interpretation of thyroid cancer by deep learning maintains consistency and, therefore, diagnostic reproducibility. Another benefit offered by our artificial intelligence system is that it could report results instantly on a graphical processing unit, and integration of the system into ultrasound equipment could help radiologists accelerate the interpretation process. Integration of this system into a portable ultrasound machine could enable flexible monitoring of disease development and progression and, thus, augment the capability of radiologists to manage individuals who are at high risk of thyroid cancer. Conferred by the high speed of a GPU, the developed DCNN model has the advantage to assess all images of a lesion, whereas a radiologist sometimes cannot do so because ultrasound image interpretation is labour-intensive. Implementation of the DCNN model could lead to a reduction in overdiagnosis and overtreatment related to thyroid cancer. However, the applicability of this proposed integration system needs to be tested in prospective clinical studies.

To the best of our knowledge, our study included the largest number of images so far for development and validation of a deep learning model. All patients in the validation sets underwent thyroid surgery and pathological examination, whereas some controls in the training set did not have surgery or a pathology report. The performance of the deep learning model is presumably lower in the validation sets because they were enriched for nodules with more typical features of

	Tianjin cohort (n=500)				Jilin cohort (n=154)				Weihai cohort (n=274)			
	Radiologist 1	Radiologist 2	Radiologist 3	DCNN model	Radiologist 1	Radiologist 4	Radiologist 5	DCNN model	Radiologist 2	Radiologist 4	Radiologist 6	DCNN model
Accuracy (95% CI)	0·786 (0·747–0·821)	0·780 (0·741–0·816)	0·796 (0·758–0·830)	0·898 (0·868–0·923)	0·747 (0·670–0·813)	0·734 (0·657–0·802)	0·708 (0·629–0·778)	0·857 (0·792–0·908)	0·818 (0·767–0·861)	0·726 (0·669–0·778)	0·777 (0·723–0·825)	0·865 (0·819–0·903)
Sensitivity (95% CI)	0·941 (0·905–0·967)	0·980 (0·955–0·994)	0·984 (0·960–0·996)	0·934 (0·896–0·961)	0·971 (0·901–0·997)	0·857 (0·753–0·929)	0·943 (0·860–0·984)	0·843 (0·736–0·919)	0·856 (0·779–0·914)	0·864 (0·789–0·920)	0·941 (0·882–0·976)	0·847 (0·770–0·907)
Specificity (95% CI)	0·623 (0·559–0·684)	0·570 (0·505–0·633)	0·598 (0·534–0·660)	0·861 (0·811–0·902)	0·560 (0·447–0·668)	0·631 (0·519–0·734)	0·512 (0·400–0·623)	0·869 (0·778–0·933)	0·788 (0·716–0·850)	0·622 (0·541–0·698)	0·654 (0·57×8)	0·878 (0·816–0·925)
Positive predictive value	0·724	0·705	0·720	0·875	0·648	0·659	0·617	0·843	0·754	0·634	0·673	0·840
Negative predictive value	0·910	0·965	0·973	0·925	0·959	0·841	0·915	0·869	0·879	0·858	0·936	0·884
Kappa*	0·569	0·555	0·588	0·796	0·510	0·476	0·436	0·712	0·634	0·466	0·567	0·725
F_1†	0·818	0·820	0·832	0·904	0·777	0·745	0·746	0·843	0·802	0·731	0·784	0·844

Patients from all validation sets were selected randomly for analysis. DCNN=deep convolutional neural network. *Measures the agreement between the DCNN model prediction and the pathological report. †Measures the accuracy of the DCNN model prediction against the pathological report.

Table 3: **Performance metrics for the DCNN model versus skilled radiologists, assessed on image sets from selected individuals from the validation sets**

malignancy and, thus, were more difficult to differentiate. The improvement in accuracy and specificity reported with the DCNN model might lead to a reduction in unnecessary fine-needle aspiration biopsy procedures. However, clinical diagnostic validity needs to be assessed in future randomised clinical trials against current standard procedures.

The trained DCNN model could correctly pinpoint malignant thyroid nodules in a weakly supervised manner through class activation map analysis. DCNN models and machine learning approaches based on conventional feature extraction have previously been investigated for discrimination of malignancy of thyroid nodules from ultrasound images. For example, Ma and colleagues[25] used DCNN and analysed 8148 manually annotated thyroid nodules and obtained an accuracy of 83·0% (95% CI 82·3–83·7) in thyroid nodule diagnosis; however, data from this study are not available so we could not assess them with our artificial intelligence system. Xia and colleagues[26] achieved an accuracy of 87·7% in differentiating malignant and benign nodules by applying extreme machine learning to radiologist-collected features that were obtained from 203 ultrasound images of 187 patients with thyroid cancer. Pereira and colleagues[27] reported an accuracy of 83% achieved by a DCNN model in distinguishing between malignant and benign thyroid nodules from 946 images of 165 patients, which was substantially higher than machine learning algorithms based on conventional feature extraction. However, these studies were limited by small sample sizes and no external validation sets. We do not know if the improvement in accuracy we reported in our study relates to the machine learning method used or to the much larger training dataset.

The **website can be accessed** at http://lixiangchun.github.io/

Our study has some limitations. We did not include training data from other hospitals, and we did not do sensitivity analyses with respect to tumour size and subtypes of malignant disease. 5651 (13%) of 42 952 individuals in the training set were true negatives, with the assumption that patients who did not undergo surgery would be mainly negative diagnoses. The performance of our artificial intelligence system is expected to increase by including more data and expanding the sets to real-world data from other hospitals. Other limitations were that a TI-RADS score of 5 was the only condition to score nodules as malignant, and that in contrast to the algorithm, radiologists in our study did not analyse lymph node images to support their diagnosis. In daily practice, a radiologist reviews approximately 300 images (from about 30 individuals) under time constraints. In our study, radiologists were asked to review images without time constraints; thus, the specificity of this group of skilled radiologists is expected to decrease in daily practice. The features of benign nodules or normal thyroid are less heterogeneous than are those of malignant nodules. Although thyroid cancer subtypes with low incidence—such as follicular thyroid cancer—were not well represented in our training set, the hierarchical features learned from papillary carcinoma should be generalisable to other subtypes since features of thyroid nodules from images of papillary carcinoma are shared with those from follicular carcinoma. Because the algorithm was trained only with images from anatomical sites that did have cancer, and the probability of cancer was calculated by averaging logarithmic transformation of one minus probabilities from each image, the algorithm could report a lower score in a clinical trial, when non-cancer site images would not be removed, leading to decreased sensitivity.

Factors that limit generalisability of the DCNN model relate mainly to an absence of multicentre training cohorts and removal of images from anatomical sites of cancer patients who have no tumours. Additionally, most patients in the cohorts are northern Han Chinese. Future multicentre investigations should mitigate this limiting factor and improve generalisability. The current artificial intelligence system was not able to account for other clinical parameters; therefore, it cannot replace manual diagnosis of thyroid cancer but could augment the ability of thyroid ultrasound radiologists in thyroid cancer diagnosis.

We are building a website to provide free access to the developed DCNN model. In our future work, we intend to link hierarchical features of thyroid ultrasound images learned by DCNN models to features of thyroid nodules that are mostly used by radiologists in interpreting thyroid cancer. Medical resources in urban and rural areas of China—and in many other countries in the world—are unbalanced; the artificial intelligence system developed in our study could contribute to reducing barriers and providing a convenient way for community hospitals to improve thyroid cancer diagnosis.

The newly developed DCNN model showed improved accuracy, sensitivity, and specificity in identifying patients with thyroid cancer at levels similar to or higher than a group of skilled radiologists. The improved technical performance obtained by the DCNN model indicates that this method is valuable to proceed with and to be tested in prospective clinical trials.

Contributors

XL did the data analysis. KC, WZ, SZ, and MG supervised the project. XL, QZ, WZ, SZ, and KC designed the experiment. XL, QZ, XiWe, WZ, and KC wrote the report. CTW, MNG, and BCP edited the report. YP curated pathological examinations. JZ, XX, XiWa, XuWa, FY, MY, QW, LZ, ZZ, YZ, XZ, and XY gathered and annotated data. XiWe, SZ, CQ, JL, YZ, and XY interpreted the validation set.

Declaration of interests

We declare no competing interests.

Acknowledgments

This study was supported by the Program for Changjiang Scholars and Innovative Research Team in University in China (grant IRT_14R40, to KC), and the National Natural Science Foundation of China (grant 31801117 to XL). WZ is supported by a fellowship from the National Foundation for Cancer Research and a Hanes and Wills Family endowed professorship in cancer at the Wake Forest Baptist Comprehensive Cancer Center. BCP and WZ are supported by the Cancer Center

Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197).

References

1 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin* 2017; **67:** 7–30.

2 Chen W, Zheng R, Baade PD, Zhang S, Zeng H. Cancer statistics in China, 2015. *CA Cancer J Clin* 2016; **66:** 115–32.

3 Tessler FN, Middleton WD, Grant EG, et al. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): White Paper of the ACR TI-RADS Committee. *J Am Coll Radiol* 2017; **14:** 587–95.

4 Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. *Eur Thyroid J* 2017; **6:** 225–37.

5 Haugen BR, Alexander EK, Bible KC, et al. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* 2016; **26:** 1–133.

6 Li Q, Lin X, Shao Y, Xiang F, Samir AE. Imaging and screening of thyroid cancer. *Radiol Clin North Am* 2017; **55:** 1261–71.

7 Tamhane S, Gharib H. Thyroid nodule update on diagnosis and management. *Clin Diabetes Endocrinol* 2016; **2:** 17.

8 Durante C, Grani G, Lamartina L, Filetti S, Mandel SJ, Cooper DS. The diagnosis and management of thyroid nodules: a review. *JAMA* 2018; **319:** 914–24.

9 American Cancer Society. Thyroid cancer survival rates, by type and stage. April 15, 2016. https://www.cancer.org/cancer/thyroid-cancer/detection-diagnosis-staging/survival-rates.html (accessed Nov 29, 2018).

10 Jegerlehner S, Bulliard J-L, Aujesky D, et al. Overdiagnosis and overtreatment of thyroid cancer: a population-based temporal trend study. *PLoS One* 2017; **12:** e0179387.

11 Park S, Oh C-M, Cho H, et al. Association between screening and the thyroid cancer "epidemic" in South Korea: evidence from a nationwide study. *BMJ* 2016; **355:** i5745.

12 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542:** 115–18.

13 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

14 Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017; **318:** 2211–23.

15 Kermany DS, Goldbaum M, Cai W, Lewis MA. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018; **172:** 1122–31.

16 Chang Y, Paul AK, Kim N, et al. Computer-aided diagnosis for classifying benign versus malignant thyroid nodules based on ultrasound images: a comparison with radiologist-based assessments. *Med Phys* 2016; **43:** 554–67.

17 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.

18 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Conf Comput Vis Pattern Recognit* 2016; published online Dec 12. DOI:10.1109/CVPR.2016.90.

19 Redmon J, Farhadi A. YOLO9000: better, faster, stronger. *Proc IEEE Conf Comput Vis Pattern Recognit* 2017; published online Nov 9. DOI:10.1109/CVPR.2017.690.

20 Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis* 2015; **115:** 211–52.

21 Chang K, Bai HX, Zhou H, et al. Residual convolutional neural network for determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin Cancer Res* 2018; **24:** 1073–81.

22 Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. *Proc IEEE Conf Comput Vis Pattern Recognit* 2016; published online Dec 12. DOI:10.1109/CVPR.2016.319.

23 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26:** 404–13.

24 Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; **76:** 378–82.

25 Ma J, Wu F, Zhu J, Xu D, Kong D. A pre-trained convolutional neural network based method for thyroid nodule diagnosis. *Ultrasonics* 2017; **73:** 221–30.

26 Xia J, Chen H, Li Q, et al. Ultrasound-based differentiation of malignant and benign thyroid nodules: an extreme learning machine approach. *Comput Methods Programs Biomed* 2017; **147:** 37–49.

27 Pereira C, Dighe M, Alessio AM. Comparison of machine learned approaches for thyroid nodule characterization from shear wave elastography images. *Proc SPIE Med Imaging Comput Aided Diagn* 2018; published online Feb 27. DOI:10.1117/12.2294572.