European Journal of Radiology 98 (2018) 14–19

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Research Paper

Inter-exam agreement and diagnostic performance of the Korean thyroid imaging reporting and data system for thyroid nodule assessment: Real-time versus static ultrasonography

Jung Min Bae, Soo Yeon Hahn*, Jung Hee Shin, Eun Young Ko

Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University, School of Medicine, 81, Irwon-ro, Gangnam-gu, Seoul, 06351, Republic of Korea

A R T I C L E I N F O

Keywords:
Thyroid
Inter-exam agreement
Ultrasonography
K-TIRADS

A B S T R A C T

Objective: To investigate the inter-exam agreement for thyroid nodule between real-time ultrasonography (US) assessment and retrospective US interpretation and to compare the diagnostic performance between two methods by using the most recently published guidelines for the US-based management of thyroid nodules, the Korean Thyroid Imaging Reporting and Data System (K-TIRADS).

Methods: The study included 253 nodules in 238 patients for the inter-exam agreement and 201 nodules in 190 patients for the diagnostic performance. Real-time and retrospectively static US images were analyzed according to the US descriptors and final categories of the K-TIRADS. Inter-exam agreements between real-time US assessments and static US image interpretation were analyzed, as was the diagnostic performance of both methods.

Results: Overall inter-exam agreements were almost perfect for orientation (κ = 0.868); and substantial for composition, spongiform appearance, echogenicity, shape, margin, calcification, and final K-TIRADS categories (κ = 0.754, 0.786, 0.747, 0.670, 0.666, 0.778, and 0.754, respectively). Specifically, moderate agreements were observed for predominantly cystic composition and ill-defined margin. The overall diagnostic performances of both real-time US assessment and retrospective US interpretation using the K-TIRADS were comparable.

Conclusions: Overall inter-exam agreements between real-time and retrospective US image interpretation for thyroid nodules using the K-TIRADS were equal or more than a substantial. Therefore, the use of K-TIRADS can provide consistent description and assessment for thyroid US regardless of the timing of interpretation.

1. Introduction

Thyroid ultrasonography (US) examination is useful for detecting thyroid nodules, discriminating between benign and malignant nodules, and guiding biopsy. According to the previous studies, diagnostic accuracy of US in thyroid nodules has been reported to be 81.5–99.1% [1–3]. However, US is subjective and operator-dependent, has low reproducibility, and it is criticized for its potential interobserver and intraobserver variations [4].

Until now, to overcome these limitations, many organizations have recommended guidelines for stratifying the risk of patients with thyroid nodules by using US and also for the selection of thyroid nodules for biopsy [1,5–10]. Among these guidelines, the most recently published guideline for the US-based management of thyroid nodules is the revised version of Korean Society of Thyroid Radiology (KSThR) consensus statement and recommendations, i.e., the Korean Thyroid Imaging Reporting and Data System (K-TIRADS) [10]. The KSThR has emphasized that the goal of the K-TIRADS is to provide the best scientific evidence available and a consensus expert-opinion regarding the US-based diagnosis and management of thyroid nodules in clinical practice. According to the K-TIRADS, we can classify the thyroid nodules into 5 categories for malignancy risk stratification by evaluating the US characteristics of thyroid nodules.

Previously, studies reported that interobserver agreement for the US assessment of thyroid nodule was low between experienced and less-experienced radiologists, and the diagnostic performance between them was significantly different [11,12]. However, to our knowledge, inter-exam agreement for the US assessment of thyroid nodule has not been examined between real-time and retrospective evaluation, although many studies on US findings of thyroid nodules have been performed retrospectively. Therefore, in this study, we investigated the inter-exam agreement for thyroid nodule between real-time US assessment and retrospective US interpretation and compared the diagnostic performance between two methods by using the K-TIRADS.

* Corresponding author.
E-mail addresses: jungmin1235.bae@samsung.com (J.M. Bae), aurore47@naver.com (S.Y. Hahn), helena35@hanmail.net (J.H. Shin), ey.ko@samsung.com (E.Y. Ko).

http://dx.doi.org/10.1016/j.ejrad.2017.10.027
Received 2 January 2017; Received in revised form 7 September 2017; Accepted 29 October 2017

2. Material and methods

2.1. Study population

This prospective study was approved by our institutional review board, and written informed consent was obtained from all participants. From November 2015 through February 2016, 309 consecutive patients who had a combined total of 353 thyroid nodules and who visited our institution for US-based thyroid evaluation were included in this study. All of them were scheduled for biopsy or surgery at the time of US. First, for the evaluation of inter-exam agreement, we excluded 87 nodules with a size of < 1 cm to prevent misinterpretation of US findings. Second, for the evaluation of diagnostic performance, we additionally excluded 34 nodules without follow up after being diagnosed as benign through fine needle aspiration (FNA) or core needle biopsy (CNB) to avoid false negative results, and 18 nodules with inconclusive FNA or CNB results (including non-diagnostic, atypia of undetermined significance/follicular lesion of undetermined significance [AUS/FLUS], or suspicious for follicular neoplasm/follicular neoplasm) due to the absence of the reference standard. Finally, the study included 253 nodules in 238 patients (183 women and 55 men; mean age ± standard deviation 49.7 ± 11.3, range 20–83) for the inter-exam agreement, and 201 nodules in 190 patients (150 women and 40 men; mean age ± standard deviation 49.3 ± 11.5, range 20–83) for the diagnostic performance. Of the 201 nodules, 102 were confirmed by surgery, 78 by FNA, and 21 by CNB.

2.2. Imaging and image analysis

Before starting this analysis, all 4 board-certified thyroid radiologists had evaluated more than 300 training cases to optimize objective image interpretation. In addition, they had a chance to control the threshold through a routinely scheduled intra-department conference.

For the real-time US evaluations, 3 of 4 thyroid radiologists were involved. All US examinations were performed by using an RS-80A with Prestige US system (Samsung Medison, Seoul, Korea) with a 5–13 MHz linear transducer. Both transverse and longitudinal scans were obtained during each examination and were stored as static images as we do in everyday practice. Each US examination was performed by 1 of 3 board-certified radiologists (with more than 8 years of experience in thyroid imaging). Then, the same radiologist who performed US examination prospectively evaluated the US images of thyroid nodule and recorded the results on the list. All image descriptors and final categories were analyzed in accordance with the K-TIRADS [10]. US descriptors included size, composition, presence of spongiform appearance, echogenicity, orientation, shape, margins, and calcification (Table 1). The final K-TIRADS category for stratifying patient risk was decided based on the US features; integrated solidity, echogenicity, and suspicious US features (including hypoechogenicity, microcalcification, nonparallel orientation, and spiculated/microlobulated margins) (Table 2).

The static US images were retrospectively evaluated by another radiologist (with 2 years of experience) who had not been involved in the real-time US examinations. This radiologist was blinded to the real-time US evaluation results, pathologic results, or clinical information of the patients. The previously stored static images were used for the retrospective analysis. All US descriptors and the final K-TIRADS categories were analyzed and recorded on the list.

As described above, we tried to design our study to be the most similar to the protocols used by radiologists to analyze the images in the actual prospective and retrospective studies.

2.3. Statistical analysis

The agreement between the assessments based on the real-time US and those based on retrospective static US images was evaluated for each descriptor examined. We used kappa statistics to calculate the degree of agreement. A kappa value (κ) of 1.0 corresponds to complete agreement; 0, no agreement; and less than 0, disagreement. Landis and Koch suggested that a kappa value ≤ 0.20 indicates slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–1.00, almost perfect agreement [13].

Imaging follow up with at least one histopathologic result was used as a reference standard for diagnostic performance. Thyroid nodules were finally classified as benign when there were at least two benign cytological or histological results without interval change in follow-up US at least 1 year or when thyroid nodules with previous benign FNA results decreased in size on follow-up US. The KSThR defined a significant size change as the nodule growth with a 20% increase in at least two nodule dimensions and a minimal increase of 2 mm, or more than a 50% change in volume [10], as previously stated by the American Thyroid Association [14]. Thyroid nodules with malignant results on surgery or with malignant/suspicious malignant results on repeat follow-up FNA were finally classified as malignant. The follow-up period included the time before the study. Finally, we calculated the diagnostic performances of the K-TIRADS. For analysis, we re-classified all thyroid nodules into two groups according to the final K-TIRADS categories: benign (K-TIRADS categories 2 and 3) versus malignant (K-TIRADS categories 4 and 5). We examined sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy.

All statistical analyses were performed by using PASW statistical software (version 23; SPSS, Chicago, IL, USA).

3. Results

3.1. Clinicopathologic characteristics

The clinicopathologic characteristics of 238 patients with 253 nodules are shown in Table 3.

3.2. Inter-exam agreement

Table 4 summarizes the inter-exam agreements between real-time US assessment and retrospective US interpretation.

Overall inter-exam agreements were almost perfect for orientation (κ = 0.868); and substantial for composition, spongiform appearance, echogenicity, shape, margin, calcification, and final K-TIRADS categories (κ = 0.754, 0.786, 0.747, 0.670, 0.666, 0.778, and 0.754, respectively). Among the descriptors evaluated on the US images, the highest level of inter-exam agreement was achieved for orientation, followed by the presence of spongiform appearance. Meanwhile, the lowest level of inter-exam agreement was achieved for margin. With respect to the final K-TIRADS categories, overall inter-exam agreement was substantial (κ = 0.754).

Specifically, for composition of nodules, almost perfect agreement was obtained for solid composition (κ = 0.810) and substantial agreements were obtained for predominantly solid (κ = 0.754) and cystic (κ = 0.738) compositions, but the agreement was moderate for predominantly cystic composition (κ = 0.492) (Fig. 1). In assessing echogenicity, almost perfect agreements were obtained for anechogenicity (κ = 1.000) and isoechogenicity (κ = 853), whereas substantial agreements were obtained for hypoechogenicity (κ = 0.752) and markedly hypoechogenicity (κ = 0.611) (Fig. 2). Agreements were substantial for smooth and spiculated/microlobulated margins (κ = 0.751 and 0.728); however, the agreement for ill-defined margins was moderate (κ = 0.465). Substantial to almost perfect agreements (κ = 0.656–0.849) were achieved in assessing calcifications. The agreement degrees of the final K-TIRADS categories varied. The least degree of agreement was observed for category 4 (substantial agreement, κ = 0.670), but almost perfect agreements were obtained for categories 2 (κ = 0.829) and 5 (κ = 0.809).