# The Accuracy of Thyroid Nodule Ultrasound to Predict Thyroid Cancer: Systematic Review and Meta-Analysis

Juan P. Brito, Michael R. Gionfriddo, Alaa Al Nofal, Kasey R. Boehmer, Aaron L. Leppin, Carl Reading, Matthew Callstrom, Tarig A. Elraiyah, Larry J. Prokop, Marius N. Stan, M. Hassan Murad, John C. Morris, and Victor M. Montori

Departments of Diabetes, Metabolism, and Nutrition (J.P.B., M.N.S., J.C.M., V.M.M.), Pediatric Endocrinology and Metabolism (A.A.N.), and Radiology (C.R., M.C.), Knowledge and Evaluation Research Unit (J.P.B., M.R.G., K.R.B., A.L.L., T.A.E., L.J.P., M.H.M., V.M.M.), Mayo Graduate School (M.R.G.), and Division of Preventive Medicine (M.H.M.), Mayo Clinic, Rochester, Minnesota 55905



**Context:** Significant uncertainty remains surrounding the diagnostic accuracy of sonographic features used to predict the malignant potential of thyroid nodules.

**Objective:** The objective of the study was to summarize the available literature related to the accuracy of thyroid nodule ultrasound (US) in the prediction of thyroid cancer.

**Methods:** We searched multiple databases and reference lists for cohort studies that enrolled adults with thyroid nodules with reported diagnostic measures of sonography. A total of 14 relevant US features were analyzed.

**Results:** We included 31 studies between 1985 and 2012 (number of nodules studied 18 288; average size 15 mm). The frequency of thyroid cancer was 20%. The most common type of cancer was papillary thyroid cancer (84%). The US nodule features with the highest diagnostic odds ratio for malignancy was being taller than wider [11.14 (95% confidence interval 6.6–18.9)]. Conversely, the US nodule features with the highest diagnostic odds ratio for benign nodules was spongiform appearance [12 (95% confidence interval 0.61–234.3)]. Heterogeneity across studies was substantial. Estimates of accuracy depended on the experience of the physician interpreting the US, the type of cancer and nodule (indeterminate), and type of reference standard. In a threshold model, spongiform appearance and cystic nodules were the only two features that, if present, could have avoided the use of fine-needle aspiration biopsy.

**Conclusions:** Low- to moderate-quality evidence suggests that individual ultrasound features are not accurate predictors of thyroid cancer. Two features, cystic content and spongiform appearance, however, might predict benign nodules, but this has limited applicability to clinical practice due to their infrequent occurrence. **(*J Clin Endocrinol Metab* 99: 1253–1263, 2014)**

Thyroid nodules are common: 4%–7% of adults in North America have palpable nodules (1). When informed by imaging studies, the prevalence rises to 30% (2), and when autopsies are conducted, approximately 60% of North American adults are found to harbor nodules (3). Thus, it can be concluded that a reservoir of subclinical disease exists in nearly two of every three Americans (4).

Because of the 5%–15% probability of malignancy in any given thyroid nodule (5,6), current thyroid guideline recommendations call for ultrasound (US) in all patients with a suspected thyroid nodule (7). A combination of clinical factors and ultrasound features determine whether the clinician should proceed with further confirmatory tests or with periodical US follow-up. For example, ac-

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in U.S.A.

Received July 22, 2013. Accepted November 8, 2013.
First Published Online November 25, 2013

Abbreviations: CI, confidence interval; DOR, diagnostic odds ratio; FNAB, fine-needle aspiration biopsy with cytology; LR, likelihood ratio; US, ultrasound.

cording to guidelines, a suspicious nodule should undergo a fine-needle aspiration biopsy with cytology (FNAB), which, depending on the results, could lead to thyroid surgery.

Within this diagnostic algorithm, the use of US evaluation has become widely accepted as a key diagnostic step in stratifying patients' risk of malignancy (7). Nevertheless, there is significant uncertainty surrounding the diagnostic accuracy of several of the features analyzed during the sonographic evaluation of thyroid nodules.

A better understanding of the US features predictive of malignancy or benign disease may avoid costly confirmatory testing and have a large impact on both guideline recommendations and clinical practice. Therefore, we conducted a systematic review and meta-analysis that appraises and summarizes the available evidence related to the diagnostic accuracy of sonographic features of thyroid nodules for thyroid cancer.

## Materials and Methods

This systematic review was conducted based on standard methods recommended by the Cochrane Collaboration (8) and followed a predefined protocol. This report follows the standards set in the Preferred Reporting Items for Systematic Reviews and Meta-analysis statement (9).

### Eligibility criteria

We searched for randomized trials and cohort studies that enrolled adults with thyroid nodules with sonography results or reported diagnostic measures of sonography We included studies in English, regardless of their sample size or publication status.

For the purpose of this study, thyroid nodules were defined as any discrete lesion that was sonographically distinguishable from the adjacent thyroid parenchyma (10). The test of interest was two-dimensional thyroid US. We considered histopathological diagnosis after surgery to be the gold standard reference test. However, because this reference test is not likely to be performed for all benign cases, we considered a hierarchy of reference standards for benign nodules: 1) core thyroid biopsies, 2) two consistent FNABs, and 3) one FNAB with a follow-up of a minimum of 6 months demonstrating reduction or stabilization of nodule size. Only pathological diagnosis and core biopsy diagnosis were considered adequate reference standards for nodules with malignant, indeterminate, or nondiagnostic cytologies. We excluded reports that had a population with a prior history of thyroid cancer or were clearly exposed to known risk factors for thyroid cancer, eg, Chernobyl survivors.

Outcomes of interest were the diagnostic accuracy of sonographic features of thyroid nodules. To that end, we used the diagnostic odds ratio (DOR) as our main outcome measure. The DOR is a measure for the discriminative power of a diagnostic test: it is the ratio of the odds of a positive test result among diseased to the odds of a positive test result among the nondiseased and it reflects the test's performance compared with the reference standard (11). The value of a DOR ranges from 0 to infinity, with higher values indicating better discriminatory test performance. A DOR value of 1 means that a test does not discriminate between patients with the disorder and those without it. Values lower than 1 point to improper test interpretation (more negative tests among the diseased) (12). Other outcomes examined were sensitivity, specificity, and the likelihood ratio (LR) of a positive and negative test.

### Study identification

We used a comprehensive search of several databases from each databases' earliest inception to December 2012. The databases searched were Ovid Medline In-Process and Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Ovid Cochrane Central Register of Controlled Trials, Ovid Cochrane Database of Systematic Reviews, and Scopus. The search strategy was designed and conducted by an experienced librarian (L.J.P.) with input from the study's principle investigator (J.P.B.). Controlled vocabulary supplemented with key words was used to search for studies of diagnostic accuracy of sonography for thyroid cancer. The reference lists from primary studies and narrative reviews were searched and consultation with experts in the field was performed to obtain any additional references that might have been missed by our initial search strategy.

Reviewers working independently and in duplicate reviewed all abstracts and titles. Upon retrieval of potentially eligible studies, the full-text publications were evaluated for eligibility. The chance-adjusted interreviewer agreement was calculated using the $\kappa$ statistic for abstract abstraction ($\kappa = 0.9$) and for full-text screening ($\kappa = 0.84$). Disagreements were resolved by either consensus or arbitration.

### Quality assessment

Reviewers working independently and in duplicate analyzed the full text of eligible articles to assess the reported quality of the methods. Using QUADAS2, the current best tool for quality assessment of studies of diagnostic accuracy in systematic reviews, we assessed the four key quality domains: 1) selection of patients; 2) conduct and interpretation of the index test (eg, lack of reliability at the moment of reporting thyroid ultrasounds); 3) type and interpretation of the reference standard (considered optimal when it consisted of thyroid surgery, core biopsy, or two consecutive FNABs and suboptimal when defined as a single FNAB with follow-up only); and 4) patient flow, timing, and exclusions (13). Chance-adjusted interrater agreement was substantial ($\kappa = 0.73$); disagreements were resolved by consensus.

### Data extraction

Reviewers working independently and in duplicate used a standardized Web-based form to extract, for each eligible study, the following data items: the country where the study was conducted; number of patients and nodules; patient age, gender, ethnicity, gland nodularity (single, multinodular, or unclear), US equipment characteristics (probe frequency on hertz), experience of the interpreting physician (in years), number of interpreting physicians, interobserver variability during the US examination, type of thyroid cancer, and total number of benign or malignant lesions confirmed by standard.

For each nodule, we evaluated the presence or absence of fourteen sonographic features mentioned as important in the American Thyroid Association Guidelines for patients with thyroid nodules (7) and the Consensus Conference Statement from the Society of Radiologists in Ultrasound (10) and the Ultra-

sound-Based Management of Thyroid Nodules: Consensus Statement and Recommendations from the Thyroid Study Group of the Korean Society of Radiology (14). Based on these guidelines, we hypothesized that the following nodule features would be predictive of malignancy: internal calcifications, hypoechogeneicity, increased blood flow centrally, taller than wider, solid, and larger size. Furthermore, we hypothesized that the following nodule features would be predictive of benignity: isoechogeneicity, increased blood flow peripherally, and spongiform or cystic in nature. Blurred or irregular margins were not considered for this analysis due to the inconsistency of definition across included studies and referenced guidelines.

We looked in each report for the features that matched our definition (Supplemental Appendix 1, published on The Endocrine Society's Journals Online web site at http://jcem.endojournals.org) and extracted true-positive, true-negative, false-negative and false-positive values to construct diagnostic $2 \times 2$ tables. In the case of the continuous variable, nodule size, we used the cutoff value chosen by the study authors.

### Author contact

To help reduce the impact of reporting bias, we contacted, via e-mail, the corresponding authors of all included studies to confirm the accuracy of the extracted data (15). We waited 2 weeks before sending out a reminder e-mail to nonresponders. When authors did not reply, we used the unconfirmed data extracted by our team.

### Subgroup analysis

A priori hypotheses to explain inconsistency across study results included differences in probe frequency in hertz ($\geq$10 MHz vs $<$10), the experience of the radiologist in years (15 or more years of experience vs less than 15 years), types of thyroid cancer evaluated (studies with high frequency of papillary thyroid cancer vs reports with lower frequency of papillary thyroid cancer), and quality of the report for the index test and reference standard.

### Statistical analysis

We used the random-effects model of DerSimonian and Laird (16) to pool sensitivities, specificities, likelihood ratios, and DORs and estimate the 95% confidence intervals for each feature. We used the $I^2$ statistic and Cochran's Q test to assess heterogeneity across individual studies. $I^2$ values of less than 25%, 25%–50%, and greater than 50% indicate low, moderate, and high heterogeneity, respectively. We also evaluated publication bias through visual analysis of a funnel plot. The analysis was done using Metadisc and Review Manager (Revman) (computer program, version 5.1; The Nordic Cochrane Center, The Cochrane Collaboration, Copenhagen, Denmark, 2011).

For each subgroup analysis, we conducted a test for interaction, with a critical value of $P < .05$ being considered significant (17).

## Results

### Included studies

We included 31 studies published between 1985 and 2012 (Figure 1). These studies enrolled 13 736 adult patients with a mean age of 47 years; they were mostly women ( 82%) (Table 1). No data on ethnicity were provided. The total number of nodules evaluated was 18 288. The average size of these nodules was 15 mm and the pooled frequency of thyroid cancer was 20%. The most common type of cancer was papillary thyroid cancer (84%), followed by follicular carcinoma (13%). Nodules with either inconclusive cytology or indeterminate nodules were exclusively investigated in seven studies.

We contacted authors to verify the extracted data. Five of 31 authors responded by confirming the accuracy of the data extracted from their studies.

### Methodological quality

Figure 2 summarizes the methodological quality of the 31 included studies. Limitations included inappropriate exclusion of patients, lack of reliability of the ultrasound evaluation, suboptimal reference standard, and differential use of reference standards for all patients. These limitations increased the likelihood of bias and reduced the reliability of the estimates of diagnostic accuracy.

### Meta-analysis

The pooled estimates of the sensitivity, specificity, DOR, and LR for the presence or absence of each of the 14 features are represented in Tables 2 and 3. The US features with the highest DOR for correctly indicating malignancy were taller than wider [11.14 (95% confidence interval [CI] 6.6–18.9)] and internal calcifications [6.78 (95% CI 4.48–10.24)]; however, taller than wider was a feature reported in only 12 of the included studies. On the other hand, the US features with the highest DOR indicating benignity were a spongiform appearance [12 (95% CI 0.61–234.3)] and being cystic [6.78 (95% CI 2.26–20.3)]. We also found that thyroid nodule size is not an accurate predictor of thyroid cancer across different cutoffs. Heterogeneity across studies was substantial in most analyses. Evaluation of publication bias could not reliably be performed due to the substantial heterogeneity (18).

### Subgroup analysis

We conducted predefined subgroup analyses within each analyzed feature. We found some interactions of potential importance (Supplemental Appendix Tables 4 and 5). The experience of the physician interpreting the US was important in the evaluation of internal calcifications. The DOR for the physician interpreting the US examination with more experience was higher than the ones with less experience [14.5(95% CI 8.5–25.14) and 5.36 (95% CI 2.72–10.57), respectively] with a pinteraction of 0.025. The type of cancer was found to influence the DOR for echogenic features of the nodules: for hypoechoic nodules

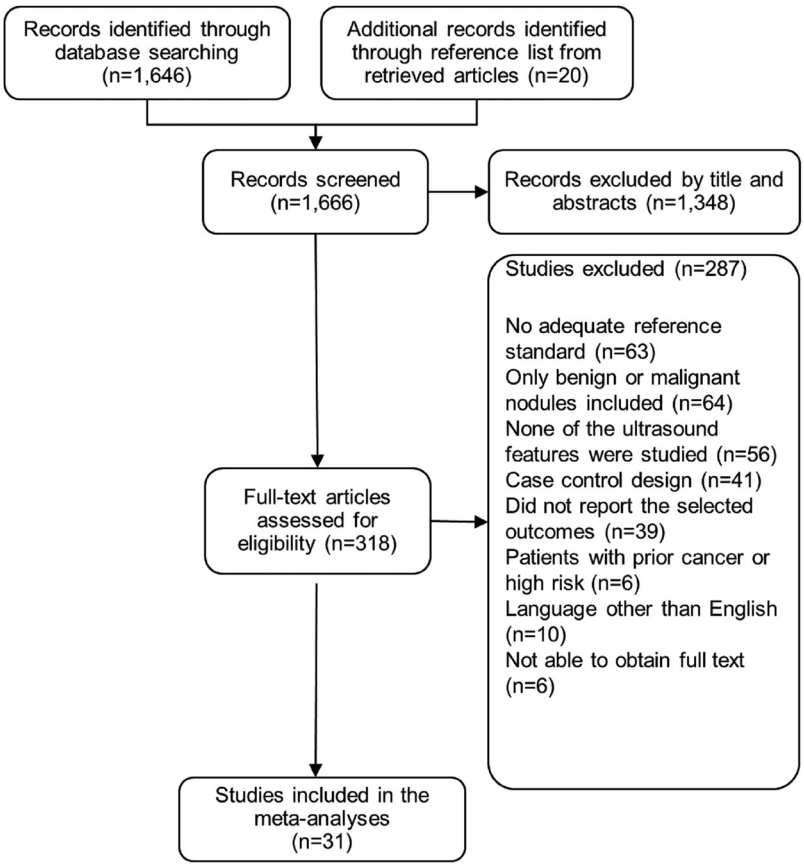


**Figure 1.** Study selection.

in studies in which more than 90% of cancers were papillary, the DOR was 9.29 (95% CI 4–21.4); in studies in which less than 90% of cancers were papillary the DOR of hypoechogenicity fell to 2.85 (95% CI 1.9–4.2) (pinteraction = 0.01). For the two features (internal calcification and hypoechogenicity) in which the analysis of US potency was possible, we found higher DOR in the group using higher US potency; however, these findings were not statistically significant.

The analysis of the subgroup of studies that exclusively included indeterminate nodules revealed a significantly lower diagnostic accuracy for several features (internal calcifications, echogenecity, infiltrative margins, and content of the nodule) compared with their accuracy in nonindeterminate nodules. However, one feature, increased blood flow centrally, had a statistically significant increase in DOR compared with the nonindeterminate nodules. The frequency of thyroid cancer in this group was higher at 46% (79% papillary type and 20% follicular neoplasia).

We conducted subgroup analyses for two quality features: 1) possible bias of the index test and 2) possible bias of the reference standard test as described in the *Quality assessment* section. There were no statistically significant differences for possible bias for index test. However, we observed an overestimation of the DOR for all the features confirmed with suboptimal reference standards. The differences were statistically significant for echogenic features (hypoechoic and isoechoic) and infiltrative margins. For instance, nodules with infiltrative features evaluated against suboptimal standards had a DOR of 13.1 (95% CI 6.1–28.1); those evaluated against an optimal standard had a DOR of 2.5 ( 95% CI 1.0–6.4).

## Discussion

We conducted a systematic review of 14 ultrasound features of thyroid nodules that are used to predict thyroid cancer. We found that two nodule features, spongiform and cystic, were significantly associated with an increased likelihood of nodule benignity. Yet the confidence that nodules with these features are not malignant is higher for cystic nodules than for spongiform nodules due to the imprecision of the CI for the latter feature.

For US features assessed for their accuracy at predicting malignancy, internal calcifications, especially when identified by experienced radiologists, are very specific for thyroid carcinoma and perhaps particularly so for the papillary subtype. In addition, nodules that are taller than wider were found to have the highest DOR among all the analyzed features, suggesting that thyroid cancer does not respect normal tissue planes and grows in a centrifugal way (14); a similar finding has been observed in breast cancer nodules (19). We also found that thyroid nodule size is not an accurate predictor of thyroid cancer across different cutoffs, a finding that holds true across all the subgroups.

Diagnostic US features, when used in nodules with indeterminate cytologies, have a lower predictive value. In this study, internal calcifications, echogenecity, infiltrative margins, and solid vs cystic content of the nodule were poor indicators of malignancy, whereas increased internal vascularity was more predictive. The explanation for this most likely lies in the increased frequency of follicular neoplasm within the analyzed population. This subtype of cancer differs from papillary thyroid cancer in that it is most frequently encapsulated with regular margins, contains high cellularity providing variation in echogenicity

**Table 1.** Included Studies

| Author, Year of Publication | Country | Study Main Objective | Age, y, Mean | Male, % | Number of Nodules, n | Features Analyzed | Patients With MNG, % | Standard for Benign Nodules |
|---|---|---|---|---|---|---|---|---|
| Atli et al, 2006 (29) | Turkey | Define the risk factors predicting malignancy | 41 | 13 | 845 | IC, HN IN, AH, SN, CN | 91 | Optimal |
| Brkljacic et al, 1994 (30) | Croatia | Sonographic features of nodules in MNG goiters that relate with malignancy | 47 | 13 | 490 | IC, HN | 100 | Optimal |
| Brunese et al, 2008 (31) | Italy | Determine whether the BFI-TS is predictive of malignancy | 42 | 32 | 539 | IC, HN | NR | Less optimal |
| Cappelli et al, 2006 (32) | Italy | To evaluate whether a nodule with shape taller than wide is a good predictor of malignancy independent of the size | NR | 31 | 6135 | IC, HN, IBC, NS | 30 | Both |
| Chen et al, 2009 (33) | China | Significance of thyroid nodule calcifications detected by US, in patients with thyroid malignancy | 44 | 18 | 999 | IC | 58 | Optimal |
| Choi et al, 2009 (34) | South Korea | Identify US features of malignancy in follicular neoplasms | 46 | 19 | 114 | IC, HN, IN, IBC, IBP, IM, AH, SN, NS, CN | 0 | Optimal |
| Chung et al, 2012 (35) | South Korea | Investigate the incidence of thyroid cancer among cases with nondiagnostic results on FNAB | 50 | 15 | 143 | IC, HN, IN, IM, TW, SN | NR | Less optimal |
| Gulcelik et al, 2008 (36) | Turkey | Identify ultrasonographic features to predict malignancy in patients with thyroid follicular neoplasm | 47 | 16 | 98 | IC, HN, SN, NS, CN | 78 | Optimal |
| Hong et al, 2012 (37) | China | Determine whether nodule size affects the differential diagnosis of benign and malignant | 45 | 12 | 329 | IC, HN, IBC, NS | 27 | Optimal |
| Kakkos et al, 2000 (38) | Greece | Investigate the value of sonographically detected thyroid calcifications in diagnosing thyroid carcinoma | 47 | 20 | 188 | IC | 56 | Optimal |
| Kim et al, 2002 (39) | South Korea | Assess role of sonography in the differentiation of benign from malignant nonpalpable thyroid lesions | 48 | 9 | 155 | IC, HN, IM, TW | 13 | Both |
| Kim et al, 2008 (40) | South Korea | Investigate the ultrasonographic and pathological findings of nonpalpable thyroid carcinomas | 52 | 33 | 140 | IC, HN, IM, IBP TW, SN, NS | NR | Both |
| Kwak et al, 2009 (41) | South Korea | Assess the diagnostic accuracy of sonographic findings of subcentimeter thyroid nodules. | 48 | NR | 815 | IC, HN, IN, IM, TW,SN | NR | Both |
| Lee et al, 2011 (42) | South Korea | Evaluate the diagnostic accuracy of a new ultrasound classification system for differentiating between benign and malignant solid thyroid nodules | 48 | 15 | 191 | IC, HN, IN, IBC, IBP, IM, TW | 77 | Both |
| Leenhardt et al, 2002 (43) | France | Improve the preoperative selection for operation of patients with solitary thyroid nodules | 42 | 25 | 155 | IC, HN, IN, AH, SN, CN | NR | Optimal |
| Mendelson et al, 2009 (44) | Canada | Determine whether preoperative variables can be used to predict malignancy for thyroid nodules with follicular or nondiagnostic cytology | NR | 17 | 77 | IC, NS | NR | Optimal |
| Mendez et al, 2008 (45) | US | Determine the clinical value of ultrasound in predicting the presence of malignancy in nodules with indeterminate cytology | NR | 19 | 180 | IC, HN,TW, SN, NS, CN | 56 | Optimal |

(*Continued*)

**Table 1.** Continued

| Author, Year of Publication | Country | Study Main Objective | Age, y, Mean | Male, % | Number of Nodules, n | Features Analyzed | Patients With MNG, % | Standard for Benign Nodules |
|---|---|---|---|---|---|---|---|---|
| Moon et al, 2012 (46) | South Korea | Investigate the factors for considering surgery on thyroid nodules that had nondiagnostic cytologies | 50 | 13 | 104 | IC, HN, IN, IM, TW, SN, CN | NR | Both |
| Moon et al, 2008 (47) | South Korea | Evaluate the diagnostic accuracy of ultrasound to predict benign and malignant thyroid nodules | 50 | 14 | 849 | IC, HN, IN, IM, TW, SN, SPN, NS, CN | NR | Both |
| Ozel et al, 2012 (48) | Turkey | Evaluate the diagnostic accuracy of ultrasound to predict benign and malignant thyroid nodules | 48 | 15 | 363 | IC, HN, IN, IBC, IBP,TW, NS | NR | Less optimal |
| Papini et al, 2002 (49) | Italy | Importance of ultrasound features as risk factors of malignancy | 48 | 13 | 402 | HN, IN, NS | 52 | Less optimal |
| Phuttharak et al, 2009 (50) | Thailand | Evaluate gray- and color-scale ultrasound in predicting malignancy of thyroid nodules | 42 | 3 | 31 | IC, HN, IN, IBC, IBP, IM, TW, AH, SN, SPN, CN | 48 | Optimal |
| Popowicz et al, 2009 (51) | Poland | Evaluate the efficacy of selected ultrasound features of thyroid focal lesions | 50 | NR | 1141 | IN, HN, IBC, NS | NR | Both |
| Rago et al, 2007 (52) | Italy | Evaluate echographic patterns predictive of malignancy in patients with follicular cytology | 45 | 21 | 505 | IC, HN, IM, NS | 40 | Optimal |
| Sahin et al, 2006 (53) | Turkey | Predict malignancy based on ultrasonographic features in indeterminate follicular thyroid lesions | 52 | 17 | 86 | IC, HN, IN, IM, SN, CN | NR | Optimal |
| Salmaslıoglu et al, 2008 (54) | Turkey | Predictive value of sonographic features in the preoperative diagnosis of malignant thyroid nodules | 47 | 19 | 1926 | IC, HN, IN, SN, CN | NR | Optimal |
| Schueller-Weidekamm et al, 2010 (55) | Norway | Assess the diagnostic value of different modalities for the characterization of cold thyroid nodules | 55 | 31 | 31 | IC, HN, IBC, CN | NR | Optimal |
| Sharma et al, 2011 (56) | US | Determine the usefulness of subcentimeter thyroid nodule evaluation | 53 | 27 | 67 | IC, HN, IBC, IBP, IM, TW, AH, SN, CN | NR | Optimal |
| Solbiati et al, 1985 (57) | Italy | Specific echographic patterns of thyroid nodules | NR | NR | 430 | HN, IN, IM, AH, CN | NR | Optimal |
| Yoon et al, 2011 (58) | Korea | Evaluate sonographic differences between benign and malignancy in thyroid nodules ≥ 3 cm | 48 | 18 | 661 | IC, HN, IN, IM, SN, CN | 45 | Both |
| Yoon et al, 2010 (59) | South Korea | Evaluate ultrasound characteristics that predict malignancy in thyroid nodules | 44 | 13 | 99 | IC, HN, IN, IM, TW, SN | 44 | Optimal |

Abbreviations: AH, absent halo; CN, cystic nodule; HN, hypoechoic nodule; IBC, increased blood flow centrally; IBP, increased blood flow peripherally; IC, internal calcification; IM, infiltrative margins; IN, isoechoic nodule; NS, nodule size; SN, solid nodule; SPN, spongiform nodules; TW, taller than wider.

and internal vascularity, and contains fewer calcifications of papillae tips (20, 21) .

## Limitations and strengths

Several limitations weaken the inferences from this review. First, not all the US features in this review were compared with the gold standard criterion, histology, which is likely due to the fact that this reference test necessitates surgery and may require unjustified expense, especially for benign-appearing nodules. To overcome this limitation, we created a hierarchy of gold standards. Standards with a greater chance for misclassification were considered to be at high risk of bias. Subsequent analysis demonstrated overestimation of the accuracy of some US features (ie, echogenicity, margins, and perhaps the internal blood flow of the nodules) that were confirmed by standards at high risk of bias. This finding suggests that prior studies reporting high diagnostic accuracy for these features might have overestimated accuracy due to inadequacy of the reference standard.

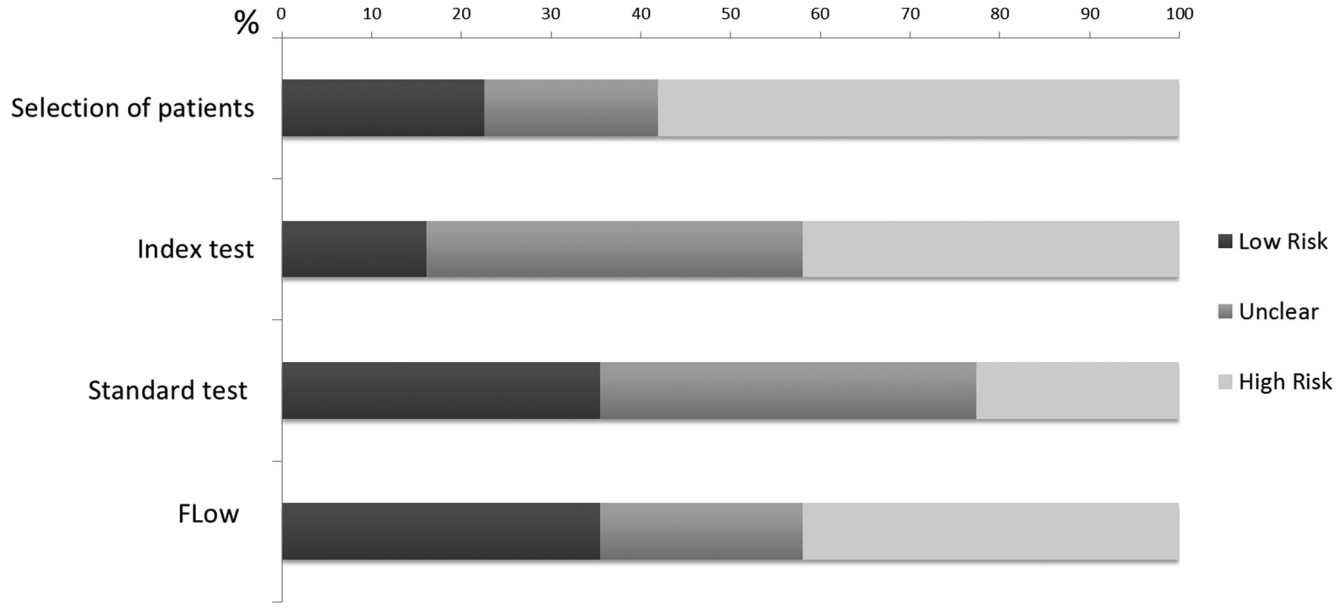


**Figure 2.** Quality assessment of included studies with the four main domains of QUADAS2 to evaluate the risk of bias: 1) selection of patients: could the selection of patients have introduced bias?; 2) index test: could the conduct or interpretation of the index test have introduced bias?; 3) standard test: could the reference standard, its conduct, or its interpretation have introduced bias?; and 4) flow: could the patient flow have introduced bias?.

Another area of concern was the manner in which the index test was conducted. Many reports did not describe the presence of interradiologist variation, whether independent radiologist review was procured or the blinding status of radiologists to clinical data (22). We again classified reports lacking these items as warranting only low confidence in the results and subsequently conducted a subgroup analysis. In this case, the subgroup analysis did not demonstrate any difference across the analyzed features.

We also acknowledge that other factors not specifically mentioned in the primary studies or in this review could have affected the overall estimation of the diagnostic properties of the US features. For example, the visibility of thyroid nodules might be affected by US machine properties beyond probe frequency (namely, power generated by the machine and differences between spacial resolution and the use of harmonics with the imaging). Optimal recognition of the sonographic features of thyroid nodules requires high-quality ultrasound machines and physician experience in interpreting the images. For example, not all echogenic foci within nodules are calcifications, and some of the foci are instead echogenic material with reverberation artifacts associated with colloid, which are strongly associated with benign nodules, so accurate interpretation is important (23). In addition, it was unknown whether the reading of the US features was conducted with real-time US imaging vs static US images. Real-time reading would provide more reliable reading, especially from nodules with ambiguous features (eg, irregular border vs infiltrative margins). This measurement bias was not accounted for in the individual papers and thus in our systematic review might have caused a misclassification of the features with a subsequent misestimating of their accuracy.

Finally, it is unknown whether differences in the diagnostic properties of each report are due to some degree of conditional nonindependence (24–26). The probability of one US feature to be positive or negative can partially depend on the presence of another US feature. For instance, markedly hypoechoic nodules could complicate the visualization of posterior acoustic shadowing and make the identification of internal calcifications more

**Table 2.** Pool Estimates of Diagnosis Parameters of the US Features to Predict Malignant Nodule

| | Number of Nodules | Sensitivity | $I^2$, % | Specificity | $I^2$, % | Positive LR | $I^2$, % | Negative LR | $I^2$, % | DOR | $I^2$, % |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Internal calcifications | 17151 | 0.54 (0.52–0.56) | 93 | 0.81 (0.8–0.82) | 98 | 3.65 (2.78–4.8) | 94 | 0.58 (0.5–0.64) | 92 | 6.78 (4.48–10.24) | 91 |
| Hypoechoic | 17014 | 0.73 (0.72–0.75) | 94 | 0.56 (0.5–0.57) | 98 | 1.85 (1.6–2.1) | 91 | 0.5 (0.4–0.6) | 88 | 4.5 (3.2–6.4) | 88 |
| Increased blood flow (centrally) | 7578 | 0.48 (0.43–0.51) | 93 | 0.53 (0.51–0.54) | 97 | 1.4 (1.2–1.6) | 32 | 0.83 (0.73–0.94) | 61 | 1.8 (1.48–2.2) | 0 |
| Infiltrative margins | 4390 | 0.56 (0.5–0.50) | 85 | 0.79 (0.77–0.8) | 92 | 3.76 (2.26–6.3) | 95 | 0.62 (0.48–0.81) | 95 | 6.89 (3.35–14.1) | 93 |
| Taller than wider | 3137 | 0.53 (0.5–0.56) | 95 | 0.93 (0.91–0.94) | 87 | 5.4 (3.86–7.60) | 57 | 0.6 (0.46–0.79) | 96 | 11.14 (6.6–18.9) | 66 |
| Absent of halo | 1646 | 0.26 (0.2–0.32) | 92 | 0.69 (0.66–0.71) | 98 | 0.83 (0.5–1.34) | 86 | 0.20 (0.91–1.5) | 78 | 0.54 (0.21–1.39) | 81 |
| Solid nodule | 6303 | 0.87 (0.85–0.89) | 97 | 0.56 (0.54–0.58) | 99 | 1.47 (1.18–1.84) | 97 | 0.35 (0.18–0.7) | 97 | 4.45 (2.63–7.5) | 80 |
| Size nodule > 1 cm | 8897 | 0.57 (0.54–0.60) | 93 | 0.4 (0.39–0.41) | 99 | 1.14 (0.78–1.66) | 97 | 1 (0.7–1.5) | 93 | 1.1 (0.48–2.5) | 95 |
| Size nodule > 3 cm | 582 | 0.37 (0.29–0.44) | 0 | 0.59 (0.54–0.64) | 0 | 0.9 (0.7–1.14) | 0 | 1.07 (0.93–1.24) | 0 | 0.94 (0.57–1.23) | 0 |
| Size nodule > 4 cm | 380 | 0.24 (0.17–0.32) | 65 | 0.77 (0.7–0.82) | 62 | 1.24 (0.57–2.68) | 75 | 0.9 (0.7–1.2) | 70 | 1.3 (0.47–3.79) | 75 |

**Table 3.** Pool Estimates of Diagnostic Parameters of the US Features to Predict Benign Nodules

| | Number of Nodules | Sensitivity | $I^2$, % | Specificity | $I^2$, % | Positive LR | $I^2$, % | Negative LR | $I^2$, % | DOR | $I^2$, % |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Isoechoic | 7181 | 0.47 (0.46–0.48) | 98 | 0.84 (0.83–0.86) | 91 | 2.35 (1.55–3.54) | 90 | 0.69 (0.59–0.80) | 93 | 3.6 (2–6.3) | 90 |
| Increased blood flow (peripherally) | 766 | 0.38 (0.34–0.41) | 94 | 0.86 (0.79–0.91) | 84 | 2.1 (0.6–7) | 89 | 0.73 (0.47–1.15) | 90 | 3 (0.56–16.3) | 88 |
| Spongiform | 880 | 0.1 (0.08–0.14) | 0 | 0.99 (0.99–1) | 0 | 10.1 (0.49–208.2) | 68.5 | 0.89 (0.87–0.93) | 0 | 12 (0.61–234.3) | 63 |
| Cystic nodule | 5559 | 0.32 (0.31–0.33) | 99 | 0.98 (0.97–0.99) | 85 | 5.5 (1.7–17.7) | 88 | 0.81 (0.69–0.96) | 99 | 6.78 (2.26–20.3) | 81 |

challenging. To account for this, we searched for, but were unable to identify, any relevant literature related to collinearity of US features. Regardless, we believe conditional nonindependence has likely affected the diagnostic accuracy estimates observed in this analysis, although we cannot determine to what degree.

The strengths of our analysis relate to the specific and a priori study selection criteria. We included a study population at standard risk of thyroid cancer and excluded case-control studies, which can exaggerate the association between diagnostic features and outcomes. These criteria were designed to avoid overestimation of the results and to provide applicable results to assist providers at the point of care, particularly in situations of diagnostic uncertainty. Furthermore, our study included an extensive literature search, reproducible judgments and data collection, and preplanned analyses, including predefined subgroup analyses.

### Implications for practice and for further research

There is a large reservoir of asymptomatic (and potentially inconsequential) thyroid nodules in the population. In an age of easily accessible and frequently used imaging technologies, the probability that these nodules will be discovered is increasing. In the absence of accurate clinical or US predictors of malignancy, many of the nodules will require FNABs, which carry their own set of costs and diagnostic challenges, eg, indeterminate and nondiagnostic cytologies. Here we have reported the diagnostic accuracy of each feature based on their LR. Therefore, tests with a low LR for negative results may rule out malignancy and the need for FNAB, whereas tests with high LR for positive results may rule in malignancy and the need for FNAB. Only two US features evaluated here (spongiform and cystic characteristics) might result in a sufficient posttest probability to help rule out cancer and avoid FNABs; however, these findings are not often present in thyroid nodules (seen in only ~2%). The other US features assessed individually might not be able to rule in or rule out malignancy due to their modest likelihood ratios. For instance, a nodule with internal calcifications would increase the probability of malignancy from 20% (pretest probability) to 50% (posttest probability), whereas in a cystic nodule, the probability of malignancy would decrease to 2%. Figure 3 illustrates the use of a Fagan nomogram to estimate the posttest probability of thyroid cancer using the pretest probability (prevalence of thyroid cancer) and the LR of the tests evaluated.

Nodules often present with more than one US features. Thus, further research should focus on understanding the level of collinearity among the various diagnostic US features and on the construction of valid prediction models that could be tested in clinical practice. For instance, if no collinearity is found between internal calcifications and the shape of the nodule, one could calculate, based on our data, a combined LR of 20 when both are present, which would generate a posttest probability of 85% for malignancy. Armed with knowledge, a clinician would gain confidence that proceeding with treatment or additional testing is appropriate.

Despite limitations precluding more definitive posttest probabilities, US features provide valuable posttest information as a component of the overall workup for thyroid nodules. Approximately 10%–30% of the aspirates of thyroid FNABs are classified as nondiagnostic cytologies (27, 28). In many of these cases, repeated FNAB or surgery is often needed. This clinical context may be the most appropriate opportunity for the utilization of US features in risk stratification and clinical decision making. For example, a patient with a solid nondiagnostic nodule by cytology that harbors internal calcifications on US might be considered for surgery as opposed to observation. The applicability of this approach could also be considered for nodules with indeterminate cytology but must take into account the lower predictive value of sonographic features within this subgroup of nodules.

Finally, we suggest that when a clinician is considering conducting a FNAB that they take into account the following factors: 1) clinical and US predictors of thyroid cancer for thyroid nodules to obtain an overall risk of malignancy; 2) the probable natural history of the lesion (eg, nodules $< 1$ cm likely to exhibit an indolent course); and 3) the patient's values and preferences and engage in a shared decision-making conversation, which considers those values and preferences and results in a course of action that fits the patient's values, preferences, goals, and context.

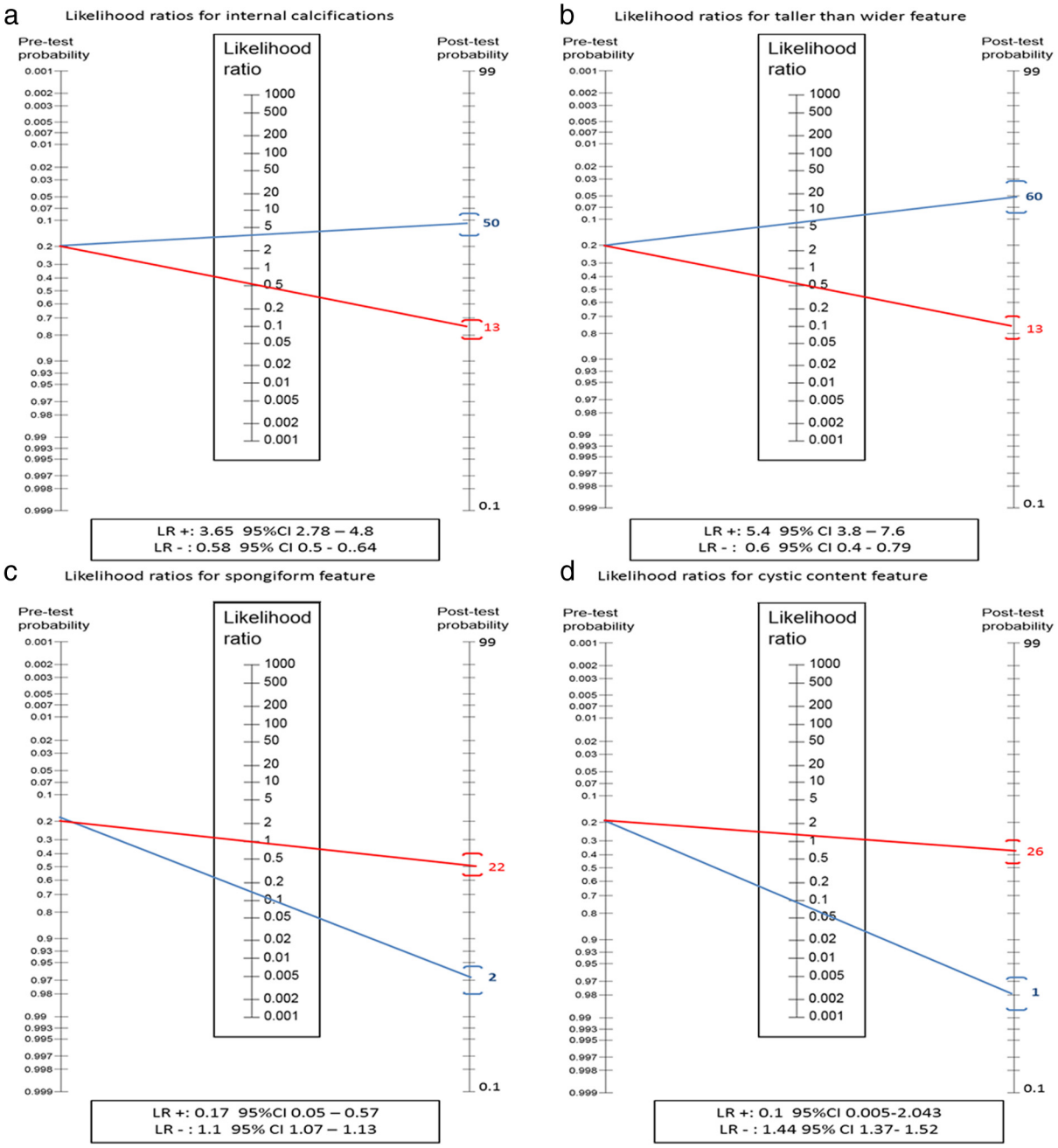


**Figure 3.** Fagan nomogram representing the LR for positive results (blue) and the LR for negative results (red) of the four most notable features. The nomogram has three components: 1) pretest probability, which is the estimated prevalence of thyroid cancer (overall prevalence of thyroid cancer in this review 20%); 2) LR of the feature; 3) posttest probability, which is the probability of having the condition given that feature is present (blue) or absent (red). The LR for spongiform and cystic feature was calculated based on the LR to predict benign nodule (eg, the false positives for benign nodules is the true positive for malignant nodule).

## Conclusion

Low- to moderate-quality evidence suggests that individual ultrasound features are not accurate predictors of thyroid cancer. Two features, cystic content and spongiform appearance, when present, might rule out malignancy. Unfortunately, this has limited applicability to clinical practice due to the infrequent occurrence of these characteristics. Nevertheless, clinicians should still use US features to determine a pretest probability of malignancy to identify the patients, who are most likely to benefit from biopsy and further analysis.

## Acknowledgments

The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official view of National Institutes of Health.

Address all correspondence and requests for reprints to: Juan P. Brito, MD, Knowledge and Evaluation Research Unit, Division of Diabetes, Metabolism, and Nutrition, Mayo Clinic, 200 First Street SW, Rochester, MN 55905. E-mail: juan.brito@mayo.edu.

This work was supported by Clinical and Translational Science Award Grant UL1 TR000135 from the National Center for Advancing Translational Sciences, a component of the National Institutes of Health. M.C. has received grants from Siemens Medical Systems and Thermedical, Inc.

Disclosure Summary: The authors have nothing to declare.

## References

1. **Tan GH, Gharib H.** Thyroid incidentalomas: management approaches to nonpalpable nodules discovered incidentally on thyroid imaging. *Ann Intern Med*. 1997;126:226–231.
2. **Reiners C, Wegscheider K, Schicha H, et al.** Prevalence of thyroid disorders in the working population of Germany: ultrasonography screening in 96 278 unselected employees. *Thyroid*. 2004;14:926–932.
3. **Mortensen JD, Woolner LB, Bennett WA.** Gross and microscopic findings in clinically normal thyroid glands. *J Clin Endocrinol Metab*. 1955;15:1270–1280.
4. **Mazzaferri EL.** Managing small thyroid cancers. *JAMA*. 2006;295:2179–2182.
5. **Hegedus L.** Clinical practice. The thyroid nodule. *N Engl J Med*. 2004;351:1764–1771.
6. **Mandel SJ.** A 64-year-old woman with a thyroid nodule. *JAMA*. 2004;292:2632–2642.
7. **Cooper DS, Doherty GM, Haugen BR, et al.** Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. *Thyroid*. 2009;19:1167–1214.
8. **Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y.** Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. Version 10. The Cochrane Collaboration, 2010: Chapter 10. Available from: http://www.cochrane-handbook.org. Accessed November 2012.
9. **Moher D, Liberati A, Tetzlaff J, Altman DG, The PG.** Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *PLOS Med*. 2009;6(7):e1000097.
10. **Frates MC, Benson CB, Charboneau JW, et al.** Management of thyroid nodules detected at US: Society of Radiologists in Ultrasound consensus conference statement. *Radiology*. 2005;237:794–800.
11. **Deville WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, Bezemer PD.** Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Med Res Methodol*. 2002;2:9.
12. **Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM.** The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*. 2003;56:1129–1135.
13. **Whiting PF, Rutjes AW, Westwood ME, et al.** QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155:529–536.
14. **Moon WJ, Baek JH, Jung SL, et al.** Ultrasonography and the ultrasound-based management of thyroid nodules: consensus statement and recommendations. *Korean J Radiol*. 2011;12:1–14.
15. **Mullan RJ, Flynn DN, Carlberg B, et al.** Systematic reviewers commonly contact study authors but do so with limited rigor. *J Clin Epidemiol*. 2009;62:138–142.
16. **DerSimonian R, Laird N.** Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177–188.
17. **Altman DG, Bland JM.** Interaction revisited: the difference between two estimates. *BMJ*. 2003;326:219.
18. **Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I.** The case of the misleading funnel plot. *BMJ*. 2006;333:597–600.
19. **Lou L, Cong XL, Yu GF, Li JC, Ma YX.** US findings of bilateral primary breast cancer: retrospective study. *Eur J Radiol*. 2007;61:154–157.
20. **Goldstein RE, Netterville JL, Burkey B, Johnson JE.** Implications of follicular neoplasms, atypia, and lesions suspicious for malignancy diagnosed by fine-needle aspiration of thyroid nodules. *Ann Surg*. 2002;235:656–662; discussion 662–664.
21. **LiVolsi VA, Asa SL.** The demise of follicular carcinoma of the thyroid gland. *Thyroid*. 1994;4:233–236.
22. **Begg CB.** Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411–423.
23. **Jun P, Chow LC, Jeffrey RB.** The sonographic features of papillary thyroid carcinomas: pictorial essay. *Ultrasound Q*. 2005;21:39–45.
24. **Gardner IA, Stryhn H, Lind P, Collins MT.** Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev Vet Med*. 2000;45:107–122.
25. **Brenner H.** How independent are multiple 'independent' diagnostic classifications? *Stat Med*. 1996;15:1377–1386.
26. **Benndorf M.** Conditional non-independence of radiographic image features and the derivation of post-test probabilities—a mammography BI-RADS example. *Radiography*. 2012;18:201–205.
27. **Gharib H, Papini E, Paschke R, et al.** American Association of Clinical Endocrinologists, Associazione Medici Endocrinologi, and European Thyroid Association medical guidelines for clinical practice for the diagnosis and management of thyroid nodules: executive summary of recommendations. *J Endocrinol Invest*. 2010;33:51–56.
28. **Degirmenci B, Haktanir A, Albayrak R, et al.** Sonographically guided fine-needle biopsy of thyroid nodules: the effects of nodule characteristics, sampling technique, and needle size on the adequacy of cytological material. *Clin Radiol*. 2007;62:798–803.
29. **Atli M, Akgul M, Saryal M, Daglar G, Yasti AC, Kama NA.** Thyroid incidentalomas: prediction of malignancy and management. *Int Surg*. 2006;91:237–244.
30. **Brkljacic B, Cuk V, Tomic-Brzac H, Bence-Zigman Z, Delic-Brkljacic D, Drinkovic I.** Ultrasonic evaluation of benign and malignant nodules in echographically multinodular thyroids. *J Clin Ultrasound*. 1994;22(2):71–76.
31. **Brunese L, Romeo A, Iorio S, et al.** Thyroid B-flow twinkling sign: a new feature of papillary cancer. *Eur J Endocrinol*. 2008;159:447–451.
32. **Cappelli C, Castellano M, Pirola I, et al.** Thyroid nodule shape suggests malignancy. *Eur J Endocrinol*. 2006;155:27–31.
33. **Chen G, Zhu XQ, Zou X, et al.** Retrospective analysis of thyroid nodules by clinical and pathological characteristics, and ultrasonographically detected calcification correlated to thyroid carcinoma in South China. *Eur Surg Res*. 2009;42:137–142.
34. **Choi YJ, Yun JS, Kim DH.** Clinical and ultrasound features of cytology diagnosed follicular neoplasm. *Endocr J*. 2009;56(3):383–389.
35. **Chung J, Youk JH, Kim JA, et al.** Initially non-diagnostic ultrasound-guided fine needle aspiration cytology of thyroid nodules: value and management. *Acta Radiol*. 2012;53:168–173.
36. **Gulcelik NE, Gulcelik MA, Kuru B.** Risk of malignancy in patients with follicular neoplasm: predictive value of clinical and ultrasonographic features. *Arch Otolaryngol Head Neck Surg*. 2008;134:1312–1315.
37. **Hong Y, Wu Y, Luo Z, Wu N, Liu X.** Impact of nodular size on the predictive values of gray-scale, color-Doppler ultrasound, and sonoelastography for assessment of thyroid nodules. *J Zhejiang Univ Sci*. 2012;13:707–716.
38. **Kakkos SK, Scopa CD, Chalmoukis AK, et al.** Relative risk of cancer in sonographically detected thyroid nodules with calcifications. *J Clin Ultrasound*. 2000;28:347–352.
39. **Kim EK, Park CS, Chung WY, et al.** New sonographic criteria for recommending fine-needle aspiration biopsy of nonpalpable solid nodules of the thyroid. *AJR Am J Roentgenol*. 2002;178:687–691.
40. **Kim JY, Lee CH, Kim SY, et al.** Radiologic and pathologic findings

of nonpalpable thyroid carcinomas detected by ultrasonography in a medical screening center. *J Ultrasound Med*. 2008;27:215–223.
41. **Kwak JY, Kim EK, Kim MJ, Son EJ.** Significance of sonographic characterization for managing subcentimeter thyroid nodules. *Acta Radiol*. 2009;50:917–923.
42. **Lee YH, Kim DW, In HS, et al.** Differentiation between benign and malignant solid thyroid nodules using an US classification system. *Korean J Radiol*. 2011;12:559–567.
43. **Leenhardt L, Menegaux F, Franc B, et al.** Selection of patients with solitary thyroid nodules for operation. *Eur J Surg*. 2002;168:236–241.
44. **Mendelson AA, Tamilia M, Rivera J, et al.** Predictors of malignancy in preoperative nondiagnostic biopsies of the thyroid. *J Otolaryngol Head Neck Surg*. 2009;38:395–400.
45. **Mendez W, Rodgers SE, Lew JI, Montano R, Solorzano CC.** Role of surgeon-performed ultrasound in predicting malignancy in patients with indeterminate thyroid nodules. *Ann Surg Oncol*. 2008;15:2487–2492.
46. **Moon HJ, Kwak JY, Choi YS, Kim EK.** How to manage thyroid nodules with two consecutive non-diagnostic results on ultrasonography-guided fine-needle aspiration. *World J Surg*. 2012;36:586–592.
47. **Moon WJ, Jung SL, Lee JH, et al.** Benign and malignant thyroid nodules: US differentiation—multicenter retrospective study. *Radiology*. 2008;247:762–770.
48. **Ozel A, Erturk SM, Ercan A, et al.** The diagnostic efficiency of ultrasound in characterization for thyroid nodules: how many criteria are required to predict malignancy? *Med Ultrason*. 2012;14:24–28.
49. **Papini E, Guglielmi R, Bianchini A, et al.** Risk of malignancy in nonpalpable thyroid nodules: predictive value of ultrasound and color-Doppler features. *J Clin Endocrinol Metab*. 2002;87:1941–1946.
50. **Phuttharak W, Somboonporn C, Hongdomnern G.** Diagnostic performance of gray-scale versus combined gray-scale with colour Doppler ultrasonography in the diagnosis of malignancy in thyroid nodules. *Asian Pac J Cancer Prev*. 2009;10:759–764.
51. **Popowicz B, Klencki M, Lewinski A, Slowinska-Klencka D.** The usefulness of sonographic features in selection of thyroid nodules for biopsy in relation to the nodule's size. *Eur J Endocrinol*. 2009;161(1):103–111.
52. **Rago T, Coscio GD, Basolo F, et al.** Combined clinical, thyroid ultrasound and cytological features help to predict thyroid malignancy in follicular and Hupsilonrthle cell thyroid lesions: results from a series of 505 consecutive patients. *Clin Endocrinol (Oxf)*. 2007;66:13–20.
53. **Sahin M, Gursoy A, Tutuncu NB, Guvener DN.** Prevalence and prediction of malignancy in cytologically indeterminate thyroid nodules. Clin *Endocrinol (Oxf)*. 2006;65:514–518.
54. **Salmaslioglu A, Erbil Y, Dural C, Issever H, et al.** Predictive value of sonographic features in preoperative evaluation of malignant thyroid nodules in a multinodular goiter. *World J Surg*. 2008;32:1948–1954.
55. **Schueller-Weidekamm C, Schueller G, Kaserer K, et al.** Diagnostic value of sonography, ultrasound-guided fine-needle aspiration cytology, and diffusion-weighted MRI in the characterization of cold thyroid nodules. *Eur J Radiol*. 2010;73:538–544.
56. **Sharma A, Gabriel H, Nemcek AA, Nayar R, Du H, Nikolaidis P.** Subcentimeter thyroid nodules: utility of sonographic characterization and ultrasound-guided needle biopsy. *AJR Am J Roentgenol*. 2011;197:W1123–W1128.
57. **Solbiati L, Volterrani L, Rizzatto G, et al.** The thyroid gland with low uptake lesions: evaluation by ultrasound. *Radiology*. 1985;155:187–191.
58. **Yoon JH, Kwak JY, Moon HJ, Kim MJ, Kim EK.** The diagnostic accuracy of ultrasound-guided fine-needle aspiration biopsy and the sonographic differences between benign and malignant thyroid nodules 3 cm or larger. *Thyroid*. 2011;21:993–1000.
59. **Yoon JH, Kwak JY, Kim EK, et al.** How to approach thyroid nodules with indeterminate cytology. *Ann Surg Oncol*. 2010;17:2147–2155.