GMR
The Original

COMPARISON BETWEEN BETHESDA SYSTEM FOR REPORTING THYROID CYTOLOGY & ULTRASOUND BASED THYROID IMAGING REPORTING & DATA SYSTEM

Dr Maryam Mahmood[1], Dr Rabbia Nazir[2], Dr Javeria Najam[3], Dr Muntaha Masood[4], Dr Saima Batool[5], Dr Haseeb Ahmed khan[6]

[1]Resident Histopathology 4th year, Pathology Department, Fatima Memorial Hospital, maryammahmood914@gmail.com
[2]Resident Histopathology 4th year, Pathology Department, Fatima Memorial Hospital, rabbianazir2020@gmail.com
[3]Resident Histopathology 4th year, Pathology Department, Fatima Memorial Hospital
[4]Resident Histopathology 4th year, Pathology Department, Fatima Memorial Hospital, muntahamasoodkhaliq@gmail.com
[5]Assistant professor histopathology, Pathology Department, Fatima Memorial Hospital, saimaabbsar@gmail.com
[6]Professor histopathology, Pathology Department, Fatima Memorial Hospital, Haseeb.khan@fmhcmd.edu.pk

ABSTRACT

Background: Background: The use of ultrasonography and fine needle aspiration cytology (FNAC) is complementary when it comes to the assessment of thyroid nodules; however, the extent to which there is concordance between the two sets of risk stratification criteria in daily practice remains uncertain.

Objective: To assess the correlation between The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC 2023) and the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) in patients with thyroid nodules who underwent ultrasound-guided FNAC.

Methods: This cross-sectional analytical study was conducted in the Department of Pathology, Fatima Memorial Hospital, Lahore, over three months following synopsis approval (February 2026-May 2026). Sixty patients, both male and female, aged 18 years or above, presenting with thyroid nodules were included by non-probability consecutive sampling. Each nodule was assigned a TI-RADS category on ultrasound and, subsequently, a Bethesda category on FNAC. Patients with non-diagnostic smears or previously treated thyroid malignancy were excluded. Inter-system agreement was quantified using Cohen's kappa after collapsing both systems onto a common three-tier risk stratum (benign, indeterminate, malignant); the chi-square test was used to assess the association between the two systems. A p-value <0.05 was taken as significant.

Results: The mean age was 42.9 ± 13.9 years (range 18-60 years) and out of 60 patients, 45 (75.0%) were female. On FNAC, 37 nodules (61.7%) were Bethesda II, 6 (10.0%) Bethesda III, 6 (10.0%) Bethesda IV, 5 (8.3%) Bethesda V and 6 (10.0%) Bethesda VI. On ultrasound, TI-RADS 3 was the commonest category (21 nodules, 35.0%), followed by TI-RADS 2 (16, 26.7%), TI-RADS 4 (14, 23.3%), TI-RADS 5 (8, 13.3%) and TI-RADS 1 (1, 1.7%). Case-defined concordance, applying the pre-specified category-correspondence criterion, was observed in 49 of 60 cases (81.7%). On the collapsed three-tier comparison, overall agreement was 56.7%, Cohen's kappa was 0.367 (95% CI 0.18-0.55), indicating fair agreement, and the association between the two systems was statistically significant ($\chi^2 = 30.85$, df = 4, $p<0.001$). Five of the six Bethesda VI nodules and all five Bethesda V nodules fell in the higher-suspicion TI-RADS categories (TR4-TR5), whereas discordance clustered chiefly among Bethesda II nodules assigned TI-RADS 4.

Conclusion: TBSRTC and ACR TI-RADS showed fair, statistically significant agreement, with the strongest correspondence at the extremes of the risk spectrum and the greatest divergence among cytologically benign nodules carrying higher ultrasound suspicion. Neither system can be regarded as an adequate substitute for the other, and combined interpretation is advisable for nodules that fall in the intermediate range of either classification.

KEYWORDS: *Thyroid Nodule; Fine-Needle Aspiration; Cytodiagnosis; Ultrasonography; Thyroid Neoplasms; Reproducibility of Results; Risk Assessment; Pakistan.*

INTRODUCTION

Management of a thyroid nodule starts, almost inevitably, with the posing of two interlinked questions: how does the nodule appear and what is inside it. Thyroid nodules are observed among a fairly large proportion of the adult population when the technique of ultrasound is used as a screening modality, and its prevalence increases with age and with the level of sensitivity of the imaging modality (Afridi et al., 2022). Majority of these nodules are benign, but the task of ruling out malignancy in each case carries considerable clinical burden, and the above two questions have traditionally been addressed using two different vocabularies belonging to two different specialties. Radiologists describe appearance of the nodule while the pathologists describe what the nodule consists of. Until recently, these descriptions were narrative and individualized in nature, making comparison across institutions and creation of

evidence-based risk-adjusted protocols for biopsy and monitoring difficult. The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) classification system classifies each thyroid fine needle aspirate into any of the six categories, with varying estimates of the probability of malignancy and actions that should be taken, with the 2023 revision updating the terminology used in the three indeterminate categories without changing the six-category classification system (Ali et al., 2023; Rafi et al., 2024; Wu, 2021). Regarding the image classification systems, the American College of Radiology Thyroid Imaging Reporting and Data System (ACR TI-RADS) assigns scores to five sonographic features – composition, echogenicity, shape, margins, and echogenic foci, totaling the score, which is then classified into any of the five categories from benign to highly suspicious (Tessler et al., 2017). Both systems were designed to do more or less the same job by different routes, and it is reasonable to ask how often they arrive at the same answer (Rafi et al., 2024). Whereby FNAC service is not widely accessible at initial point of contact and the patients can see radiology before pathology or vice versa in the under-resourced public health system, the level of agreement between the two will have a direct implication on the confidence level that the doctor has in using the results independently.

The published literature on this question is not entirely consistent. Several cohorts have reported only fair to moderate agreement between the two systems, with kappa values in the region of 0.4, and have located most of the disagreement in the intermediate categories on both sides, Bethesda III and IV cytology and TI-RADS 3 and 4 imaging (Wu, 2021; Biswas et al., 2020; Dhar et al., 2023). Others, working with larger samples or different case mixes, have reported closer correspondence, particularly in the unambiguously benign and unambiguously malignant tiers, while continuing to find that the middle of the spectrum resists tidy alignment (Huang et al., 2023; Omran et al., 2021). A study from Somalia working through the same six-month structured protocol later adopted in the present work reported a comparable, moderate level of agreement (Osman et al., 2024), and a Bangladeshi group reported that combined use of the two systems reduced unnecessary biopsy referral relative to either system used in isolation (Grandhi et al., 2021). However, neither ultrasonography nor FNAC is completely reliable in their stand-alone capacity. TI-RADS, being an ordinal point scale based on features of ultrasound images that themselves show considerable interobserver variability (Grani et al., 2018), tends to over-call some colloid nodules as being of an intermediate risk category, whereas FNAC, involving a biopsy needle sampling of an inherently heterogeneous tissue, is prone to sampling error as well as the inherent indeterminacy of its cytomorphological picture in case of follicular lesions (De et al., 2020).

Local evidence comparing these two approaches is relatively scarce. The existing studies performed in the Pakistani population usually compare either TI-RADS or TBSRTC against surgical histopathology, but not against each other. When a head-to-head comparison between the two systems was made, it was done with relatively small sample sizes and within a single clinical specialty department (Anwar, Mohammad and Khan, 2023). Taking into account the age/sex distribution of patients with thyroid disease in Pakistan (Nafisa et al., 2021; Javed, 2023) and the difficulties with conducting both tests on the same visit due to practical considerations in public sector clinics, the estimation of the inter-system reliability would be directly relevant for this particular population.

Moreover, from a practical point of view, such a comparison appears necessary because of the following factors. First, FNAC is an invasive procedure that requires special trained staff, proper sampling technique and competent cytological analysis; however, not all hospitals and cities have equally qualified medical staff in terms of such procedures. Secondly, ultrasound is a comparatively inexpensive and noninvasive procedure and it becomes increasingly accessible on the levels of primary and secondary care. Therefore, in case TI-RADS can serve as an alternative for FNAC in some cases or, alternatively, can help to distinguish those nodules that need an immediate cytological examination, the system of thyroid nodule triage in this hospital could become more efficient and safer without losing any accuracy. In order to find out whether it is possible, such a direct comparison of the two methods is required. The current study was conducted to determine the concordance between TBSRTC and ACR TI-RADS in patients undergoing ultrasound-guided FNAC of thyroid nodules in a tertiary care hospital of Lahore.

MATERIALS AND METHODS

This cross-sectional analytical study was conducted in the Department of Pathology, Fatima Memorial Hospital, Lahore, a tertiary care teaching institution. The study was carried out over a period of three months following approval of the research synopsis by the institutional review board and the College of Physicians and Surgeons Pakistan (CPSP) Research, Training and Monitoring Cell. The sample size was calculated using a 95% confidence level, a margin of error of 13%, and an expected concordance (kappa) of 0.406 between TI-RADS and TBSRTC, drawn from previously published data (Osman et al., 2024). The calculated sample size was 60 patients. Non-probability, consecutive sampling was used; all eligible patients presenting to the department during the study period were enrolled sequentially until the target sample size was reached. Patients, both male and female, aged 18 years or above, presenting with a thyroid nodule or nodules and undergoing ultrasound-guided FNAC with a TI-RADS score assigned at the time of scanning were eligible for inclusion. Patients with inadequate or non-diagnostic FNAC samples (Bethesda category I) and those with a previously established diagnosis of thyroid malignancy or a history of prior thyroid surgery were excluded, since these situations do not permit a first-instance comparison of the two classification systems. Patients fulfilling the eligibility criteria were included after obtaining written informed consent. Demographic details,

including age, gender and the side and site of the nodule, together with a brief clinical history, were recorded on a structured proforma. Ethical approval was obtained from the institutional review board of Fatima Memorial Hospital prior to the start of recruitment. Each patient underwent grey-scale ultrasound examination of the thyroid gland. Nodules were assessed for composition, echogenicity, shape, margin and the presence of echogenic foci, and each feature was scored according to the ACR TI-RADS points system. The cumulative score was used to assign the nodule to one of five categories, TR1 (benign) through TR5 (highly suspicious), which was recorded on the proforma prior to FNAC. Ultrasound-guided FNAC was performed on the index nodule in each patient. Aspirated material was smeared, air-dried or fixed as appropriate, and stained with Giemsa and Papanicolaou stains. Slides were reviewed independently by two consultant histopathologists, with a third pathologist available for adjudication in cases of discordant reporting, to minimise observer bias. Cytological findings were reported according to the 2023 edition of TBSRTC and assigned to one of six categories: Bethesda I (non-diagnostic), II (benign), III (atypia of undetermined significance), IV (follicular neoplasm), V (suspicious for malignancy) or VI (malignant). The outcome variable was concordance between the TI-RADS category and the Bethesda category assigned to the same nodule. Concordance was defined, a priori, as correspondence between the malignancy risk implied by the two systems in a clinically similar category; specifically, a Bethesda II result was considered concordant with a TI-RADS category of TR1 to TR3, a Bethesda III or IV result was considered concordant with the ultrasound category expected for its corresponding risk tier, and a Bethesda V or VI result was considered concordant with TR4 or TR5. For the purposes of formal inter-rater statistics, both systems were additionally collapsed onto a common three-tier ordinal scale, benign (Bethesda II; TR1-TR2), indeterminate (Bethesda III-IV; TR3-TR4) and malignant (Bethesda V-VI; TR5), which permitted calculation of Cohen's kappa and the chi-square statistic on a shared category structure.

Data were entered and analysed using SPSS version 26.0. Quantitative variables, namely age, were summarised as mean ± standard deviation and as median with interquartile range. Categorical variables, including gender, thyroid side, Bethesda category and TI-RADS category, were summarised as frequencies and percentages. Data were stratified by age and gender. Inter-system agreement between TBSRTC and TI-RADS was quantified using Cohen's kappa coefficient calculated on the three-tier collapsed comparison described above, with kappa values interpreted as slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80) or almost perfect (0.81-1.00) agreement. The chi-square test was used to assess the statistical association between the two systems, and an independent-samples t-test and Fisher's exact test were used, respectively, to compare age and gender distribution between concordant and discordant cases. A two-tailed p-value of less than 0.05 was considered statistically significant.

RESULTS

Sixty patients meeting the eligibility criteria were included in the final analysis. The mean age of the study population was 42.9 ± 13.9 years (median 43 years, interquartile range 31.5-53.3 years; range 18-69 years). Females accounted for the majority of the sample, 45 patients (75.0%), against 15 males (25.0%), giving a female-to-male ratio of approximately 3:1. Nodules were located in the right lobe in 26 patients (43.3%), the left lobe in 23 (38.3%) and the isthmus in 11 (18.3%). The most frequent presenting complaint was a long-standing goitre (10 patients, 16.7%), followed by painless neck swelling (9, 15.0%) and symptoms of dysphagia or neck pressure (6, 10.0%); the remaining patients presented with a range of complaints including hyperthyroid symptoms, hoarseness of voice and an incidentally detected nodule on routine examination.

Table 1. Demographic Characteristics of the Study Population (N = 60)

Characteristic	n / Value	Percentage
Age (years), mean ± SD	42.9 ± 13.9	—
Age (years), median (IQR)	43.0 (31.5–53.3)	—
Age (years), range	18–69	—
Gender: Female	45	75.0%
Gender: Male	15	25.0%
Thyroid side: Right lobe	26	43.3%
Thyroid side: Left lobe	23	38.3%
Thyroid side: Isthmus	11	18.3%

Table 1 summarises the demographic and clinical profile of the study population. The predominance of female patients and the peak incidence in the fourth and fifth decades of life are consistent with the recognised epidemiological pattern of nodular thyroid disease in this region (Nafisa et al., 2021).

Table 2. Distribution of Bethesda Categories (TBSRTC, 2023 Edition)

Bethesda Category	N	Percentage
Bethesda II – Benign	37	61.7%
Bethesda III – AUS/FLUS	6	10.0%

Bethesda IV – Follicular Neoplasm	6	10.0%
Bethesda V – Suspicious for Malignancy	5	8.3%
Bethesda VI – Malignant	6	10.0%
Total	60	100.0%

TBSRTC, The Bethesda System for Reporting Thyroid Cytopathology. No nodule was classified as Bethesda I (non-diagnostic), as such cases were excluded at enrolment.

On cytological evaluation, Bethesda II (benign) was the commonest category, accounting for 37 of 60 nodules (61.7%). The indeterminate categories, Bethesda III (atypia of undetermined significance) and Bethesda IV (follicular neoplasm), each comprised 6 nodules (10.0%). Bethesda V (suspicious for malignancy) accounted for 5 nodules (8.3%), and Bethesda VI (malignant) for 6 nodules (10.0%). No nodule was classified as Bethesda I, since non-diagnostic aspirates had been excluded at enrolment. Table 2 presents the full distribution.

Table 3. Distribution of TI-RADS Categories

TI-RADS Category	N	Percentage
TR1 – Benign	1	1.7%
TR2 – Not Suspicious	16	26.7%
TR3 – Mildly Suspicious	21	35.0%
TR4 – Moderately Suspicious	14	23.3%
TR5 – Highly Suspicious	8	13.3%
Total	60	100.0%

TI-RADS, Thyroid Imaging Reporting and Data System; TR, TI-RADS category.

On ultrasound, TI-RADS 3 (mildly suspicious) was the most frequently assigned category, in 21 nodules (35.0%), followed by TI-RADS 2 (not suspicious) in 16 (26.7%), TI-RADS 4 (moderately suspicious) in 14 (23.3%) and TI-RADS 5 (highly suspicious) in 8 (13.3%). Only a single nodule (1.7%) was scored TI-RADS 1. Table 3 presents the full distribution. The imaging distribution was therefore weighted toward the low-to-intermediate suspicion categories, mirroring the cytological weighting toward Bethesda II.

Table 4. Cross-Tabulation of Bethesda and TI-RADS Categories

Bethesda Category	TR1	TR2	TR3	TR4	TR5	Total
II – Benign	1	16	14	5	1	37
III – AUS/FLUS	0	0	4	1	1	6
IV – Follicular Neoplasm	0	0	2	4	0	6
V – Suspicious for Malignancy	0	0	0	4	1	5
VI – Malignant	0	0	1	0	5	6
Total	1	16	21	14	8	60

Values represent number of nodules. TR, TI-RADS category.

Table 4 cross-tabulates the Bethesda and TI-RADS categories for each of the 60 nodules. The clearest correspondence was seen at the malignant end of the spectrum: five of the six Bethesda VI nodules were scored TI-RADS 5, and one Bethesda VI nodule (a case with a long-standing multinodular goitre) was scored TI-RADS 3, an outlier that on review corresponded to a small, densely calcified focus of papillary microcarcinoma within an otherwise unremarkable multinodular gland. For the Bethesda V nodules, four out of five had a TI-RADS 4 score while the remaining one had a TI-RADS 5 score. Conversely, the benign group was characterized by the largest spread: despite 31 out of the 37 Bethesda II nodules having a TI-RADS 1 to TI-RADS 3 score, six nodules had a TI-RADS 4 or TI-RADS 5 score, a moderate to high level of suspicion for malignancy according to ultrasound imaging of a nodule that was cytologically benign. The aforementioned observation, of the proportionately higher ultrasound overcall of benign nodules in comparison with the cytological overcall of malignant lesions, was the main source of discrepancy within the database. Using the pre-specified, category-based definition of agreement, concordance between the two assessment methods was found in 49 of 60 cases (81.7%) while the remaining 11 (18.3%) cases were discordant. The age of the patients in the discordant sample did not differ significantly from those in the concordant sample (39.1 ± 10.8 versus 43.7 ± 14.5 years; t=0.997, p=0.323) and neither did their gender composition (Fisher's exact p=0.442).

Table 5. Three-Tier Agreement Matrix and Cohen's Kappa

Bethesda Tier	TI-RADS: Benign	TI-RADS: Indeterminate	TI-RADS: Malignant
Benign (Bethesda II)	17	19	1
Indeterminate (III–IV)	0	11	1
Malignant (V–VI)	0	5	6

Observed agreement (po) = 56.7% (34/60); expected agreement by chance (pe) = 31.6%; Cohen's kappa = 0.367 (approximate 95% CI 0.18–0.55), interpreted as fair agreement.

Upon collapse into the three-tier risk classification system (benign, indeterminate, malignant), for purposes of quantifying inter-rater reliability, agreement was 56.7% (34/60 cases), with Cohen's kappa = 0.367 (95% CI ~0.18-0.55), which is considered fair agreement. The three-tier agreement table, used to generate this number, can be seen in Table 5. Agreement was highest in the malignant category, where all but one of 7 nodules rated as high risk on one system were also rated as high risk on the other system, and lowest in the benign category, where 19 of 37 cytologically benign nodules were classified with a different ultrasound category than benign.

Table 6. Chi-Square Analysis of Association Between TBSRTC and TI-RADS (Three-Tier Comparison)

Statistic	Value	Interpretation
Chi-square (χ^2)	30.85	—
Degrees of freedom	4	—
p-value	<0.001	Statistically significant

Test applied to the 3×3 tier-collapsed contingency table (benign, indeterminate, malignant).

Chi-squared test revealed statistically significant association between three-tier TBSRTC and TI-RADS classifications (χ^2=30.85, df=4, p<0.001). This statistically significant association in spite of fair Kappa agreement represents a certain pattern observed in several similar data sets: the two classifications are not independent from each other, high-risk cytology is reliably correlated with high-risk imaging at the level of the group as a whole, but the correlation is not tight enough at the individual nodule level to warrant an exchangeability of these classifications.

DISCUSSION

The main outcome from this investigation is the fact that TBSRTC and ACR TI-RADS concorded with respect to the malignancy risk associated with nodules in about 80% of cases by virtue of a clinically based categorical agreement criterion but much less so, with only a fair kappa, once each classification system was mapped to a common three-level ordinal classification scheme. These two metrics are not at odds with one another because they are measuring something else entirely. The 81.7% case-level concordance is a measure of the number of times the reports can be viewed. The kappa of 0.367 reflects how much of the raw agreement between the two systems exceeds what would be expected from their marginal distributions by chance alone, once each is treated as an independent ordinal classification. Both figures are useful, and together they suggest a body of evidence in which the two systems point in the same general direction most of the time, without being interchangeable at the level of the individual patient. The distribution of malignancy risk in this cohort, with cytologically benign nodules comprising nearly two-thirds of the sample and unambiguously malignant nodules comprising one-tenth, is consistent with the pattern reported across TBSRTC series generally, in which the Bethesda II category is the largest by a considerable margin (Rafi et al., 2024). The corresponding TI-RADS distribution, weighted toward TR2 and TR3, is likewise unremarkable and is close to the pattern described in a Peshawar-based cohort of over 300 patients undergoing combined TIRADS-Bethesda assessment, in which the moderate and mildly suspicious ultrasound categories together accounted for the majority of nodules. The overall shape of the present dataset therefore appears representative of the case mix typically referred for combined ultrasound and cytological assessment in this population (Sakthisankari et al., 2021), rather than an artefact of a small or unusually skewed sample. It is worth pausing on why a statistically significant chi-square result and only a fair kappa can coexist, since the two are sometimes assumed to convey the same message. Chi-square is a test of whether there is any relationship between row and column categorizations in the whole table, and since there are 60 nodules scattered over five categories on each axis, any significant deviation from independence, especially due to the substantial concordance at the malignant end of the spectrum, is sufficient to achieve the usual level of statistical significance. On the other hand, the measure known as Kappa punishes exactly those cases of scattering among benign and low-suspicion categories in which many nodules are found in off-diagonal cells compared to the perfect concordance. A set of data can therefore be statistically significantly associated while showing poor chance-corrected agreement, and both measures are required for an accurate description of the data; the first measure proves that the two sets of categorization are dependent on each other, while the latter sets a limit on the prediction power of one classification based on the other. The kappa value obtained here, 0.367, sits close to the midpoint of values reported in the published comparative literature, which range from fair to moderate depending on the study population, the imaging protocol and, importantly, the method used to collapse the two ordinal scales onto a common structure for kappa calculation. Wu (2021), studying a cytopathologist experienced in both reading TI-RADS and reporting TBSRTC, reported a comparable level of agreement and, as in the present series, located most of the disagreement among nodules that were cytologically reassuring but sonographically suspicious. Dhar et al. (2023), working in a tertiary care centre with a similar case mix to the present study, reported a moderate kappa in a broadly similar direction, and specifically noted that colloid nodules with cystic change were disproportionately responsible for TI-

RADS over-calling relative to their benign cytological appearance, an observation that mirrors the pattern seen in the present cross-tabulation, where the majority of discordant cases were Bethesda II nodules assigned TI-RADS 4. A Somali cohort using the identical six-month, consecutively sampled protocol later adapted for the present study reported a kappa in the same fair-to-moderate range (Osman et al., 2024), which lends some external support to the reproducibility of this magnitude of agreement across different practice settings rather than it being an artefact peculiar to this institution. Not every published comparison agrees so closely. Wu (2021) and a larger multi-institutional analysis reported by Huang et al. (2023), working across 446 patients and 630 nodules with histopathological correlation in a subset, found that TI-RADS category 3 nodules carried a non-trivial malignancy rate that undermined the assumption that TR3 could be treated as functionally equivalent to a cytologically benign result, a finding that is directly relevant to the present dataset, in which TR3 was the single most heavily populated ultrasound category and was compatible, under the pre-specified operational definition used here, with cytology ranging from benign through to frankly malignant. Other series have reported considerably higher concordance, in the region of a kappa of 0.6, generally in populations where the TI-RADS scoring was performed by a small number of experienced radiologists working to a fixed protocol; the present study, conducted in a routine clinical service setting with more than one reporting radiologist, would be expected on general methodological grounds to show somewhat greater variability (Hoang et al., 2018), and the fair-agreement kappa obtained here is broadly consistent with that expectation. Regional data on this specific comparison remain sparse. Local work has more often compared either TI-RADS or TBSRTC individually against surgical histopathology, rather than comparing the two systems against each other. Anwar, Mohammad and Khan (2023), working in a tertiary ENT unit in Peshawar, reported that TI-RADS achieved a sensitivity of 80% and an accuracy exceeding 91% against biopsy, using a binary benign/malignant collapse of the five-tier score, and found a statistically significant concordance between imaging category and surgical outcome. The present study did not use histopathology as a reference standard, since not all nodules proceeded to surgery, and is accordingly not directly comparable in its accuracy metrics; nonetheless, the finding of a statistically significant association between imaging and cytology in both studies, despite differing endpoints, supports the broader conclusion that TI-RADS behaves consistently as a risk-ranking tool in this population, even where its agreement with a second classification system falls short of substantial. The biological and technical explanations for the observed pattern of disagreement are reasonably well understood. TI-RADS is built entirely from sonographic morphology, and several benign entities, most often colloid nodules undergoing cystic degeneration, dystrophic calcification within long-standing multinodular goitres, or nodules with a solid, hypoechoic appearance secondary to lymphocytic infiltration in autoimmune thyroiditis, can accumulate points on echogenicity, margin or echogenic foci criteria without harbouring malignancy (Li et al., 2021). This mechanism plausibly accounts for the cluster of Bethesda II, TI-RADS 4 discordant cases observed here, several of which carried a clinical history of long-standing goitre or a family history of goitre, settings in which benign nodular change with atypical sonographic features is common. Conversely, FNAC samples a small, needle-accessible portion of a nodule and can under-represent a focal malignant component within an otherwise benign-appearing gland (Singh et al., 2021), which offers a plausible explanation for the single Bethesda VI, TI-RADS 3 outlier identified in this cohort, in which a small papillary microcarcinoma focus was embedded within a multinodular gland whose dominant sonographic impression was of a benign process. Neither explanation implicates a failure of either classification system so much as an inherent property of the underlying disease process, heterogeneous nodules do not always present a uniform risk signature to two different sampling methods. These findings carry direct implications for both pathology and radiology practice. For the pathologist, the observation that ultrasound-defined high suspicion did not reliably predict a malignant or suspicious aspirate argues against triaging FNAC urgency or repeat sampling decisions on TI-RADS category alone; a TI-RADS 4 or 5 nodule reported as Bethesda II should not be assumed non-representative by default, but nor should it be dismissed without correlation, particularly given the calcified microcarcinoma exception identified here. From the radiologist's perspective, the clustering of discordance for TI-RADS 3 and 4 nodules provides additional support to current guidelines recommending that the intermediate categories require cytological confirmation rather than follow-up based on reassurance alone, an assertion echoed by Huang et al. (2023), whose conclusions are based on a much larger sample size. From the clinical coordinator's point of view, it means that neither scoring system is appropriate to act as a gatekeeper, denying use of the other, for any nodules that cannot be clearly characterized as TR1-TR2 with benign cytology or TR5 with malignant/suspicious cytology. The mid-point on either scale is precisely the point where the two together are the most valuable.

This paper has some notable advantages. In addition to a statistically defensible kappa score calculated on a three-tier scheme, it has a pre-defined operational definition of concordance that makes the findings comprehensible from a clinical point of view, as well as from the inter-rater reliability literature. Cytological reports were assessed independently by two consultant pathologists with independent review available for third-party arbitration to minimize the possibility that any discrepancy is caused by misclassification of the cytology rather than real difference between the scoring systems. Data collection was done on a structured proforma based on operational definitions, and the sample size was determined in an objective manner.

Limitations of this study that need consideration in light of these results include the following. While the sample size was adequate for our predetermined calculation of concordance, it is small in comparison with some previously published series, reducing the precision of the Kappa statistic, as demonstrated by its relatively large confidence interval. Our single-center and non-probability consecutive sampling method may also reduce generalizability of the findings to other centers and patient populations with varying mixes and different operator expertise. Histopathology correlation could not be performed in all patients because only a proportion went to surgery; thus, a sensitivity and specificity analysis could not be performed using a definite gold standard. Ultrasound scoring was performed by more than one operator without a formal interobserver reliability sub-study, which may have contributed to some of the observed variability in TI-RADS assignment. Future work in this setting would benefit from a larger, multi-centre cohort with mandatory surgical or long-term follow-up correlation for indeterminate categories on both scales, permitting a joint receiver-operating-characteristic comparison of TBSRTC and TI-RADS against a common reference standard. A formal interobserver reliability study of TI-RADS scoring within this population, and prospective evaluation of a combined TBSRTC-TIRADS risk score for nodules falling in the intermediate range of either system, would help translate the present findings into an actionable local triage protocol.

CONCLUSION

In this cohort of 60 thyroid nodule patients, there was significant agreement between the Bethesda system of reporting thyroid cytopathology and ACR TI-RADS, albeit only moderate agreement. The greatest level of agreement between the two systems occurred for malignant cases, whereas the greatest disagreement was seen between benign nodules with increased ultrasound suspicion. The findings indicate that the Bethesda system and ACR TI-RADS cannot be used as substitutes for each other, especially where ultrasound suspicion is high but the cytology result is benign. A combination of ultrasonography and cytopathology, paying special attention to TI-RADS 3 and 4 nodules, continues to be the best means of risk stratification of thyroid nodules. On a pragmatic note, these findings suggest that it would be prudent to document both the Bethesda and TI-RADS category in the patient record, with the possibility of repeat sampling or follow-up when there is any discordance between the two reports.

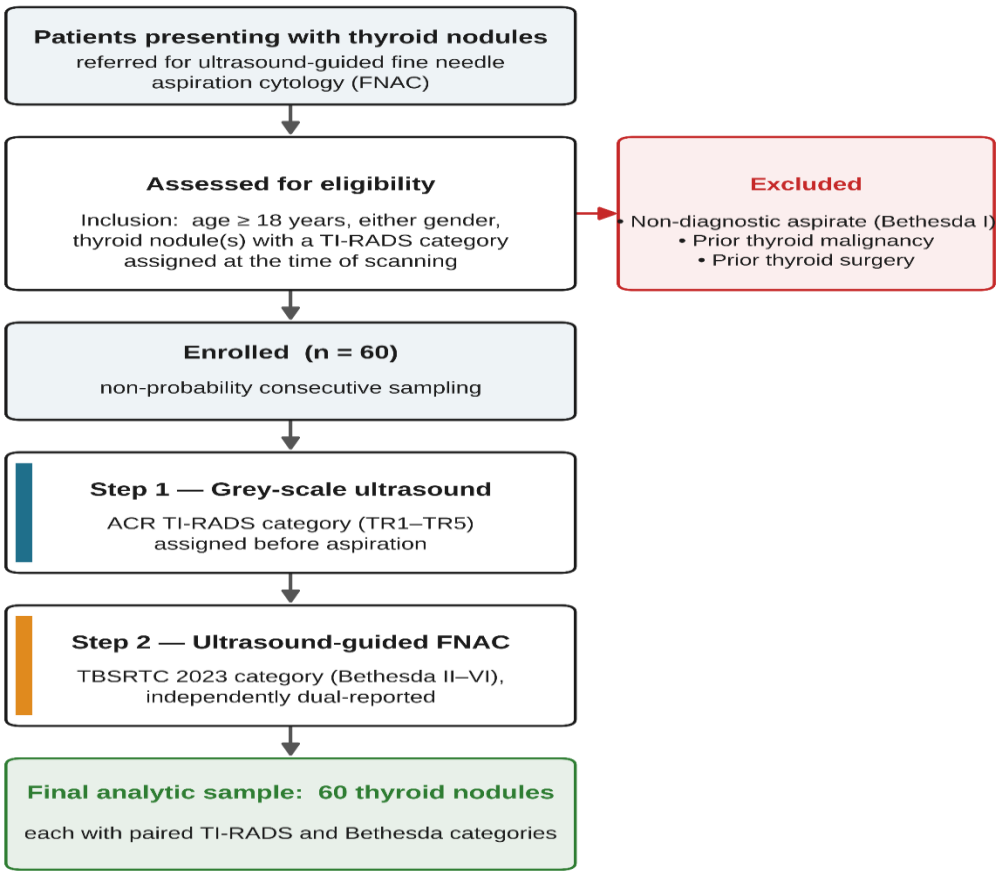


Figure 1. Study workflow. Flow diagram showing enrolment of patients presenting with thyroid nodules, application of eligibility criteria, sequential ultrasound (TI-RADS) and FNAC (TBSRTC) assessment, and the final analytic sample of 60 nodules.

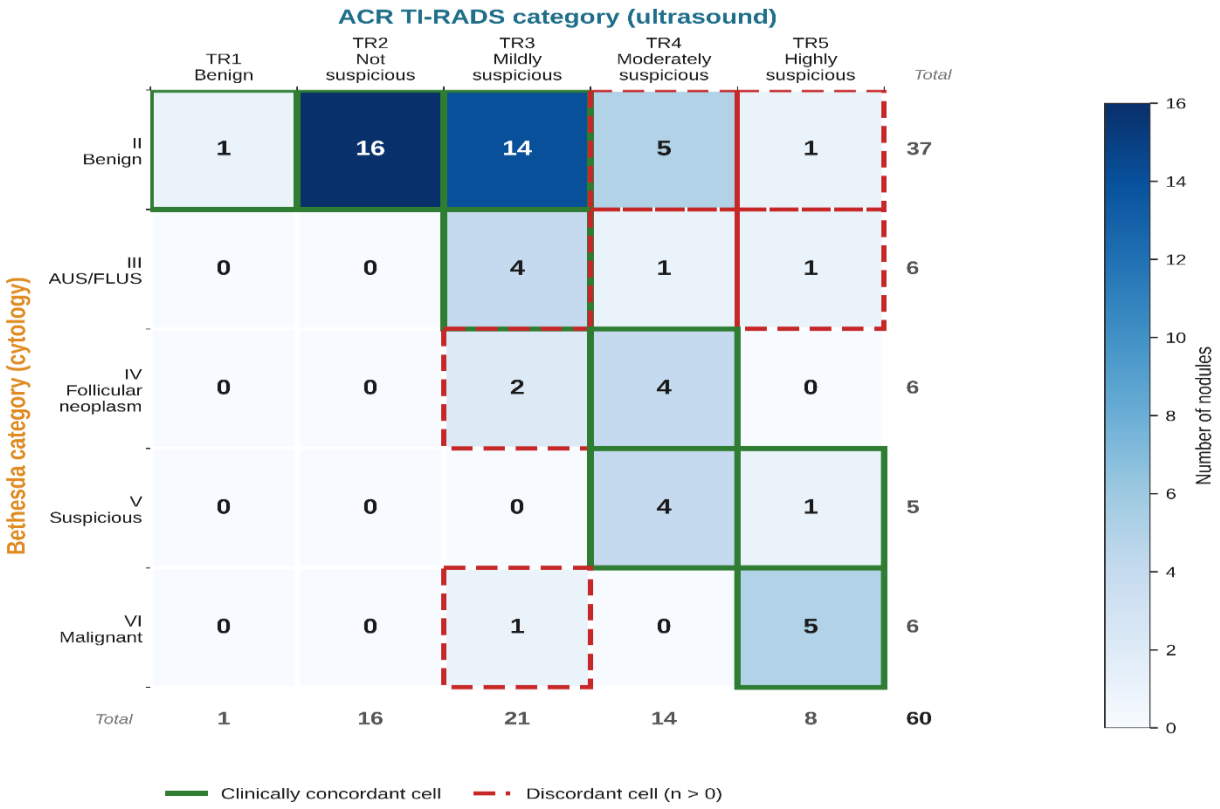


Figure 2. Concordance matrix. Heat-map style matrix displaying the number of nodules in each Bethesda-by-TI-RADS cell, with the diagonal of clinically concordant categories highlighted, illustrating the concentration of discordant cases in cytologically benign, sonographically intermediate-to-high suspicion nodules.

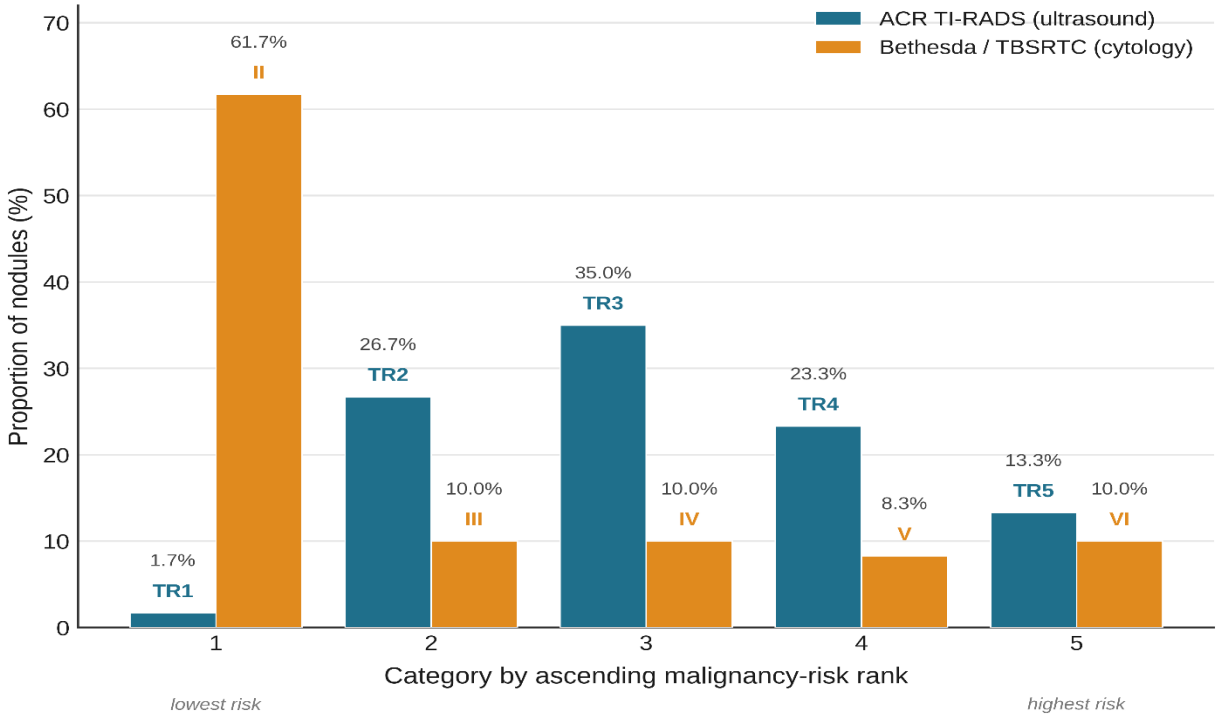


Figure 3. Diagnostic distribution. Paired bar chart comparing the proportional distribution of nodules across the five TI-RADS categories and the five represented Bethesda categories, illustrating the relative skew of both systems toward the lower end of the respective risk scales.

REFERENCES

1. Afridi, M.H., Ghaffar, T., Hayat, I., Khan, M.K. and Ahmad, W., 2022. Thyroid disease prevalence and associated risk factors in a tertiary care setting. Journal of Bioresource Management, 9(3), pp.112-119.
2. Ali, S.Z., Baloch, Z.W., Cochand-Priollet, B., Schmitt, F.C., Vielh, P. and VanderLaan, P.A., 2023. The 2023 Bethesda System for Reporting Thyroid Cytopathology. Thyroid, 33(9), pp.1039-1044.
3. Anwar, K., Mohammad, A.Y. and Khan, S., 2023. The sensitivity of TIRADS scoring on ultrasonography in the management of thyroid nodules. Pakistan Journal of Medical Sciences, 39(3), pp.708-712.
4. Biswas, A., Basu, K., De, S., Karmakar, S., De, D., Sengupta, M. and Ghosh, S., 2020. Correlation between thyroid imaging reporting and data system and Bethesda system of reporting of thyroid cytopathology of thyroid nodule: a single centre experience. Journal of Cytology, 37(4), pp.193-199.
5. De, D., Dutta, S., Tarafdar, S., Kar, S.S., Das, U., Basu, K., Mukhopadhyay, P. and Ghosh, S., 2020. Comparison between sonographic features and fine needle aspiration cytology with histopathology in the diagnosis of solitary thyroid nodule. Indian Journal of Endocrinology and Metabolism, 24(4), pp.349-354.
6. Dhar, L., Singh, M., Jain, S.L., Rana, D., Kumar, J., Meher, R., Khurana, N., Verma, N., Mohammed, A. and Chellani, G., 2023. Correlation of ultrasound-based TIRADS and the Bethesda system for reporting of thyroid cytopathology: a study in a tertiary care centre. Cytopathology, 34(6), pp.590-596.
7. Grandhi, B., Durga, K., Mohan-Rao, N., Syamasundara-Rao, B., Vijayalakshmi, M. and Sunandha-Lakshmi, G.V., 2021. Study of thyroid lesions: co-relation of TIRADS with Bethesda system. Saudi Journal of Pathology and Microbiology, 6(4), pp.128-131.
8. Grani, G., Lamartina, L., Cantisani, V., Maranghi, M., Lucia, P. and Durante, C., 2018. Interobserver agreement of various thyroid imaging reporting and data systems. Endocrine Connections, 7(1), pp.1-7.
9. Haugen, B.R., Alexander, E.K., Bible, K.C., Doherty, G.M., Mandel, S.J., Nikiforov, Y.E., Pacini, F., Randolph, G.W., Sawka, A.M., Schlumberger, M., Schuff, K.G., Sherman, S.I., Sosa, J.A., Steward, D.L., Tuttle, R.M. and Wartofsky, L., 2016. 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer. Thyroid, 26(1), pp.1-133.
10. Hoang, J.K., Middleton, W.D., Farjat, A.E., Teefey, S.A., Abinanti, N., Boschini, F.J., Bronner, A.J., Dahiya, N., Hertzberg, B.S., Newman, J.R., Scanga, D., Vogler, R.C. and Tessler, F.N., 2018. Interobserver variability of sonographic features used in the American College of Radiology Thyroid Imaging Reporting and Data System. American Journal of Roentgenology, 211(1), pp.162-167.
11. Huang, E.Y.F., Kao, H.W., Lin, C.H., Jang, T.C., Venkatanarasimha, N., Kiong, K.L., See, A., Lee, K.M.A. and Lim, C.M., 2023. Concordance of the ACR TI-RADS classification with Bethesda scoring and histopathology risk stratification of thyroid nodules. JAMA Network Open, 6(9), p.e2331612.
12. Javed, R., 2023. Incidence of thyroid diseases in local population. Pakistan Journal of Medical & Health Sciences, 17(1), pp.17-19.
13. Li, W., Wang, Y., Wen, J., Zhang, L. and Sun, Y., 2021. Diagnostic performance of American College of Radiology TI-RADS: a systematic review and meta-analysis. American Journal of Roentgenology, 216(1), pp.38-47.
14. Middleton, W.D., Teefey, S.A., Reading, C.C., Langer, J.E., Beland, M.D., Szabunio, M.M. and Desser, T.S., 2017. Multiinstitutional analysis of thyroid nodule risk stratification using the American College of Radiology Thyroid Imaging Reporting and Data System. American Journal of Roentgenology, 208(6), pp.1331-1341.
15. Nafisa, A., Ikram, N., Khursheed, S., Anjum, R. and Akhtar, N., 2021. Epidemiologic profile of thyroid disorders in a tertiary care hospital: a five year analysis. Journal of Rawalpindi Medical College, 25(4), pp.517-521.
16. Omran, H.M., El-Shafei, M.A., El-Badawy, N.M., Bassiouny, A.M., Abd Elrahman, M.A. and El-Rifai, A.Y., 2021. Correlation between thyroid image reporting and data system and fine-needle aspiration cytology in the assessment of suspicious thyroid nodules. The Egyptian Journal of Surgery, 40(4), pp.1289-1294.
17. Osman, A.A., Ibrahim, I.G., Tahtabaşı, M., Kaya, V. and Shermatova, S., 2024. Comparison of sonographic-based TI-RADS classification and cytological-based Bethesda scoring: six-month experience in Somalia. Thyroid Research and Practice, 4(5), pp.1-6.
18. Rafi, M., Kanwal, M.A., Awais, M., Ahmad, T., Tayyab, M., Naem, M.S., Ghaloo, S.K., Hussain, S.R. and Faisal, M., 2024. Unveiling the malignancy risk of surgically treated Bethesda III thyroid nodules: a 10-year retrospective study comparing fine-needle aspiration cytology and histopathology. Cureus, 16(12), p.e75855.
19. Sakthisankari, S., Vidhyalakshmi, S., Shanthakumari, S., Devanand, B. and Nagul, U., 2021. The combination of ACR-Thyroid Imaging Reporting and Data System and the Bethesda System for Reporting Thyroid Cytopathology in the evaluation of thyroid nodules: an institutional experience. Cytopathology, 32(4), pp.472-481.
20. Singh, P., Gupta, N., Dass, A., Handa, U. and Singhal, S., 2021. Correlation of fine needle aspiration cytology with histopathology in patients undergoing thyroid surgery. Otolaryngologia Polska, 75(4), pp.33-39.
21. Tessler, F.N., Middleton, W.D., Grant, E.G., Hoang, J.K., Berland, L.L., Teefey, S.A., Cronan, J.J., Beland, M.D., Desser, T.S., Frates, M.C., Hammers, L.W., Hamper, U.M., Langer, J.E., Reading, C.C., Scoutt, L.M. and Stavros, A.T., 2017. ACR Thyroid Imaging, Reporting and Data System (TI-RADS): white paper of the ACR TI-RADS Committee. Journal of the American College of Radiology, 14(5), pp.587-595.

22. Wu, M., 2021. A correlation study between thyroid imaging report and data systems and the Bethesda system for reporting thyroid cytology with surgical follow-up: an ultrasound-trained cytopathologist's experience. Diagnostic Cytopathology, 49(4), pp.494-499.

Abbreviations

ACR: American College of Radiology. AUS/FLUS: Atypia of Undetermined Significance / Follicular Lesion of Undetermined Significance. CI: Confidence Interval. FNAC: Fine Needle Aspiration Cytology. SD: Standard Deviation. SPSS: Statistical Package for the Social Sciences. TBSRTC: The Bethesda System for Reporting Thyroid Cytopathology. TI-RADS: Thyroid Imaging Reporting and Data System. TR: TI-RADS Category.

Declarations

Ethics Approval and Consent to Participate

The study was approved by the institutional review board of Fatima Memorial Hospital, Lahore, and by the Research Evaluation Unit of the College of Physicians and Surgeons Pakistan. Written informed consent was obtained from all participants prior to enrolment.

Consent for Publication

Not applicable; no identifying patient information is presented in this manuscript.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualisation, methodology and supervision were carried out jointly by the authors. Data curation, formal analysis, investigation, and writing of the original draft were carried out by the corresponding author, with review and editing contributed by the supervising author. All authors approved the final manuscript.

Data Availability

The dataset analysed during the current study is available from the corresponding author on reasonable request.

Acknowledgements

The authors thank the technical staff of the Department of Pathology, Fatima Memorial Hospital, Lahore, for their assistance with sample processing.