Journal of Advanced Scientific Research

Available online through https://sciensage.info

ISSN

0976-9595

Research Article

Interobserver Concordance in The Bethesda System Categorization of Thyroid Nodules: A Critical Evaluation

Sukriti, Ankita, Sanjeev Kumar, Sudipta Chakrabarti*

Department of Pathology, ESIC Medical College and Hospital, Bihta, Patna, Bihar, India

Corresponding author: sudiptach@gmail.com

Received: 24-09-2026; Accepted: 09-10-2026; Published: 30-10-2026

© Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

https://doi.org/10.55218/JASR.2026171002

ABSTRACT

The Bethesda System for Reporting Thyroid Cytopathology 2023 (TBSRTC) offers a standardized, economical diagnostic approach. The maintenance of reproducibility of the TBSRTC categories in spite ofinterobserver variabilityis critical.In this study, interobserver diagnostic consistency and reproducibility were evaluated alongside category distribution to assess the institutional diagnostic reliability. This cross-sectional, observational, hospital-based retrospective study was carried out over a period of 2-year. Cases of thyroid nodules undergoing FNAC were reviewed and categorized according to the TBSRTC (2023) guidelines. To evaluate interobserver concordance, two cytopathologists independently reviewed the smears. A total of 82 cases were analysed. The mean age was 36.91 (± 10.82) years with a distinct female predominance (F:M ratio of 19.2:1). The majority of cases (62 cases, 75.6%) were benign (Bethesda Category II), followed by 12 cases (14.6%) of Atypia of Undetermined Significance/Follicular Lesion of Undetermined Significance (AUS/FLUS; Category III). Follicular Neoplasm/Suspicious for a Follicular Neoplasm (Category IV) was reported in 3 cases (3.7%), 2 cases (2.4%) were Suspicious for Malignancy (Category V), and 1 case (1.2%) was diagnosed as Malignant (Category VI). Interobserver concordance was highest in the benign group (Category II), yielding a 95.2% absolute agreement, while the highest variability was concentrated within the indeterminate cohorts (Categories III and IV). High interobserver concordance observed in this study underscores the reproducibility and diagnostic reliability of the Bethesda System in routine thyroid cytopathology practice. Interobserver validation may serve as an important quality assurance measure in institutions where comprehensive cytohistological follow-up is limited.

Keywords: Fine-needle aspiration cytology (FNAC); The Bethesda System for Reporting Thyroid Cytopathology; Thyroid nodules; Interobserver concordance.

INTRODUCTION

Thyroid nodules are frequently encountered in clinical practice, with a rising global prevalence attributable to the widespread use of high-resolution imaging.[1] While the vast majority of these nodules are benign, a small but significant proportion harbour malignancy, necessitating highly precise pre-operative evaluation.[2] Globally, thyroid carcinoma represents the fifth most common malignancy among women, showing a global incidence rate of 5.5%.[3]

In routine clinical practice, FNAC is universally accepted as the most cost-effective and valuable first-line tool for the pre-operative triage of thyroid lesions. [4] Over recent years, the widespread adoption of standardized reporting through the Bethesda System for Reporting Thyroid Cytology 2023 (TBSRTC) has significantly improved diagnostic uniformity, aided clinicians in therapeutic decision-making, and promoted evidence-based management algorithms. [4,5]

Despite these diagnostic improvements, considerable controversy persists regarding the management of thyroid nodules belonging to the indeterminate categories, specifically AUS/FLUS (Category III) and Suspicious for Follicular Neoplasm (Category IV).[5,6] These grey-zone categories present unique diagnostic challenges because their cytomorphological features overlap significantly, resulting in substantial interobserver variability. [6]Recent studies have shown that interobserver variability remains greatest in the indeterminate Bethesda categories, particularly AUS/FLUS and follicular neoplasm, despite widespread adoption of standardized reporting.[7]

Analysing the reproducibility of the Bethesda categories and evaluating interobserver concordance is critical to maintaining high-quality cytological reporting. This ensures that institutional diagnostic thresholds remain accurate, uniform, and reliable. The present study aims to evaluate the distribution of Bethesda categories with primary emphasis on evaluating interobserver diagnostic concordance to establish institutional reliability.

MATERIALS AND METHODS

Thisprospective cross-sectional observational study was conducted in the Department of Pathology at a tertiary care teaching hospital in Bihar, India, over a two-year period from January 2024 to December 2025. Consecutive patients presenting with thyroid nodules who underwent FNAC during the study period were eligible for inclusion. Cases with inadequate clinical records, non-palpable diffuse lesions,

absence of diagnostic material or confirmed non-thyroid lateral neck masses were excluded from analysis.

The minimum required sample size for this cross-sectional study was determined usingCochran's single proportion formula:$n = [Z^2 \times p \times (1 - p)] / d^2$

Where n represents the target sample size, Z is the standard normal deviate set at 1.96 to achieve a 95% confidence interval, p represents the expected proportion of the most prevalent diagnostic category (Bethesda Category II / Benign) in a tertiary care setting. Based on contemporary multi-tier reporting literature, the baseline regional prevalence (p) of benign thyroid nodules on fine-needle aspiration cytology (FNAC) is expected to range between 70% and 85%[7]. For this calculation, p was conservatively estimated at 0.75 (75%).Substituting these values into the formula yielded:$n = [(1.96)^2 \times 0.75 \times (1 - 0.75)] / (0.10)^2 \approx 72.03$ – denoting calculated minimum sample size required to achieve the desired statistical precision.

FNAC was performed by experienced clinicians using 22-gauge needles attached to 10 ml disposable syringes. Aspirated material was immediately expressed onto glass slides. Smears were both air-dried for Giemsa staining and also promptly fixed in 95% ethyl alcohol for Papanicolaou (Pap) staining.Demographic, clinical, and cytological data were noted and analyzed. Thyroid aspirates were categorized according to the Bethesda System for Reporting Thyroid Cytopathology, 2023 (TBSRTC).

Rather than calculating a speculative positive predictive statistical significance without universal histopathological confirmation, the researchers' focus was shifted toward diagnostic reproducibility. To assess interobserver reproducibility, all cytology smears were independently reviewed in a blinded manner by two experienced cytopathologists, and discordant cases were subsequently subjected to consensus review. All observations were recorded and analysed.

Statistical Methods

Statistical analysis was performed using GraphPad software. Descriptive statistics were utilized to summarize demographic characteristics and the distribution of Bethesda categories. To evaluate interobserver diagnostic consistency and reporting reproducibility, two cytopathologists independently reviewed the smears in a blinded manner. The degree of interobserver concordance was quantified using descriptive percentage agreement along with Cohen's kappa and weighted Cohen's kappa statistics to account for the ordinal nature of the 6-tier Bethesda system.

RESULTS AND STATISTICAL ANALYSIS

A total of 103 cases were provisionally suitable for the study. However, after applying the exclusion criteria, 82 eligible cases of thyroid lesions were analysed.The median age of the patients was 37.0 years, with a mean age of 36.91 ± 10.82years. A marked female predominance was noted, with 77 female cases and 4 male cases, yielding a female-to-male ratio of 19.2:1.

Bethesda Category II (Benign) lesions constituted the vast majority of cases (62 cases, 75.6%), followed by Category III (AUS/FLUS) lesions (12 cases, 14.6%). Category IV lesions were noted in 3 cases (3.7%), Category V accounted for 2 cases (2.4%), and Category VI accounted for 1 case (1.2%). Only 2 cases (2.4%) were categorized as non-diagnostic/unsatisfactory (Category I) (Table 1).The observer matrixhas beenrepresented in Table 2.

Interobserver Concordance

Interobserver agreement between the two cytopathologists was assessed using estimated Cohen's kappa statistics. Based on the documented distribution of concordant and discordant Bethesda categories, the overall observed agreement was 92.7% (76/82 cases). The estimated Cohen's kappa coefficient was 0.80, indicating substantial agreement between observers. Interobserver concordance was highest in the Benign category (Category II), achieving an absolute agreement of 95.2% (59/62 concordant cases).

The discordance (6 cases, 7.3%) were predominantly focused within the indeterminate groups. Specifically, 2 cases were disputed between Category I (Unsatisfactory) and Category II (Benign) due to low cellularity accompanied by thin colloid fluid, while 3 cases exhibited discordance between Category II and Category III (AUS/FLUS) due to focal nuclear enlargement and architectural crowding that one observer attributed to benign repair (Fig. 1 & 2). One case was disputed between Category II and Category IV (Follicular Neoplasm) (Fig 3). Blinded multi-observer panel review of these indeterminate categories successfully resolved these variations, eliminating the risk of diagnostic inflation or subjective misclassifications. A summary of diagnostic pitfalls and consequent diagnostic discordance is shown in Table 3, confirming that most discrepancies occurred within the indeterminate categories reflecting

Table 1: Distribution of Bethesda Categoriesand Diagnostic Reproducibility

Bethesda category	*Number of cases (n=82)*	*Percentage (%)*
I(Non-diagnostic / Unsatisfactory)	2	2.4
II(Benign)	62	75.6
III(AUS / FLUS)	12	14.6
IV(FollicularNeoplasm/Suspicious for FN)	3	3.7
V(Suspicious for Malignancy)	2	2.4
VI(Malignant)	1	1.2

Table 2: Interobserver Agreement Matrix for Bethesda Category Assignment by Two Independent Cytopathologists (observer 1 and 2). Discordant cases has been highlighted in bold font.

The Bethesda Categories of Observer 1	*The Bethesda Categories of Observer 2*					
	I	*II*	*III*	*IV*	*V*	*VI*
I	0	**2**	0	0	0	0
II	0	59	**3**	1	0	0
III	0	0	9	0	0	0
IV	0	0	0	**2**	0	0
V	0	0	0	0	2	0
VI	0	0	0	0	0	1

Total cases = 82, Concordant = 76, Discordant = 6 (Shown in bold)
I ↔ II = 2 cases, II ↔ III = 3 cases, II ↔ IV = 1 case

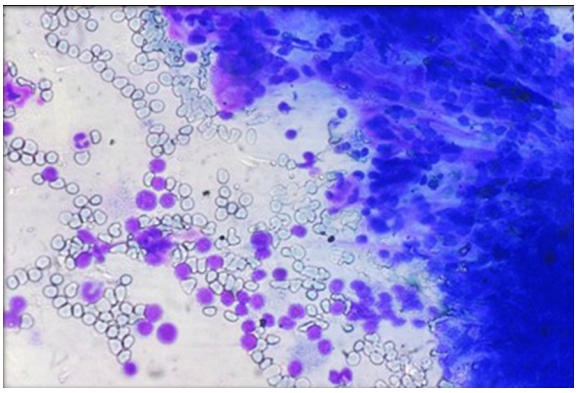


Figure 1: Giemsa, 10X showing Cytomorphological appearance of Atypia of Undetermined Significance / Follicular Lesion of Undetermined Significance (AUS/FLUS, Bethesda Category III) showing focal cytological atypia

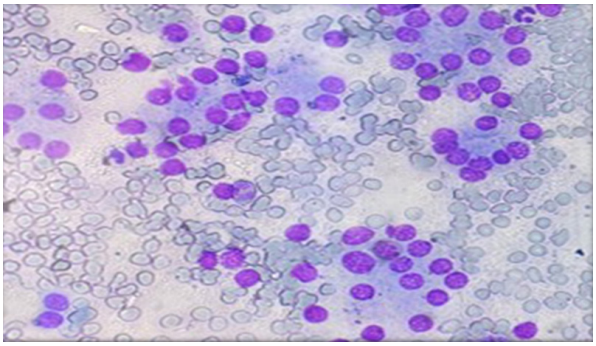


Figure 2: Giemsa, 40X showing Microphotograph of an indeterminate thyroid lesion (Bethesda Category III) demonstrating cellular features – this category has the highest proportion of interobserver variability.

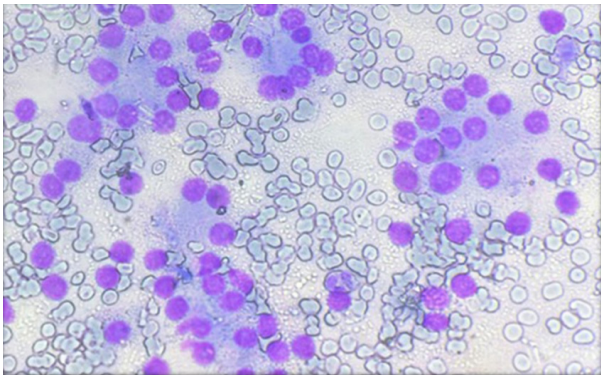


Figure 3: Giemsa, 40X showing Cytology smear of a Follicular Neoplasm (Bethesda Category IV) displaying characteristic microfollicular architecture and overlapping crowding patterns.

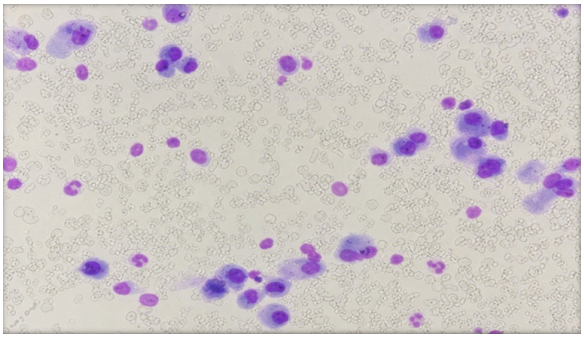


Figure 4: Giemsa, 40X showing Cytology smear of a Malignant Neoplasm (Bethesda Category VI) displaying characteristic round, ovoid, plasmacytoid cells having abundant cytoplasm and eccentric nuclei.

the known diagnostic challenges associated with these groups. No discordance was observed in Category VI (Fig 4).

Interobserver agreement was further assessed using estimated weighted Cohen's kappa statistics to account for the ordinal nature of the Bethesda categories. The weighted kappa coefficient was estimated to be 0.93, indicating almost perfect agreement between the two cytopathologists. Most disagreements occurred between adjacent diagnostic categories (Category I vs II and Category II vs III), while major category shifts were uncommon, contributing to the high weighted agreement.

DISCUSSION

In resource-limited regional healthcare systems, optimizing the triage of thyroid nodules via reliable screening is vital to prevent unnecessary surgical interventions while identifying malignant disease early.[8] Because extensive cytopathological correlation can be restricted by high patient migration rates and low local follow-up awareness, establishing a high level of interobserver concordance serves as an indispensable proxy for validating institutional diagnostic standards. [8,9]

Demographic and Category Profiling

The demographic profile of the 82-case cohort, showing a median age of 37 years and a strong female predominance (F: M - 19.2:1), aligns closely with established regional epidemiological patterns, noting peak incidence during the third and fourth decades of life. [3,8] Bethesda Category II (Benign) lesions comprise the overwhelming majority of cases (75.6%), which is strongly consistent with broad cytopathologic reviews, such as those by Rong et al. (51.2%)[10] and Anand *et al.* (75.9%)[9], confirming that most palpable thyroid nodules are non-malignant. Our non-diagnostic/unsatisfactory rate (Category I) was successfully limited to 2.4%, falling well within the standard international quality threshold (<10 %) and comparing favourably to the 15.7% rate reported by Rong *et al.*[10] owing to the dual-staining methodology (Pap and Giemsa) which maximizes slide evaluation accuracy.

Interobserver Concordance and Category Discrepancies

The estimated Cohen's kappa value of 0.80 observed in the present study is comparable to previously published (Table 4)interobserver reproducibility studies employing the Bethesda System. This finding indicates substantial agreement and supports the reproducibility of thyroid FNAC reporting in routine practice. Similar to earlier reports, most diagnostic variation was concentrated within the indeterminate categories (Bethesda III and IV), where overlapping cytomorphological features frequently contribute to observer-dependent interpretation.

Similar observations have been reported by Sripodok and Benjakul,[7] who evaluated interobserver variability in inconclusive thyroid FNAC categories at a tertiary care centre. Their study demonstrated that diagnostic disagreements were concentrated

Table 3: Common Cytomorphological Drivers of Interobserver Discordance and their suggested remedial measures.

Discrepant Diagnostic Path	*Main Cytomorphological Drivers / Pitfalls*	*Recommended Institutional Solution*
Category I vs. Category II	Thin background colloid mistaken for tissue fluid; cyst fluid containing sparse macrophages with low follicular cell yields.	Standardize adequacy rules: sparse cells are acceptable as Benign *only* if abundant colloid is present.
Category II vs. Category III	Focal nuclear enlargement, clearing, or grooving caused by hyperplastic/reparative nodules or thyroiditis.	Use multi-observer panel consensus review; avoid over-interpreting focal changes lacking widespread atypia.
Category II vs. Category IV	High cellularity with crowded microfollicular sheets and minimal visible colloid in hyperplastic adenomatoid nodules.	Maintain strict thresholds for microfollicular crowding; recommend ultrasound-guided repeat FNAC.
Category V vs. Category VI	Sparse sample cellularity containing few highly atypical cells or focal nuclear inclusions suspicious for papillary carcinoma.	Classify as Category V if diagnostic features are qualitative but sample volume is insufficient for an absolute diagnosis.

Table 4: Comparison of results and statistical significance of studies on Interobserver Concordance of TBSRTC

Study	*Cohort Size (N)*	*Overall Concordance / Kappa Metric (κ)*	*Highest Concordance Categories*	*Lowest Concordance / High Discordance Categories*
Present Study (2026)	82	92.7% Substantial Agreement (kappa = 0.80)	Category II (95.2%)	Category III (AUS/FLUS) & Category IV
Goel *et al.* (2018)[11]	107	91.5% Agreement (kappa = 0.613)	Category I & Category III	Category IV (FN/SFN) & Category V (50%)
Pattar *et al.* (2024)[12]	156	Substantial Agreement (kappa = 0.735 - 0.841)	Category II & Category VI	Category III (AUS/FLUS) & Category IV
Bahattin *et al.* (2024)[15]	173	Moderate to Good (kappa = 0.429 - 0.606)	Category II (kappa= 0.606)	Category III (kappa = 0.429) & Category V

predominantly within the indeterminate Bethesda categories, particularly AUS/FLUS and follicular neoplasm/suspicious for follicular neoplasm. The authors emphasized that subjective interpretation of subtle cytomorphological features contributes substantially to variability among observers. These findings support the observations of the present study, where the majority of discordant cases were likewise confined to Categories III and IV despite an overall high level of agreement.

The present study emphasized that while diagnostic reproducibility is excellent at the polar ends of the Bethesda scale (Category II and Category VI), it undergoes a noticeable decline in the indeterminate "grey zones" (Categories III and IV). This distribution trend is widely supported by global literature.[6,10]

The higher rate of discordance within Categories III and IV highlighted in the literature stems directly from overlapping cytomorphological criteria, such as focal microfollicular arrangements, mild nuclear atypia, and variation in background colloid thickness. Goel et al.[11] reported that five benign cases were discrepant because thin background colloid was misinterpreted as tissue fluid, shifting them into Category I, while focal nuclear features driven by benign inflammatory repair were frequently over-diagnosed as Category III atypia. To further illustrate the practical basis of such discrepancies, Table 4 summarizes the principal cytomorphological pitfalls and drivers of interobserver variability identified across published comparative studies.

By deploying an immediate, blinded internal multi-observer panel review for any borderline or gray-zone diagnoses, authors effectively resolved these focal cytomorphological variations, successfully avoiding diagnostic inflation and achieving strict compliance with the TBSRTC standards.Suggested steps of remedial measures has been depicted in Table 3.

The present study has certain limitations, including a relatively small sample size from a single tertiary care centre and a lack of systematic cyto-histopathological correlation caused by frequent patient migration in rural populations.

CONCLUSION AND CLINICAL RECOMMENDATIONS

The implementation of routine interobserver validation measures ensures high diagnostic reproducibility and prevents isolated single-observer errors. While FNAC serves as the benchmark screening test for thyroid lesions, it cannot be interpreted in absolute isolation.[13] International evidence underscores that diagnostic screening accuracy is maximized only when cytological findings are carefully integrated with high-resolution ultrasonographic risk assessments (such as EU-TIRADS profiling) and clinical risk indicators.[14] Fostering this collaborative clinic-radiological-cytological approach within rural healthcare frameworks, alongside efforts to improve long-term patient follow-up compliance, is vital to optimizing the management of thyroid nodules.

ACKNOWLEDGEMENT

Authors expresses special thanks to MA Masum Mallick, DMLT for his sincere help and support during procedure and staining

Informed consent

From the patients were obtained during the time of aspiration.

REFERENCES

1. Gharib H, Goellner JR. Fine-needle aspiration biopsy of the thyroid: an appraisal. Ann Intern Med. 1993;118(4):282-289.
2. Hegedüs L. Clinical practice. The thyroid nodule. N Engl J Med.

2004;351(17):1764-1771.
3. Oo ZP, Hlaing AM, Kyi KCS, Fukuoka J, Bychkov A. An overview of thyroid fine-needle aspiration practice in Myanmar. Gland Surg. 2020;9(5):1747-1753. doi:10.21037/gs-20-414
4. Ravetto C, Colombo L, Dottorini ME. Usefulness of fine-needle aspiration in the diagnosis of thyroid carcinoma. Cancer. 2000;90(6):357-363.
5. Lind P, Jacobson A, Nordenström E, Johansson L, Wallin G, Daskalakis K. Diagnostic sensitivity of fine-needle aspiration cytology in thyroid cancer. Sci Rep. 2024;14(1):24216. doi:10.1038/s41598-024-75677-7
6. Padmanabhan V, Barkan GA, Young NA, et al. Reproducibility of atypia of undetermined significance/follicular lesion of undetermined significance category using the Bethesda System for Reporting Thyroid Cytology when reviewing slides from different institutions: a study of interobserver variability among cytopathologists. Diagn Cytopathol. 2017;45(5):399-405. doi:10.1002/dc.23681
7. Sripodok S, Benjakul N. Interobserver variability in inconclusive diagnostic categories of thyroid fine needle aspiration cytology: an urban-based tertiary hospital experience. Ann Diagn Pathol. 2023;63:152083. doi:10.1016/j.anndiagpath.2022.152083
8. Alshahrani AS, Algazlan AG, Junaid M, Aldosari AA, Amer KA, Mubarki MH, Alharbi SM, Al-Qannass AM. Association of fine-needle aspiration of thyroid nodules with final histopathology in diagnosing thyroid malignancy: a single institute retrospective study. Cureus. 2022;14(11):e31733. doi:10.7759/cureus.31733
9. Anand B, Ramdas A, Ambroise MM, Kumar NP. The Bethesda System for Reporting Thyroid Cytopathology: a cytohistological study. J Thyroid Res. 2020;2020:8095378. doi:10.1155/2020/8095378
10. Rong R, Chaudhuri R, Wang X, et al. Fine needle aspiration cytology of thyroid nodules: a cytopathologic study of 2,043 cases. Zhonghua Bing Li Xue Za Zhi. 2016;45(6):380-384. doi:10.3760/cma.j.issn.0529-5807.2016.06.003
11. Goel G, Awasthi P, Khurana U, Joshi D, Majumdar K, Kapoor N. Reproducibility of the Bethesda System for Reporting Thyroid Cytopathology: a retrospective analysis of 107 patients. J Cytol. 2018;35(1):33-37. doi:10.4103/joc.joc_215_16
12. Pattar PM. Interobserver variability in diagnosing thyroid fine-needle aspiration cytology using the 2023 Bethesda System for Reporting Thyroid Cytopathology. Asian J Med Sci. 2024;15(11):75-81.
13. Erkinuresin T, Demirci H. Diagnostic accuracy of fine needle aspiration cytology of thyroid nodules. Diagnosis (Berl). 2020;7(1):35-40. doi:10.1515/dx-2019-0039
14. Russ G, Bonnema SJ, Erdogan MF, Durante C, Ngu R, Leenhardt L. European Thyroid Association guidelines for ultrasound malignancy risk stratification of thyroid nodules in adults: the EU-TIRADS. Eur Thyroid J. 2017;6(5):225-237. doi:10.1159/000478927
15. Bahattin E, Emine D, Çivi ÇK, Fatih Y. Inter- and intra-observer reproducibility of thyroid fine needle aspiration cytology: an investigation of Bethesda 2023 using immunohistochemical BRAFV600E antibody. J Cytol. 2024;41(4):221-228. doi:10.4103/joc.joc_26_24

HOW TO CITE THIS ARTICLE: Sukriti, Ankita, Kumar S, Chakrabarti S. Interobserver Concordance in The Bethesda System Categorization of Thyroid Nodules: A Critical Evaluation. *J Adv Sci Res.* 2026;17(10): 6-10 **DOI:** 10.55218/JASR.2026171002