Med Pregl 2025; LXXVIII (5-8): 127-131. Novi Sad: maj-avgust.

EDITORIAL
UVODNIK

THE ILLUSION OF DIAGNOSTIC AGREEMENT – INSTITUTIONAL BIAS AS A HIDDEN FORCE IN MODERN MEDICINE

ILUZIJA DIJAGNOSTIČKE SAGLASNOSTI – KAKO INSTITUCIONALNA PRISTRASNOST UTIČE NA SAVREMENU MEDICINU

Franz FOGT

ORCID NUMBER Franz Fogt – 0009-0009-4914-499X

Pennsylvania Hospital, University of Pennsylvania Department of Pathology and Laboratory Medicine

Editorial
Uvodnik
UDK 614.253.1:616-07
https://doi.org/10.2298/MPNS2508127F

Abstract
This editorial explores how diagnostic agreement within medical institutions may reflect shared interpretive culture rather than true diagnostic accuracy. Drawing from pathology, psychiatry, internal medicine, and medical decision science, the article argues that institutional diagnostic cultures, groupthink dynamics, cognitive imprinting during training, and anchoring effects can create the illusion of consistency that does not necessarily correspond to biological reality. The editorial also examines how central review panels and artificial intelligence systems may reproduce institutional biases unless diversity of input and multi-institutional calibration are incorporated into diagnostic processes.
Key words: Diagnosis; Bias; Institutional Practice; Artificial intelligence; Risk Factors

Sažetak
U tekstu se razmatra ideja da se visoka saglasnost među lekarima u okviru iste ustanove može zasnivati pre na zajedničkim profesionalnim navikama, obrascima razmišljanja i interpretativnim okvirima nego na stvarnoj dijagnostičkoj tačnosti. Polazeći od oblasti patologije, psihijatrije, interne medicine i nauke o medicinskom odlučivanju, članak tvrdi da faktori poput institucionalne kulture, grupnog mišljenja, obrazaca usvojenih tokom specijalističke obuke i kognitivnog „usidravanja" mogu stvoriti privid doslednosti dijagnoza koji ne mora uvek odgovarati biološkoj stvarnosti. Rad dalje analizira kako centralne ekspertne komisije i sistemi veštačke inteligencije mogu nesvesno preuzeti i reprodukovati istu vrstu institucionalne pristrasnosti, ukoliko se u dijagnostičke procese ne uključe raznovrsniji izvori stručnog mišljenja i međuinstitucionalna kalibracija.
Ključne reči: dijagnoza; pristrasnost; institucionalna praksa; veštačka inteligencija; faktori rizika

Diagnostic accuracy is a core premise of modern medical practice. Every decision – whether to biopsy, treat, surveil, discharge, or escalate care – depends on the assumption that the diagnostic categories we use are both meaningful and stable. Yet the reality is far more complex. Diagnostic error has traditionally been understood as underdiagnosis – situations in which important information is absent, overlooked, or not recognized. However, advances in highly sensitive diagnostic technologies have introduced a parallel challenge: overdiagnosis, in which abnormalities are detected that would never have caused symptoms or clinical harm. Overdiagnosis is increasingly recognized in population-based screening programs and in situations where imaging or laboratory tests are performed without clear medical indication. Together, these dual forms of diagnostic error demonstrate that diagnostic categories are not fixed biological truths but dynamic constructs shaped by evolving technology and interpretive judgment. As a result, the stability of these categories is more fragile than clinicians often realize [1–3].

Agreement among observers does not always reflect diagnostic truth. Instead, it may arise from shared assumptions and institutional habits that act as invisible forces shaping diagnostic behavior. The phenomenon is not limited to a single organ system; studies across multiple diagnostic fields show that clinicians who work within the same institutional environment, share similar case volumes, and engage in ongoing informal learning tend to demonstrate higher internal concordance [4]. Such agreement, however, often reflects a common interpretive culture rather than objective correctness. Furthermore, clinicians themselves frequently sense when their interpretations fall near diagnostic boundaries: discordance is far more likely when observers report low confidence, uncertainty, or a feeling that the case does

✉ Corresponding author: Franz Fogt, E-mail: fogt@mail.med.upenn.edu

not neatly fit established patterns. In real-world practice, these moments of uncertainty often prompt a second opinion or additional consultation – mechanisms known to improve diagnostic accuracy – but such safeguards are absent in controlled study environments. A recent study of "indefinite for dysplasia" in gastrointestinal mucosal biopsies illustrates this phenomenon vividly: strong internal agreement masked the possibility that shared diagnostic thresholds, rather than biological reality, were shaping the classification [5].

In that study, the authors sought to understand the natural history of indefinite lesions in the esophagus and stomach - cases in which atypia is present but insufficient for a definitive diagnosis of dysplasia. The reported progression rates were surprisingly high, particularly in the esophageal cohort. Instead of behaving like reactive or inflammatory mucosa, which theoretically should form the bulk of the indefinite category, many lesions progressed to definite dysplasia, including a meaningful number that advanced to high-grade dysplasia or carcinoma. This unexpected behavior raised an important question: were these lesions truly "indefinite," or were some of them, at baseline, subtle examples of low-grade dysplasia that were not recognized during initial interpretation [6].

To answer this, the investigators performed a central review of the original biopsies, using two expert pathologists from the same institution. Critically, nearly all cases remained classified as indefinite upon review, only a small percentage were downgraded, and none were upgraded. At first glance, this very high level of agreement appears reassuring. It suggests strong interobserver reliability, a consistent approach to diagnosis, and diagnostic confidence. However, when placed in the context of the broader gastrointestinal pathology literature, this result is atypical. Many multi-institutional studies of Barrett-related dysplasia have demonstrated extremely high variability in the evaluation of borderline lesions. When cases diagnosed as low-grade dysplasia in community practice are reviewed by panels of expert gastrointestinal pathologists from different institutions, the majority – sometimes more than sixty percent – are downgraded to either reactive or indefinite, and a smaller but important fraction are upgraded. This long-recognized variability has shaped modern surveillance algorithms and is the reason expert external review is now recommended before ablative therapy is offered. Further complicating interpretation, reviewers in prior studies have noted that when they perceived a case as difficult or diagnostically ambiguous, their inclination was to classify it as indefinite for dysplasia – even when cytologic atypia was substantial – rather than risk "overdiagnosis". As a result, a surprisingly high percentage of cases without a clear consensus diagnosis, and with relatively low average dysplasia scores, ultimately progressed to cancer at rates higher than expected [7]. These observations highlight the vulnerability of histologic interpretation when it operates without clinical context and underscore the inherent limitations of "interpretation divorced from clinical data". Yet, despite these challenges, a consistent finding across the literature remains: as the degree of dysplasia increases, so does the likelihood of detecting invasive carcinoma, with higher-grade categories associated with progression in greater numbers of patients and over shorter intervals.

In that context, a reclassification rate of only four percent in a single-institution review is not simply a sign of precision – it is almost certainly a reflection of shared diagnostic culture. Pathologists working within the same institution, attending the same consensus meetings, and reviewing difficult cases together inevitably internalize each other's interpretive boundaries. They share a similar understanding of what constitutes dysplasia, what features deserve the designation of "indefinite", and how to evaluate atypia in the setting of inflammation or regenerative change. This alignment is not intentional; it develops organically over years of shared practice, mutual teaching, and institutional tradition. It becomes an unspoken diagnostic dialect – a locally consistent set of expectations about what specific histologic patterns signify. When pathologists who share this dialect review cases, even when blinded to clinical outcomes, their diagnoses will naturally converge. Agreement within this setting is therefore an echo of shared assumptions, not evidence that those assumptions reflect biological truth. This may be supported by various studies in which diagnostic accuracy or agreement improved after a consensus training in which specific diagnostic criteria were taught and agreed upon [5,7].

The implications of institutional diagnostic harmonization are profound. First, it creates an illusion of reproducibility. High agreement within a single center, even among recognized experts, may reflect internal coherence reinforced by group-level cognitive and organizational forces rather than true diagnostic objectivity. Numerous studies in health care have shown that groupthink, driven by cohesiveness, shared training backgrounds, and hierarchical team structures, is common and promotes alignment within clinical groups. This fosters distinct diagnostic "dialects" that may diverge substantially from those of other institutions, meaning that internal concordance does not ensure external validity.

When diagnostic categories are shaped primarily by local norms rather than universally applied criteria, institutions may develop distinct thresholds for borderline lesions, resulting in clinically meaningful variation. In pathology, for example, a biopsy considered indefinite for dysplasia in one center may be interpreted as definite low-grade dysplasia in another, reflecting how departments collectively learn and internalize diagnostic thresholds. Psychiatry shows a similar pattern: diagnostic boundaries along continua – such as mood or personality disorders – are influenced by local training traditions, supervisory styles, and the diagnostic "language" of a residency program. Clinicians frequently engage in confirmatory information search, reinforcing initial impressions, and such tendencies are amplified within cohesive teams where shared mental models, longstanding habits, and groupthink dynamics encourage conformity [9].

Internal medicine provides additional evidence that hierarchy and culture shape diagnostic reasoning. Qualitative studies reveal that residents often attribute diagnostic decisions to consultants or senior team members, perceiving limited opportunities to contribute their own assessments. This hierarchical structure reduces cognitive diversity and fosters groupthink, causing junior clinicians to adopt prevailing interpretations even when alternative hypotheses come to mind. Practitioners may also diverge from guidelines when new evidence conflicts with impressions formed during training - a phenomenon rooted in cognitive imprinting and the hidden curriculum, through which the diagnostic habits and priorities of senior clinicians become enduring defaults.

Across specialties, the consequence is consistent: divergence in diagnostic philosophy directly affects treatment decisions, follow-up strategies, and patient outcomes – whether the lens is a microscope, a mental status examination, or a multidisciplinary ward round [10,11].

A second consequence of institutional diagnostic culture is the effect of anchoring bias on the observed natural history of disease. When a department adopts a particular interpretive threshold – for example, a conservative standard for diagnosing dysplasia – that threshold becomes a cognitive anchor against which all subsequent cases are judged. Borderline lesions that might be classified as dysplastic elsewhere become anchored to the local "indefinite" category. As a result, the indefinite group at that institution will inevitably show higher progression rates, not because the biology differs, but because true dysplasia was underclassified at baseline. Conversely, institutions anchored to a lower threshold for diagnosing dysplasia will observe lower progression rates in their indefinite category because fewer biologically dysplastic lesions remain within it. These anchoring effects can produce systematic, institution-specific distortions in progression data, leading to misleading impressions of disease behavior that may be incorrectly generalized to broader populations [12–14].

Third, institutional bias undermines research quality in ways that mirror well-documented group-level distortions in expertise recognition. As shown in Bunderson's 2003 study on status dynamics in work groups, groups tend to rely on familiar internal cues and shared interpretive habits rather than objective indicators of expertise, particularly when power is centralized or members share a long-standing institutional culture. Applied to pathology research, when both the original diagnosis and the "expert" review come from the same institution, reviewers are not evaluating correctness – they are replicating the institution's own status hierarchy and interpretive norms. This produces an illusion of agreement driven by shared culture rather than true diagnostic accuracy. Such designs inflate measures of concordance, obscure misclassification risk, and generate evidence that reflects the institution's diagnostic ideology rather than biologic reality. When these internally reinforced findings are then used to shape clinical guidelines or risk models, institutional bias propagates outward, amplifying local patterns of over- or under-diagnosis into system-wide practice [15].

This issue is not limited to gastrointestinal pathology. Radiology provides a particularly revealing parallel. In a recent study of nearly 1.9 million radiology reports, only 114 intra-institutional discrepant-opinion alerts were recorded – an astonishingly low rate of 0.006%, or roughly one disagreement per 16,563 reports. Yet despite this apparent harmony, more than half of these rare alerts represented major interpretive discrepancies, and over fifty percent resulted in changes to patient management. This extremely low internal discrepancy rate stands in striking contrast to the much higher discordance documented when radiologists reinterpret studies from other institutions, where disagreement rates commonly range from 7% to over 30% and where second opinions are more accurate in the vast majority of cases. This pattern suggests that intra-departmental agreement reflects a shared diagnostic culture more than true diagnostic precision [16].

In psychiatry, diagnostic tendencies align closely with training environments and training setting with consequent high agreement within departments revealing that what looks like reliability may instead be the product of local diagnostic philosophy [17].

Institutional diagnostic culture is therefore a pervasive, cross-disciplinary phenomenon. Yet it remains largely unacknowledged, perhaps because it challenges medicine's belief in its own objectivity. We are accustomed to thinking that diagnoses reflect intrinsic biological categories. But for many conditions, especially those on diagnostic continua – reactive versus dysplastic mucosa, benign versus atypical nevi, mild versus moderate psychiatric symptoms, borderline valvular stenosis – the distinction depends as much on interpretive thresholds as on biology. Those thresholds do not arise in isolation. They are shaped by training, experience, local tradition, and shared norms [18].

Recognizing institutional bias is the first step toward mitigating its effects, and solutions are well within reach. One essential approach is to increase the use of multi-institutional review panels in studies evaluating borderline or subjective diagnostic categories. When pathologists or radiologists from different institutions independently evaluate the same cases, the resulting consensus is more likely to reflect underlying biological patterns rather than the idiosyncrasies of any single department. Standardized diagnostic criteria, structured scoring systems, and ancillary testing – such as immunohistochemical markers in pathology or quantitative assessment tools in radiology – can further reduce interpretive drift. Digital pathology and cloud-based image sharing now make such cross-institutional consultation feasible at a scale that was previously unimaginable.

However, it is equally important to acknowledge that central review panels are not immune to the very biases they aim to counteract. Once a panel meets regularly, discusses cases, and negotiates shared thresholds, it begins to function like an institutional group with its own internal diagnostic culture. Empirical evidence illustrates this limitation. In a central pathology review by Spoor and Cross involving seven pathologists and 630 biopsy specimens, concordance with the original diagnoses for atypical hyperplasia was only 68%. Even more revealing, when the central review diagnoses were compared with final hysterectomy outcomes, 4.7% of cases were still misclassified despite multilayered expert review. These findings underscore that while multi-institutional review improves reliability, it does not eliminate group-level interpretive bias. Instead, it shifts the locus of bias from individual departments to newly formed expert collectives – groups that may themselves develop shared assumptions and diagnostic dialects over time [19,20].

Artificial intelligence introduces both new risks and new possibilities [21]. Machine-learning models trained on data from a single institution – or from narrow, homogeneous datasets – inevitably absorb and reproduce the diagnostic habits, thresholds, and blind spots embedded in that environment. Studies in medical AI have already shown that such models can perpetuate institution-specific biases and perform poorly when applied elsewhere, echoing the well-documented problems of algorithmic bias in criminal justice, hiring, and loan approval systems. In those domains, biased algorithms have created self-reinforcing interpretive worlds in which human insight is diminished and skepticism becomes difficult; questioning the algorithm is often perceived as questioning objectivity itself. Medicine is not immune to this risk. If uncritically deployed, AI could magnify institutional diagnostic philosophies and obscure the very uncertainty clinicians need to recognize. Yet when designed with diverse, multi-institutional training data and used deliberately, AI can serve the opposite role – quantifying interobserver variability, flagging discordant cases, and exposing where local diagnostic cultures diverge. In this way, AI can help map and ultimately correct the institutional biases that threaten diagnostic accuracy, rather than silently hardwiring them into future practice [21,22].

Ultimately, the medical community must confront a simple but uncomfortable truth: agreement within an institution is not synonymous with accuracy. It is entirely possible for a group of experts to agree consistently on a diagnosis that reflects their shared interpretive culture rather than biological fact. Only through external challenge, cross-institutional collaboration, and deliberate exposure to diverse diagnostic viewpoints can we ensure that our categories reflect disease rather than tradition. The illusion of diagnostic agreement is one of the most subtle – and most consequential – sources of variability in modern medicine. Acknowledging it is the first step toward building diagnostic systems that are truly reliable, reproducible, and aligned with the realities of patient biology.

References

1. Graber ML, Wachter RM, Cassel CK. Bringing diagnosis into the quality and safety equations. JAMA. 2012;308(12):1211-2.

2. Zwaan L, Singh H. The challenges in defining and measuring diagnostic error. Diagnosis (Berl). 2015;2(2):97-103.

3. Schröder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009;360(13):1320-8.

4. Elmore JG, Longton GM, Carney PA, Geller BM, Onega T, Tosteson AN, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA. 2015;313(11):1122-32.

5. Angerilli V, Galuppini F, Brignola S, Businello G, Filippin B, Pennelli G, et al. Predictors of neoplastic progression in gas-

troesophageal lesions indefinite for dysplasia. Histopathology. 2025;87(3):424-35.

6. Montgomery E, Goldblum JR, Greenson JK, Haber MM, Lamps LW, Lauwers GY, et al. Dysplasia as a predictive marker for invasive carcinoma in Barrett esophagus. Hum Pathol. 2001;32(4): 379-88.

7. Curvers WL, ten Kate FJ, Krishnadath KK, Visser M, Elzer B, Baak LC, et al. Low-grade dysplasia in Barrett's esophagus: overdiagnosed and underestimated. Am J Gastroenterol. 2010;105(7): 1523-30.

8. van den Einden LC, de Hullu JA, Massuger LF, Grefte JM, Bult P, Wiersma A, et al. Interobserver variability and the effect of education in the histopathological diagnosis of differentiated vulvar intraepithelial neoplasia. Mod Pathol. 2013;26(6):874-80.

9. Mendel R, Traut-Mattausch E, Jonas E, Leucht S, Kane JM, Maino K, et al. Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses. Psychol Med. 2011;41(12):2651-9.

10. Choi JJ, Mhaimeed N, Al-Mohanadi D, Mahmoud MA. Medical residents' perceptions of group biases in medical decision making: a qualitative study. BMC Med Educ. 2024;24(1):661.

11. Saini V, Garcia-Armesto S, Klemperer D, Paris V, Elshaug AG, Brownlee S, et al. Drivers of poor medical care. Lancet. 2017; 390(10090):178-90.

12. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak. 2016;16(1):138.

13. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science. 1974;185(4157):1124-31.

14. Joseph MM, Mahajan P, Snow SK, Ku BC, Saidinejad M. Optimizing pediatric patient safety in the emergency care setting. J Emerg Nurs. 2022;48(6):652-65.

15. Bunderson JS. Recognizing and utilizing expertise in work groups: a status characteristics perspective. Admin Sci Q. 2003;48(4):557-91.

16. DiPiro PJ, Licaros A, Zhao AH, Glazer DI, Healey MJ, Curley PJ, et al. Frequency and clinical utility of alerts for intrainstitutional radiologist discrepant opinions. J Am Coll Radiol. 2023;20(4):431-7.

17. Jensen-Doss A, Hawley KM. Understanding clinicians' diagnostic practices: attitudes toward the utility of diagnosis and standardized diagnostic tools. Adm Policy Ment Health. 2011; 38(6):476-85.

18. Clark LA, Cuthbert B, Lewis-Fernández R, Narrow WE, Reed GM. Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's research domain criteria (RDoC). Psychol Sci Public Interest. 2017;18(2):72-145.

19. McCoy CA, Coleman HG, McShane CM, McCluggage WG, Wylie J, Quinn D, et al. Factors associated with interobserver variation amongst pathologists in the diagnosis of endometrial hyperplasia: a systematic review. PLoS One. 2024;19(4):e0302252.

20. Kepp KP, Aavitsland P, Ballin M, Balloux F, Baral S, Bardosh K, et al. Panel stacking is a threat to consensus statement validity. J Clin Epidemiol. 2024;173:111428.

21. Javaid MK. Role of artificial intelligence and fracture liaison service setting. Med Pregl. 2025;78(1-2):7-9.

22. Koçak B, Ponsiglione A, Stanzione A, Bluethgen C, Santinha J, Ugga L, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagn Interv Radiol. 2025;31(2):75-88.

23. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. NPJ Digit Med.2023;6(1):113.

Rad je primljen 1. XII 2025.
Recenziran 1. XII 2025.
Prihvaćen za štampu 1. XII 2025.
BIBLID.0025-8105:(2025):LXXVIII:5-8:127-131.