Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

REG2024: a benchmark challenge and curated dataset for pathology report generation from giga- pixel whole-slide images of bladder cancer

Harim Oh
Korea University Anam Hospital, Korea University
Eunsu Kim
Korea University College of Medicine
Hyeseong Lee
Korea University College of Medicine
Kyung A Kim
Yonsei University College of Medicine
Sungman Hong
Korea University
Sehwan Yoo
Lee Kong Chian School of Medicine, Nanyang Technological University
Yousun Kang
Tokyo Polytechnic University
Hyo Sup Shim
Yonsei University College of Medicine
Yumi Lee
Ewha Womans University
Thanaporn Viriyasaranon
Ewha Womans University
Wonkyeong Lee
Stanford University
Jang-Hwan Choi
Ewha Womans University
Eunbeen Jo
Korea University College of Medicine
Daehong Lee
AIVIS Inc
Keonwoo Noh
AIVIS Inc
Yeongjae Shin
Korea University Guro Hospital
Kyeongman Park
Seoul National University

Joonbeom Park
Seoul National University
Jong-Ho Kim
Korea University Research Institute for Medical Bigdata Science, Korea University College of Medicine
Hyun Yee Cho
Korea University Anam Hospital, Korea University
Won-Ki Jeong
Korea University
Sung Hak Lee
The Catholic University of Korea
Sangjeong Ahn
vanitas80@korea.ac.kr

Korea University Anam Hospital, Korea University

Research Article

Keywords: Pathology report generation, Whole slide image, Computational pathology, Vision-language model, Multi-modal medical AI, Evaluation metrics

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10824156/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

We aimed to construct a curated whole-slide image (WSI)-report dataset for bladder cancer and establish a clinically grounded benchmark for evaluating vision–language models through an open community challenge.

Methods

We curated a dataset of 902 deidentified bladder cancer WSIs paired with corresponding pathology reports. Seven teams submitted distinct modeling frameworks. Performance was assessed using a clinically weighted composite metric integrating word-overlap scores (ROUGE-L, BLEU-4), keyword matching, and embedding-based semantic similarity. Expert mean opinion scores (MOS) were obtained, and the correlation between the weighted-sum score and expert-assigned MOS was analyzed.

Results

Two modeling strategies emerged: end-to-end generative architectures and multi-step classification-then-generation pipelines. Among individual submissions, the top-ranked multi-step pipeline achieved the highest main-diagnosis accuracy (48/50), whereas end-to-end models showed greater sensitivity to rare pathological features. The weighted-sum metric correlated strongly with MOS within each team (Spearman's ρ, 0.65–0.91) and, on reports with clinically important misclassifications, agreed with MOS more closely than any single metric.

Conclusions

Performance variability was influenced by architectural design and the integration of pretrained foundation models. Our proposed evaluation framework addressed key limitations of conventional natural language processing metrics in evaluating pathology reports. This challenge establishes a benchmark for clinical-grade pathology report generation by providing vision-language pathology datasets and introducing a clinically aligned evaluation framework, thereby advancing computational pathology.

1. Background

Histopathology remains the gold standard for cancer diagnosis and treatment strategy determination across various organs [1]. Diverse histopathologic findings are integrated into comprehensive pathology reports in accordance with international guidelines [2, 3]. These reports are critical for communication between pathologists and clinicians, shaping therapeutic strategies and guiding prognostic assessment. Yet, report generation remains laborious and heuristic, requiring substantial expertise and time. Given the steadily increasing workload for pathologists [4], autonomous report generation systems represent a critical unmet need; they could reduce this workload and allow pathologists to devote more time to cases that require specialized expertise.

In recent years, artificial intelligence (AI) has driven remarkable progress in computational pathology (CPath), with deep learning models achieving near-expert performance in tasks such as cancer subtyping [5], grading [6], survival prediction [7, 8], and biomarker prediction [9, 10]. However, most existing models are task-specific rather than task-agnostic and rely on unimodal, image-based approaches confined to disease-specific settings [11], limiting their generalizability across diverse diagnostic contexts. In addition, the scarcity of multimodal pathology datasets, particularly vision–language benchmarks, remains a major bottleneck.

The rapid development of vision–language models (VLMs) has enabled new approaches in medical AI [12, 13]. In pathology, the emergence of foundation models (FMs) has enabled significant advances in histopathological image representation learning [14, 15], creating new opportunities for automated pathology report generation [12]. A growing research direction now explores post-training strategies to adapt these FMs for diverse downstream tasks. Nonetheless, key challenges persist, including representation learning at the whole-slide image (WSI) level, weak image–text alignment, the detectability of rare fine-grained features, and the lack of standardized evaluation metrics [16].

In CPath, several community challenges have been held across various organs such as prostate [6], ovary [17], and breast [18, 19], addressing diverse tasks and advancing method development. Despite this progress, most prior efforts have focused on narrowly defined tasks within single-disease contexts.

To address these gaps, we organized REG2024, the first community-wide challenge for end-to-end pathology report generation from WSIs. Participants were tasked with generating structured pathology reports directly from bladder-cancer WSIs. The challenge makes three primary contributions: (i) the release of a curated WSI-level vision–language pathology dataset, (ii) explicit support for pathology foundation models to promote the development of post-training strategies, and (iii) the introduction of a clinically weighted evaluation framework, together with an empirical assessment of its agreement with expert ratings.

2. Methods

2.1 Organization

The REG2024 challenge was organized into four stages—data preparation, training, testing, and results analysis—forming a structured workflow (**Additional file 1**: **Figure S1**). The challenge was hosted on our in-house website (www.k-medicon.com; currently closed; **Additional file 1: Figure S2**). The detailed timeline of the challenge is summarized in **Additional file 1: Figure S1** legend.

2.2 Datasets

The REG2024 dataset was jointly provided by Korea University Anam Hospital and the Korea National Information Society Agency. It consists of specimens from patients who underwent transurethral resection for bladder cancer at Korea University Anam Hospital between 2020 and 2023, encompassing multiple disease categories. All 902 WSIs were scanned using an Aperio AT2 slide scanner at 20x magnification (Leica Biosystems, Deer Park, IL, USA). The study protocol was approved by the Institutional Review Board of Korea University Anam Hospital (IRB No. 2023AN0243).

Each WSI was paired with a corresponding pathology report, and data curation was conducted by two board-certified pathologists (H.O. and S.A.), with additional review by a genitourinary subspecialty pathologist (H.C.). The pathology reports were reconstructed based on the College of American Pathologists (CAP) guidelines (Urinary Bladder, Biopsy, v4.2.0.0), with discrepancies resolved by consensus. The reports include organ (bladder), procedure (transurethral resection), histologic type, histologic grade, tumor extent, and presence of muscle proper (Fig. 1 and **Additional file 1: Supplementary Data 1**). Histologic typing followed the WHO Classification of Tumors, 5th Edition. When multiple WSIs were obtained from a single patient, a separate report was generated for each WSI.

WSIs were provided in .tif format at 0.5 μm/pixel resolution (scanned at 20x magnification), with only the level-0 (highest-resolution) layer available from the pyramidal structure. Paired pathology reports were provided in .txt format for the training set. The dataset (N = 902) was split into a training set (n = 752) and a test set (n = 150), which included 100 public-test and 50 private-test cases, with equal class distribution across subsets. The detailed distribution is presented in Table 1.

Table 1
Distribution of training and test dataset

Category	Subcategory	Total (N = 902)	Training set (N = 752)	Public test (N = 100)	Private test (N = 50)
Main diagnosis	Invasive urothelial carcinoma	419 (46.45%)	350 (46.54%)	46	23
	Non-invasive urothelial carcinoma, high grade	239 (26.5%)	208 (27.66%)	20	11
	Non-invasive urothelial carcinoma, low grade	154 (17.07%)	125 (16.62%)	22	7
	Urothelial carcinoma in situ	90 (9.98%)	69 (9.18%)	12	9
Invasion depth	Absent	483 (53.55%)	402 (53.46%)	54	27
	Subepithelial connective tissue_minor	104 (11.53%)	87 (11.57%)	11	6
	Subepithelial connective tissue_major	167 (18.51%)	140 (18.62%)	18	9
	Muscle proper	148 (16.41%)	123 (16.36%)	17	8
Urothelial carcinoma in situ	Present	230 (25.5%)	183 (24.34%)	29	18
	Absent	672 (74.5%)	569 (75.66%)	71	32
Muscle proper	Present	723 (80.16%)	604 (80.32%)	75	44
	Absent	179 (19.84%)	148 (19.68%)	25	6
Squamous differentiation	Present	36 (3.99%)	28 (3.72%)	4	4
	Absent	866 (96.01%)	724 (96.28%)	96	46
Granulomatous inflammation	Present	36 (3.99%)	31(4.12%)	2	3
	Absent	866 (96.01%)	721(95.88%)	98	47

For cases with subepithelial connective tissue invasion, we introduced two categories, minor and major, based on whether the invasion could be identified at 200x magnification by board-certified pathologists (**Additional file 1: Figure S3**). This scheme enables fine-grained evaluation of feature detection: minor invasion reflects the ability to capture subtle, localized features, whereas major invasion reflects recognition of broader, global

features. This categorization was introduced specifically for the challenge and is not typically used in pathology reports.

2.3 Performance evaluation

Generated reports were assessed using both automatic metrics and expert review.

Automatic metrics

● Word score: This metric evaluates the textual similarity between generated and ground-truth reports using ROUGE-L [20] and BLEU-4 [21] scores.

● Keyword score: Assessed using ScispaCy [22], this metric measures the overlap of key medical terms between the generated and ground-truth reports.

● Embedding score: Using OpenBioLLM-Llama3-70B [23], this metric measures semantic similarity, focusing on the clinical relevance between the generated and ground-truth reports.

Expert metric

● MOS: five board-certified pathologists independently rated private-test outputs (n = 50) for diagnostic accuracy and completeness.

2.3.1 Mean Opinion Score (MOS)

In addition to automatic metrics, five board-certified pathologists evaluated generated pathology reports to provide a comprehensive assessment (Table 2). The diagnostic accuracy component focused on clinically relevant elements, with weights assigned according to their impact on staging and prognosis: Main diagnosis (0.20) and invasion depth (0.20) were prioritized because both directly inform tumor T stage (Urothelial carcinoma *in situ* [Tis], Non-invasive papillary urothelial carcinoma [Ta], Invasive urothelial carcinoma with involvement of subepithelial connective tissue [T1], and Invasive urothelial carcinoma with involvement of muscle proper [T2 or higher]). Histologic grade received a weight of 0.10, reflecting its prognostic and therapeutic importance. The presence/absence of urothelial carcinoma in situ received 0.10 due to its distinct prognostic implications. Muscle proper inclusion received 0.05 as a marker of surgical adequacy rather than direct prognosis. Squamous differentiation and granulomatous inflammation, considered minor pathologic features, were each assigned a weight of 0.05 to reflect their diagnostic relevance while limiting their contribution to the overall score.

Table 2
Mean Opinion Score for private test set

1. Diagnostic Accuracy and Key Feature identification (Total = 0.75)		
Category	Subcategory	Score (Sum of each category)
Main diagnosis	Non-invasive urothelial carcinoma	0.2 point
	Invasive urothelial carcinoma	
	Urothelial carcinoma in situ	
Invasion depth	Absent	0.2 point
	Subepithelial connective tissue	
	Muscle	
Grade	Low	0.1 point
	High	
Urothelial carcinoma in situ	Present	0.1 point
	Absent	
Muscle proper	Include	0.05 point
	Not included	
Squamous differentiation	Yes	0.05 point
	No	
Granulomatous inflammation	Yes	0.05 point
	No	
2. **Report completeness** (Total = 0.25)		
Category	Description	Score
Complete	Includes all required content and adheres perfectly to the report format.	0.25
Nearly complete	Includes all required content but may lack readability or deviate slightly in format.	0.2
Partial	Missing at least one required element (e.g., invasion depth not included).	0.15
Incomplete	Missing critical elements such as organ name or procedure name.	0.1
Poorly completed	Does not follow the format and lacks most content, making interpretation difficult.	0.05
The diagnostic accuracy criteria are calculated by summing the scores from seven categories, with a total possible score of 0.75. The completeness criteria are evaluated out of a total score of 0.25, where five pathologists assign scores to the respective categories, and the scores are averaged. The final score is the sum of the diagnostic accuracy and completeness scores, ranging from 0 to 1.		

The completeness component assessed thoroughness, structural coherence, and readability using a five-category ordinal scale, with each category mapped to a fixed score ranging from 0.05 to 0.25 in 0.05-point increments (Table 2). Given the subjective nature of this evaluation, five board-certified pathologists independently assigned each report to one of the five categories, and the average of the corresponding scores across raters was taken as the final completeness score.

The overall MOS was computed by combining the diagnostic accuracy score (75% weight) and the completeness score (25% weight), yielding a final value from 0 to 1. This approach captures both objective factor and expert judgment relevant to clinical use.

2.3.2 Ranking scores

For the public-test set, model performance was assessed using the following weighted formula:

Ranking score_Pu = 0.15 ROUGE-L + 0.15*BLEU-4 + 0.4*Keyword score + 0.3*Embedding score*

The weights were prespecified in the challenge rules and announced to participants before the test phase. Ranking score_Pu was updated daily and displayed on a public leaderboard to provide real-time feedback, enabling participants to use the public test set for intermediate evaluation and model refinement.

For the private-test set, the same prespecified automatic-metric formula was applied:

*Ranking score_Pr = 0.15*ROUGE-L + 0.15*BLEU-4 + 0.4*Keyword score + 0.3*Embedding score*

During the private phase, Ranking score_Pr was also updated daily and displayed on a leaderboard to ensure transparency. After the competition concluded, private-set predictions underwent expert evaluation, integrating automated metrics with MOS.

The final score was derived by combining the public and private testset scores, with the public-test score contributing 20% and the private-test score contributing 80%. Within the private-test contribution, the automatic ranking score was weighted 70% and the MOS 30%. The final score was:

*Final score = 0.2*Ranking score_Pu + 0.56*Ranking score_Pr + 0.24*MOS*

Final scores were publicly released to ensure transparency and fairness in the evaluation process.

2.4 Statistical analysis

Associations between each automatic metric and the MOS were assessed using Spearman's rank correlation coefficient (ρ), computed within each team (n = 50 cases) and pooled across all team–case pairs (7 teams × 50 cases; n = 350). To evaluate metric behavior on clinically consequential errors, we identified team–case pairs in which the main diagnosis, invasion depth, or histologic grade was misclassified relative to the ground-truth report and computed, for each metric, the mean absolute deviation from the MOS (|metric − MOS|; both scales range from 0 to 1). The same quantity was computed over all 350 pairs, over pairs in which every feature was correct, and within each single-feature error subset. Analyses were performed in R version 4.4.1 (R Foundation for Statistical Computing).

3. Results

3.1 Competing solutions

This section describes the participating approaches. **Additional file 1: Table S1** summarizes methods that achieved a final score > 0.7. The original abstracts submitted by each team are provided in **Additional file 2**.

ADCT [24] The winning team employed a two-step strategy that decoupled slide-image analysis from report generation. For analysis, they built an ensemble of two multiple-instance learning (MIL) models for multi-label classification: one attention-based [25], and one Transformer-based [26], the latter integrated into a teacher-student knowledge distillation framework. For feature extraction, they utilized two pathology foundation models (PFMs), GigaPath [27] and UNI [28], processing patches at 1.0x and 0.5x magnification, respectively. Preprocessing included background filtering via RGB thresholds and noise removal using Canny edge detection [29]. For the generation phase, they used a T5 text-to-text model [30]. To overcome data limitations, the input to this model was a simplified string of classification predictions, and the training data was augmented to enhance stability and performance.

Ewha Medical AI Team The second-place team adopted a fully end-to-end vision–language model approach, integrating a pathology-specific image encoder with a pretrained language model. They used CONCH [12] as the foundational patch-level feature extractor. Their architecture combined HistGen's [31] local-global hierarchical visual encoder to aggregate patch features into a WSI-level representation, which was then passed to a BLIP [32] language decoder to generate the final report. For preprocessing, they adapted the CLAM [5] methodology but omitted contour verification, which they found enabled capture of more detailed cellular features.

Joeunbeen The third-ranked team implemented a multi-model classification pipeline rather than a single end-to-end generative model. They decomposed report generation into four classification problems and trained four Inception_v3-based models [33], pretrained on ImageNet [34]: (1) a multi-label model for main diagnosis and granuloma; (2) a binary model for squamous differentiation; (3) a multi-class model for invasion level; and (4) a binary model for the presence of muscle proper. Their image preprocessing involved resizing each WSI, splitting it into 300x300 pixel patches, and filtering out background-heavy patches. Final reports were constructed by retrieving predefined text snippets from a dictionary based on the collective predictions of the four models, rather than generated end-to-end.

3.2 Performance

The REG2024 challenge received registrations from 10 teams, with 6 submissions for the public test and 7 submissions for the private test. Performance across multiple metrics (Section 2.3) is summarized in Table 3. Violin and box plots illustrating the ranking score and MOS are presented in **Additional file 1: Figure S4**. A detailed quantitative evaluation on the private-test set (n = 50) was performed for each model.

Table 3
Results of scores by team

Team	Ranking score_Pu	Ranking score_Pr	MOS	Accuracy	Completeness	Final score
ADCT	0.8258	0.8842	0.9178	0.684	0.2338	0.8806
Ewha Medical AI Team	0.8378	0.8618	0.908	0.669	0.239	0.8681
Joeunbeen	0.8218	0.8601	0.8816	0.647	0.2346	0.8576
AIVIS	0.6797	0.8642	0.9048	0.665	0.2398	0.8417
Byulbit	0.5856	0.8067	0.8946	0.672	0.2226	0.7836

The scores achieved by each participating team, highlighting their performance across the different evaluation metrics and final rankings. Only teams with a final score > 0.7 are shown.

Abbreviations: MOS: Mean opinion score, Ranking score_Pu: ranking scores on the public-test sets, Ranking score_Pr: ranking scores on the private-test sets

ADCT achieved the highest final score of 0.8806. The model demonstrated strong results in major diagnostic categories, particularly main diagnosis (48/50) and invasion depth (41/50). For minor features, there were no false positives, but all positives were missed: squamous differentiation 0/4 detected and chronic granulomatous inflammation 0/3 detected.

The Ewha Medical AI Team secured second place with a final score of 0.8681. The model demonstrated strong performance in main diagnostic categories (46/50) and detected all cases of granulomatous inflammation (3/3). **Joeunbeen** ranked third with a final score of 0.8576. The model showed stable performance in invasion-depth classification and also achieved complete detection of granulomatous inflammation (3/3).

On the private test, both AIVIS and Byulbit performed strongly, with high main-diagnosis accuracy and invasion-depth concordance; however, their lower public-test scores placed them fourth (AIVIS) and fifth (Byulbit) overall. Detailed category-wise accuracy for each team is presented in Table 4 and **Additional file1: Table S2**.

Table 4
Category-Wise Accuracy for Each Team on Private Test Set

Team	Main diagnosis	Invasion depth	Grade	Carcinoma in situ	Muscle proper	Granulomatous inflammation	Squamous differentiation
ADCT	96 (48/50)	82 (41/50)	94 (47/50)	92 (46/50)	98 (49/50)	94 (47/50)	92 (46/50)
Ewha Medical AI Team	92 (46/50)	80 (40/50)	94 (47/50)	86 (43/50)	98 (49/50)	100 (50/50)	92 (46/50)
Joeunbeen	82 (41/50)	84 (42/50)	90 (45/50)	86 (43/50)	90 (45/50)	98 (49/50)	90 (45/50)
AIVIS	90 (45/50)	86 (43/50)	94 (47/50)	78 (39/50)	92 (46/50)	96 (48/50)	94 (47/50)
Byulbit	90 (45/50)	90 (45/50)	90 (45/50)	80 (40/50)	96 (48/50)	96 (48/50)	92 (46/50)

Only teams with a final score > 0.7 are shown.

3.3 Qualitative evaluation

Additional file 1: Supplementary Data 2 presents representative WSIs, ground-truth (GT) pathology reports, and AI-generated reports submitted by participating teams. This qualitative evaluation reveals both notable successes and technical limitations.

In Case 1, a focus of subepithelial connective tissue invasion was identifiable only at high magnification. Both **AIVIS** and **Byulbit** generated reports stating “invasive urothelial carcinoma with involvement of subepithelial connective tissue,” along with concomitant “urothelial carcinoma in situ.” In contrast, the other three teams failed to detect these localized features, leading to incorrect main diagnoses. Similarly, in Case 7, which included squamous differentiation confined to a single high-power field, two teams (**Ewha** and **AIVIS**) successfully identified this focal feature, which is easily overlooked on routine review.

In Cases 2 and 3, the GT labeled invasion depth as subepithelial connective tissue; however, four models (Case 2) and three models (Case 3) generated reports describing invasion of the muscle proper. These discrepancies were considered to reflect the boundary ambiguity between subepithelial connective tissue and muscle proper in bladder histology. In Case 2, the GT report stated that the specimen did not include muscle proper, whereas all five models reported its presence. Subsequent WSI review confirmed that muscle proper was indeed present. These findings emphasize the inherent limitations of human interpretation during GT curation.

Cases 4

6 were examples in which chronic granulomatous inflammation was observed. Given the rarity of this long-tail pattern, model sensitivity was reduced. Misclassification occurred in Cases 4 and 5, where chronic inflammation was erroneously identified as granulomatous inflammation. Conversely, in Case 6, where well-formed granulomatous inflammation was present, several teams failed to detect it. Such errors reflect challenges posed by rare histopathologic patterns during training and inference.

Overall, these case-based analyses demonstrate that while AI models can achieve pathologist-level performance for certain subtle features, they remain vulnerable to limitations arising from histologic ambiguity and long-tail patterns.

3.4 Statistical analysis of evaluation metrics

The weighted-sum score showed a high Spearman correlation with MOS (ρ = 0.65–0.91; ROUGE-L, 0.68–0.96); per-team coefficients are presented in Fig. 2. Pairwise correlations among the metrics within each team are presented in **Additional file 1: Figure S5**. Across all team–case pairs from the seven teams (n = 350), ROUGE-L showed the strongest overall correlation with MOS (ρ = 0.92), followed closely by the weighted-sum score (ρ = 0.90); because the same 50 cases contribute to every team, these pooled coefficients are descriptive.

On the 132 of 350 team–case pairs in which the main diagnosis, invasion depth, or grade was misclassified, the weighted-sum score showed the closest absolute agreement with the MOS among the five metrics (mean |metric − MOS| = 0.093, versus 0.103 for ROUGE-L, 0.167 for embedding, 0.169 for keyword, and 0.249 for BLEU-4; Fig. 3a). The same ordering was observed for main-diagnosis errors (0.087 vs. 0.100 for ROUGE-L), invasion-depth errors (0.063 vs. 0.094), and grade errors (0.119 vs. 0.125) (**Additional file 1: Table S3**). The deviations also differed in direction (Fig. 3b): on reports with clinically important errors, the embedding score overestimated the MOS by 0.14–0.20 on average, and ROUGE-L overestimated it when invasion depth was misclassified, whereas the weighted-sum score was nearly unbiased. By contrast, BLEU-4 and the keyword score underestimated the MOS even on fully correct reports.

Additional file 1: Table S4 presents eight representative examples of generated and ground-truth reports, along with the MOS and five automatic measures: BLEU-4, ROUGE-L, keyword score, embedding score, and the weighted-sum score. In the first example, the model incorrectly generated "invasive urothelial carcinoma with involvement of subepithelial connective tissue" instead of the correct diagnosis, "non-invasive papillary urothelial carcinoma, high grade." This misclassification failed to capture both the main diagnosis and the invasion depth, resulting in an MOS of 0.57. Notably, BLEU-4 was low (0.25), whereas the embedding score was high (0.86); the weighted-sum score (0.56) aligned closely with the MOS.

In the second example, the proposed model failed to detect “chronic granulomatous inflammation with foreign body reaction.” Because this is a pathologically minor feature, the MOS remained high (0.94) despite lower scores across most metrics except embedding.

4. Discussion

4.1 Analysis of models

The REG2024 challenge highlighted a key divergence in strategy: end-to-end generative models versus multi-step classification-then-generation pipelines. Teams such as Ewha, AIVIS, and Byulbit adopted end-to-end vision–language architectures that generate reports directly from image features. In contrast, the winning team, ADCT, used a two-step approach that first performed multi-label classification with an ensemble of MIL models and then generated text from the structured predictions. This decoupled strategy proved highly effective, achieving the top MOS and final score. Its strength lay in optimizing diagnostic accuracy for heavily weighted, clinically significant categories such as main diagnosis. However, by excluding rare, imbalanced categories (e.g., squamous differentiation, granulomatous inflammation) from the classification task to maintain stability, the

model showed a 100% false negative rate for these findings. Conversely, end-to-end models, particularly Ewha's, demonstrated a stronger capacity for detecting these long-tail features, successfully identifying all cases of granulomatous inflammation.

A second consistent finding was the importance of PFMs as feature extractors. Four of the top five teams leveraged powerful PFMs, including GigaPath [27], UNI [28], and CONCH [12], which consistently supported high performance in major diagnostic categories. Architectural innovations built upon these PFMs were also crucial. For instance, the hierarchical encoders used by Ewha and Byulbit, based on the HistGen [31] architecture, effectively aggregated local patch information into a global, slide-level context. Such strategies for handling spatial and contextual information at the WSI-level appear essential for achieving high diagnostic fidelity in automated report generation.

4.2 Evaluation metrics

Each automatic Natural Language Generation (NLG) evaluation metric has strengths and, importantly, limitations. Word-based scores (BLEU-4 and ROUGE-L) focus on lexical overlaps and may not fully capture semantic similarity. Keyword score prioritizes key-term matching but does not assess overall text quality or contextual coherence; for example, when the model incorrectly predicted the presence of muscle proper, Keyword score assigned a perfect 1.0 solely because the term appeared in the output.

Embedding score emphasizes semantic consistency rather than exact word matching; for instance, when "Non-invasive papillary urothelial carcinoma, low grade" was misclassified as "high grade," the embedding score remained exceptionally high (0.9966). These discrepancies highlight the limitations of individual metrics and support a weighted-sum approach that integrates multiple signals to provide a more balanced, clinically meaningful assessment of AI-generated pathology reports.

In REG2024, which focused on bladder-cancer report generation, both the report formats and the vocabulary were relatively constrained; accordingly, the applicable "pathology text dictionary" was restricted. Under these conditions, ROUGE-L—which evaluates similarity via the longest common subsequence— achieved a marginally higher correlation with MOS than the weighted-sum score (0.92 vs. 0.90). However, on cases with diagnostically critical misclassifications—in main diagnosis, invasion depth, and grade—the weighted-sum score showed the closest absolute agreement with the MOS among all metrics.

Notably, "Chronic granulomatous inflammation with foreign body reaction," although pathologically minor, is a relatively long phrase; when the model failed to detect it, automatic metrics decreased (range, 0.4545–0.8388), whereas MOS remained high (0.94). This discrepancy underscores the need for evaluation metrics in the medical domain that more directly reflect clinical significance.

4.3 Strengths and Limitations of the REG2024 challenge

Our results showed a high correlation between the weighted-sum automatic evaluation metrics and expert-based MOS, and closer agreement with expert judgment than any single metric on clinically consequential errors. Moreover, the curated dataset provides a standardized benchmark for the field, thereby facilitating reproducibility and further development of automated pathology report-generation models.

Despite these contributions, the REG2024 challenge has several limitations. First, the dataset consisted exclusively of bladder-cancer tissue, eliminating the need for organ-prediction tasks and potentially reducing task

complexity, which may have contributed to the overall high accuracy of the generated reports. Second, all cases were sourced from a single institution, which may limit generalizability across diverse clinical settings. Third, the limited number of minor features led most models to exhibit lower detection rates for such findings, highlighting the difficulty of capturing subtle pathology; such long-tail problems are common in medical AI, where rare but clinically significant features are underrepresented in training data. Fourth, although the pathologist-based evaluation captured overall diagnostic accuracy, hallucinations—findings unsupported by the input WSI—were not separately quantified.

To address the first two limitations, we subsequently organized a follow-up challenge, REG2025, at MICCAI 2025, using a multi-organ, multi-center Pan-Asia dataset (https://reg2025.grand-challenge.org/). Notably, the same prespecified weighted-sum metric was adopted as the primary evaluation metric in REG2025, supporting the transferability of the proposed composite beyond a single-organ benchmark. Addressing the remaining limitations will require more sophisticated FMs to mitigate distributional bias and improve rare-feature detection, together with dedicated factuality metrics for hallucination assessment. Its successor, REG² (the 2026 edition), currently underway, takes a step in this direction by extending the benchmark toward the evaluation of diagnostic reasoning over WSIs (https://reg2026.grand-challenge.org/).

5. Conclusions

The REG2024 challenge was designed to advance pathology report generation by leveraging high-quality WSIs paired with expertly curated pathology reports. Participants submitted competitive, innovative solutions, with top models achieving high accuracy on key diagnostic categories. We proposed a weighted-sum metric, used alongside MOS, as a complementary approach to standard NLP measures, which may not fully capture diagnostic accuracy and clinical significance.

We expect this benchmark to advance CPath research, stimulate future community challenges, and support the development of robust report-generation models suitable for integration into real-world diagnostic workflows.

Abbreviations

AI: artificial intelligence; CAP: College of American Pathologists; CPath: computational pathology; FM: foundation model; GT: ground truth; MIL: multiple-instance learning; MOS: mean opinion score; NLG: natural language generation; PFM: pathology foundation model; VLM: vision–language model; WHO: World Health Organization; WSI: whole-slide image

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Review Board of Korea University Anam Hospital (IRB No. 2023AN0243). The requirement for informed consent was waived by the IRB owing to the retrospective use of fully de-identified data.

Consent for publication

Not applicable.

Availability of data and materials

The datasets are available from the corresponding author upon reasonable request. Data will be shared with qualified researchers under a restricted-access agreement after review and approval. The REG2024 dataset, comprising the whole-slide images and paired pathology reports, is available at https://drive.google.com/drive/folders/1IdMQisW7URqhSTY2CwEXhYO8EHwIdOEU?usp=drive_link. The evaluation code is available at https://github.com/PathfinderLab/REG.

Competing interests

The authors declare no competing interests.

Funding

This research was supported by the Digital-Bio AI + X Global Innovative Talent Nurturing Project with Hands-on Experience of the National Research Foundation (NRF), funded by the Korean government (MSIT) (No. RS2024-00441029); the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02215813); a grant of Korea University Anam Hospital, Seoul, Republic of Korea (O2400091); and the Medical data-driven hospital support project through the Korea Health Information Service (KHIS), funded by the Ministry of Health & Welfare, Republic of Korea. Also, this research was supported by a grant from Korea University Anam Hospital, Seoul, Republic of Korea (O2616711). The funders had no role in the study design, collection, analysis, and interpretation of data, writing of the report, or decision to submit the article for publication.

Authors' Contributions

SA, SHL, WKJ, and JHK conceived and designed the study. KAK, SH, SY, YK, HSS, YL, TV, WL, JHC, and EJ contributed to the acquisition, analysis, or interpretation of data. HO and EK drafted the manuscript and performed the statistical analysis. SA and JHK obtained funding. EK and HL provided administrative, technical, or material support. All authors critically revised the manuscript for important intellectual content. SA supervised the study. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Srinidhi CL, Ciga O, Martel AL: **Deep neural network models for computational histopathology: A survey**. *Medical image analysis* 2021, **67**:101813.
2. Amin MB, Edge SB, Greene FL, Byrd DR, Brookland RK, Washington MK, Gershenwald JE, Compton CC, Hess KR, Sullivan DC: **AJCC cancer staging manual**, vol. 1024: Springer; 2017.
3. **Cancer Protocol Templates: CAP Protocols and Guidelines** [https://www.cap.org/protocols-and-guidelines/cancer-reporting-tools/cancer-protocol-templates]
4. Märkl B, Füzesi L, Huss R, Bauer S, Schaller T: **Number of pathologists in Germany: comparison with European countries, USA, and Canada**. *Virchows Archiv* 2021, **478**(2):335-341.

5. Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F: **Data-efficient and weakly supervised computational pathology on whole-slide images**. *Nature biomedical engineering* 2021, **5**(6):555-570.
6. Bulten W, Kartasalo K, Chen P-HC, Ström P, Pinckaers H, Nagpal K, Cai Y, Steiner DF, Van Boven H, Vink R: **Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge**. *Nature medicine* 2022, **28**(1):154-163.
7. Wang X, Zhao J, Marostica E, Yuan W, Jin J, Zhang J, Li R, Tang H, Wang K, Li Y: **A pathology foundation model for cancer diagnosis and prognosis prediction**. *Nature* 2024, **634**(8035):970-978.
8. Chen RJ, Lu MY, Williamson DF, Chen TY, Lipkova J, Noor Z, Shaban M, Shady M, Williams M, Joo B: **Pan-cancer integrative histology-genomic analysis via multimodal deep learning**. *Cancer cell* 2022, **40**(8):865-878. e866.
9. Wagner SJ, Reisenbüchler D, West NP, Niehues JM, Zhu J, Foersch S, Veldhuizen GP, Quirke P, Grabsch HI, van den Brandt PA: **Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study**. *Cancer cell* 2023, **41**(9):1650-1661. e1654.
10. Shamai G, Livne A, Polónia A, Sabo E, Cretu A, Bar-Sela G, Kimmel R: **Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer**. *Nature Communications* 2022, **13**(1):6753.
11. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ: **Multimodal biomedical AI**. *Nature medicine* 2022, **28**(9):1773-1784.
12. Lu MY, Chen B, Williamson DF, Chen RJ, Liang I, Ding T, Jaume G, Odintsov I, Le LP, Gerber G: **A visual-language foundation model for computational pathology**. *Nature medicine* 2024, **30**(3):863-874.
13. Saab K, Tu T, Weng W-H, Tanno R, Stutz D, Wulczyn E, Zhang F, Strother T, Park C, Vedadi E: **Capabilities of gemini models in medicine**. *arXiv preprint arXiv:240418416* 2024.
14. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A: **Emerging properties in self-supervised vision transformers**. In: *Proceedings of the IEEE/CVF international conference on computer vision: 2021*; 2021: 9650-9660.
15. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, Fernandez P, Haziza D, Massa F, El-Nouby A: **Dinov2: Learning robust visual features without supervision**. *arXiv preprint arXiv:230407193* 2023.
16. Li D, Wan G, Wu X, Wu X, Nirmal AJ, Lian CG, Sorger PK, Semenov YR, Zhao C: **A Survey on Computational Pathology Foundation Models: Datasets, Adaptation Strategies, and Evaluation Tasks**. *arXiv preprint arXiv:250115724* 2025.
17. Asadi-Aghbolaghi M, Farahani H, Zhang A, Akbari A, Kim S, Chow A, Dane S, Consortium OC, Consortium O, Huntsman DG: **Machine learning-driven histotype diagnosis of ovarian carcinoma: insights from the OCEAN AI challenge**. *medRxiv* 2024:2024.2004. 2019.24306099.
18. Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, Marami B, Prastawa M, Chan M, Donovan M: **Bach: Grand challenge on breast cancer histology images**. *Medical image analysis* 2019, **56**:122-139.
19. van Rijthoven M, Aswolinskiy W, Tessier L, Balkenhol M, Bogaerts JM, Drubay D, Blesa LC, Peeters D, Stovgaard ES, Lænkholm A-V: **Tumor-infiltrating lymphocytes in breast cancer through artificial intelligence: biomarker analysis from the results of the TIGER challenge**. *medRxiv* 2025:2025.2002. 2028.25323078.
20. Lin C-Y: **Rouge: A package for automatic evaluation of summaries**. In: *Text summarization branches out: 2004*; 2004: 74-81.

21. Papineni K, Roukos S, Ward T, Zhu W-J: **Bleu: a method for automatic evaluation of machine translation**. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics: 2002*; 2002: 311-318.
22. Neumann M, King D, Beltagy I, Ammar W: **ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing**. In*: August 2019; Florence, Italy*: Association for Computational Linguistics; 2019: 319-327.
23. Pal MSA, Sankarasubbu M: **Openbiollms: Advancing open-source large language models for healthcare and life sciences**. In.; 2024.
24. Kim KA, Hong S, Yoo S, Kang Y, Shim HS: **Enhancing Structured Pathology Report Generation With Foundation Model and Modular Design**. *IEEE Access* 2025.
25. Ilse M, Tomczak J, Welling M: **Attention-based deep multiple instance learning**. In: *International conference on machine learning: 2018*: PMLR; 2018: 2127-2136.
26. Shao Z, Bian H, Chen Y, Wang Y, Zhang J, Ji X: **Transmil: Transformer based correlated multiple instance learning for whole slide image classification**. *Advances in neural information processing systems* 2021, **34**:2136-2147.
27. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, Wong C, Gero Z, González J, Gu Y: **A whole-slide foundation model for digital pathology from real-world data**. *Nature* 2024, **630**(8015):181-188.
28. Chen RJ, Ding T, Lu MY, Williamson DF, Jaume G, Song AH, Chen B, Zhang A, Shao D, Shaban M: **Towards a general-purpose foundation model for computational pathology**. *Nature medicine* 2024, **30**(3):850-862.
29. Canny J: **A computational approach to edge detection**. *IEEE Transactions on pattern analysis and machine intelligence* 2009(6):679-698.
30. Ni J, Abrego GH, Constant N, Ma J, Hall KB, Cer D, Yang Y: **Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models**. *arXiv preprint arXiv:210808877* 2021.
31. Guo Z, Ma J, Xu Y, Wang Y, Wang L, Chen H: **HistGen: Histopathology Report Generation via Local-Global Feature Encoding and Cross-Modal Context Interaction**. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI: 2024//; Cham*: Springer Nature Switzerland: 189-199.
32. Li J, Li D, Xiong C, Hoi S: **Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation**. In: *International conference on machine learning: 2022*: PMLR; 2022: 12888-12900.
33. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: **Rethinking the inception architecture for computer vision**. In: *Proceedings of the IEEE conference on computer vision and pattern recognition: 2016*; 2016: 2818-2826.
34. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L: **Imagenet: A large-scale hierarchical image database**. In: *2009 IEEE conference on computer vision and pattern recognition: 2009*: Ieee; 2009: 248-255.

Figures

Image not available with this version

Figure 1

Examples of paired whole slide images and pathology report

a. Subepithelial connective tissue invasion. b. Presence of muscle proper. c. Muscle proper invasion. d. Squamous differentiation. e. Urothelial carcinoma in situ. f. Non-invasive papillary urothelial carcinoma. g. Chronic granulomatous inflammation with foreign body reaction.

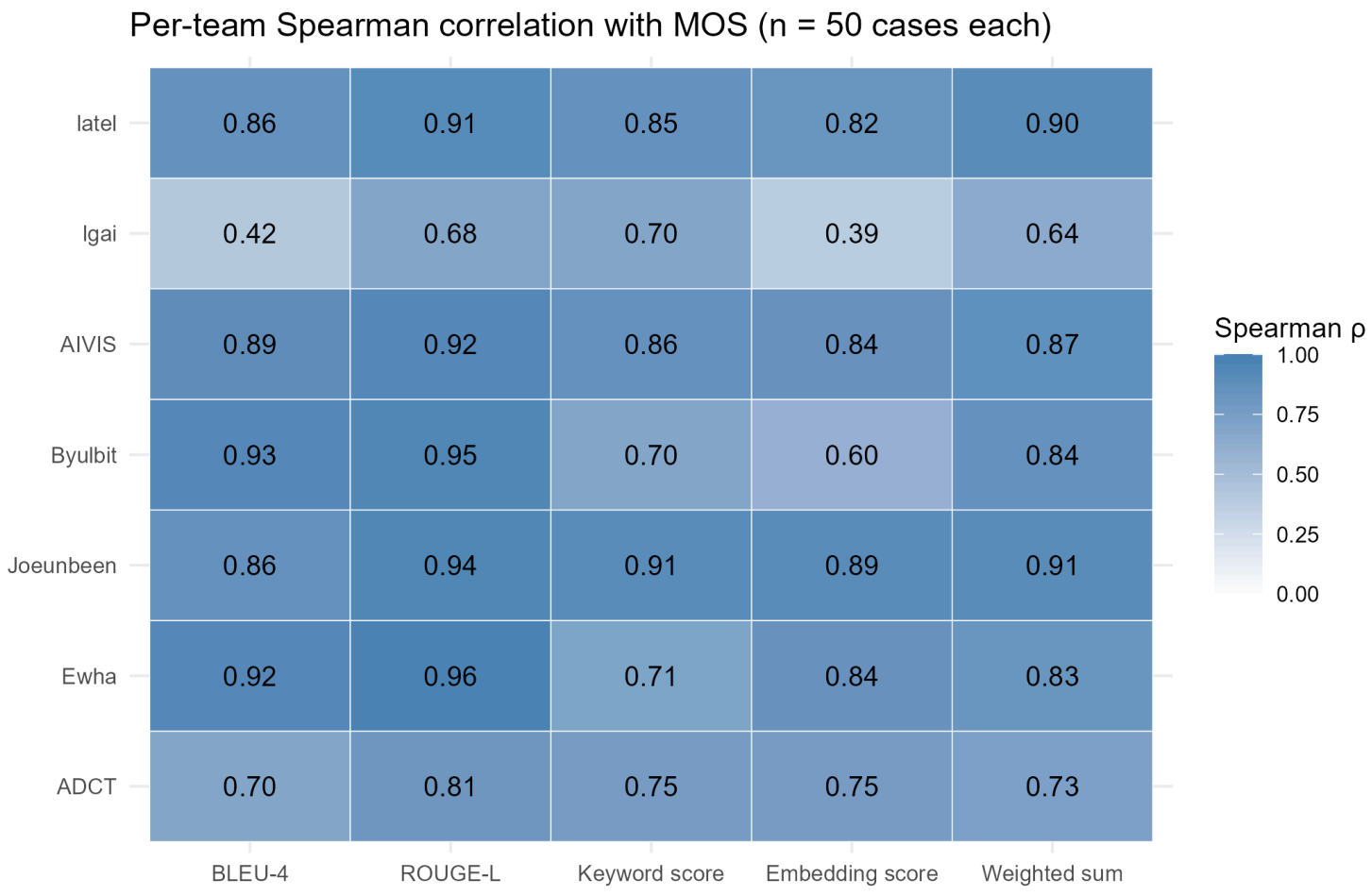


Figure 2

Per-team spearman correlations between automatic metrics and the expert mean opinion score

Each cell shows Spearman's rank correlation coefficient (ρ) between the indicated automatic metric and the expert mean opinion score (MOS), computed across the 50 private-test cases separately for each of the seven participating teams.

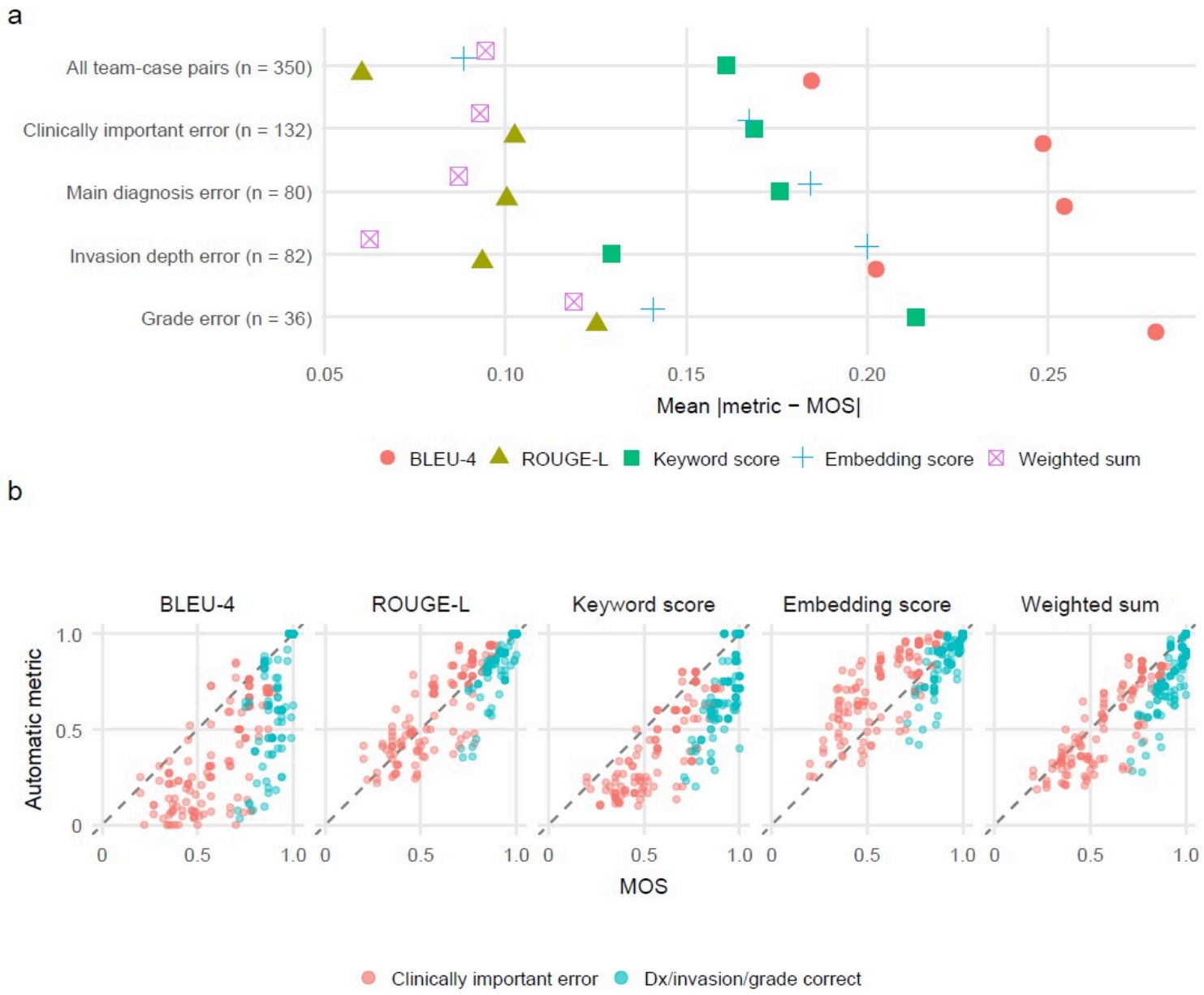


Figure 3

Agreement between automatic metrics and the expert mean opinion score (MOS).

a. Mean absolute deviation from the MOS (|metric − MOS|) over all pairs and within subsets with a clinically important misclassification. Lower values indicate closer agreement with expert judgment. Results for all error strata are provided in **Additional file 1: Table S3.** b. Each point represents one team–case pair (n = 350; 7 teams × 50 private-test cases), colored by whether the generated report contained a clinically important misclassification (main diagnosis, invasion depth, or histologic grade).

Abbreviations: Dx; main diagnosis

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalData2Team4byulbit.docx
- AdditionalData2Team3EunbeenJo.docx
- AdditionalData2Team5AIVIS.docx
- AdditionalData2Team2EWHA.docx
- AdditionalData2Team1ADCT.docx
- AdditionalMaterials.docx
- Additionalfiles.docx