Overview of #SMM4H-HeaRD 2026 – Task 6: Predicting TNM staging from pathology reports

Jose Miguel Acitores Cortina[1,2], Jacob Berkowitz[1,2], Nadine A. Friedrich[1,2,3], Nicholas P. Tatonetti[1,2],

[1]Department of Computational Biomedicine, Cedars-Sinai
[2]Cedars-Sinai Cancer, Cedars-Sinai
[3]Department of Urology, Cedars-Sinai

Correspondence: nicholas.tatonetti@csmc.edu

Abstract

This paper provides an overview of Task 6 from the Social Media Mining for Health/Health Real-World Data shared task (#SMM4H-HeaRD 2026), which focused on predicting TNM staging from pathology reports from TCGA. Seven teams submitted systems spanning fine-tuned clinical encoders, open-source generative LLMs, and closed-source API models. On a straightforward test set, most teams achieved near-perfect F1 scores (average 0.993, 0.972, and 0.957 for T, N, and M). However, on a harder tiebreak set where explicit TNM notation was removed and staging had to be inferred from clinical descriptions, performance dropped substantially (average 0.725, 0.783, and 0.846). Notably, the two teams using large closed-source API models generalized best to the harder set, achieving the highest T and N scores despite not leading on the easy set. These results suggest that while fine-tuned domain-specific encoders excel at surface-level extraction, larger general-purpose LLMs may be more robust when staging must be inferred from contextual clinical findings. All teams surpassed baseline overall performance on both test sets.

1 Introduction

The TNM staging system is the global standard for describing the extent of cancer spread (Amin et al., 2017). Its three components capture the size and local extent of the primary tumor (T), the involvement of regional lymph nodes (N), and the presence of distant metastasis (M). Stage drives prognosis, guides treatment selection, and determines eligibility for clinical trials (Kefeli et al., 2024; Kefeli and Tatonetti, 2024). Despite this central role, stage is rarely recorded in structured fields of the electronic health record. Instead, staging information is often embedded in unstructured pathology report text, where it may appear in varying formats, abbreviations, and levels of explicitness across institutions and pathologists.

Tumor registries are the current solution. Trained specialists read pathology reports and clinical notes and assign stage by hand. This process can take up to six months from the date of diagnosis to the point at which a structured stage value becomes available (White et al., 2017; Edwards et al., 2022), and the workforce of registry specialists is shrinking (Rollison et al., 2022). By the time a patient's stage is recorded, the window for trial enrollment may already be closed, and the patient may be missing from retrospective analyses that filter on stage. Automating extraction directly from pathology report text would reduce this delay, broaden the pool of patients available for trial matching, and make large retrospective cohort construction more tractable.

The Cancer Genome Atlas (TCGA) provides a large public source of de-identified diagnostic pathology reports that have been converted into cleaned, machine-readable text. The TCGA-TNM corpus links a subset of these reports to pathologic T, N, and M component labels. This resource makes it possible to train and compare models for TNM extraction at scale, while also raising an important evaluation challenge: because TCGA is public, held-out TCGA reports may have been seen by modern language models during pretraining.

Prior work has approached TNM extraction with rule-based methods, traditional machine learning, and more recently transformer encoders (Yala et al., 2017; Glaser et al., 2018; Abedian et al., 2021; Preston et al., 2023). The most directly relevant system is BB-TEN (Kefeli et al., 2024), which fine-tuned a long-context clinical encoder on TCGA-TNM pathology reports and generalized off-the-shelf to an external institution, achieving AU-ROC scores from 0.815 to 0.942 across the three components. Since BB-TEN was published, the set of plausible candidates for this task has expanded considerably. Open-weight generative LLMs are now available

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 332–337

at a range of sizes and with biomedical pretraining, longer-context clinical encoders have been released, and frontier API models can be prompted with full-length reports. However, no shared benchmark exists for comparing these approaches head to head on TNM extraction, and most existing evaluations cannot separate surface extraction of explicit staging tokens from genuine inference based on clinical findings.

We organized Task 6 of #SMM4H-HeaRD 2026, which focused on predicting cancer staging from pathology reports using training data drawn from the TCGA-TNM corpus and evaluating each component independently. Seven teams submitted systems spanning fine-tuned clinical encoders, parameter-efficient adaptation of open generative LLMs, and closed-source API models. The remainder of this paper describes the task and data, the BB-TEN baseline, the participant systems, the results across the evaluation and tiebreaker sets, and what we learned about the current state of TNM extraction from pathology reports.

2 Shared Task

2.1 Dataset generation

Both training and test data derive from the TCGA pathology report corpus and its associated T, N, and M labels. The training set consists of real TCGA reports released to participants with their labels. The released development data were split into training and validation files. Because T, N, and M labels are evaluated as separate prediction tasks and not every report has all three component labels, the number of labeled examples differs by component. The training file contained 5,853 reports with T labels, 4,826 with N labels, and 3,916 with M labels. The validation file contained 1,034 reports with T labels, 852 with N labels, and 692 with M labels, for totals of 6,887, 5,678, and 4,608 labeled examples for T, N, and M, respectively. The test sets are synthetic notes generated from TCGA labels and styled after TCGA reports. We chose synthesis over a real held-out split because TCGA is publicly available, and a real test set drawn from the same corpus would be vulnerable to external memorization. Generating new notes under controlled labels lets us release an evaluation set with known ground truth that the participating models could not have seen during pretraining.

Each synthetic note was produced by sampling one TCGA row to supply the target T, N, and M values and a different TCGA row to serve as a style exemplar. A generator LLM was prompted to write a new note expressing the target labels in the exemplar's style. Returned outputs were parsed as JSON and validated against the target labels. If the returned labels did not match, we resampled the style exemplar and retried up to a fixed number of attempts. Generation used Azure OpenAI with temperature 0.8, JSON-mode output, and a fixed sampling seed. We tracked every (label-source, style-source) pair and rejected any combination that had appeared in a prior run, including across the easy and hard sets. Only TCGA rows with all three labels present were eligible as label sources or style exemplars. T values are indexed from 0 in the released CSVs but were presented to the generator as clinical T1 through T4 and mapped back on save.

Test set 1 was generated with a prompt that instructed the model to embed target T, N, and M values naturally in the note. Note that the prompt did not require explicit staging tokens such as pT2 or cM0, but the generator produced them occasionally. We generated 100 notes with GPT-5.4 for scoring and 2,500 additional notes with GPT-5.4-mini that we released alongside as decoys, with no indication to participants of which subset would be used for evaluation. The total released for test set 1 set was 2,600 notes. The hard set was designed as a tiebreaker for teams achieving perfect scores on the easy set, with the surface-extraction shortcuts removed. Its prompt forbids any T, N, M, Stage, TNM, or AJCC notation. The note must instead describe tumor size and depth of invasion, the count and size of positive lymph nodes, and the presence or absence of distant lesions, so that a clinician reader can infer the stage from findings alone. We required distractors such as prior cancer history, family history, and differential diagnoses, and we required staging-relevant findings to be scattered across sections rather than consolidated in one place. We generated 50 notes with GPT-5.4 for scoring and 250 with GPT-5.4-mini as decoys, releasing 300 notes in total.

A single clinician reviewed the 100 scored easy notes and the 50 scored hard notes, confirming that the described findings were consistent with the target T, N, and M values. Notes that failed review were excluded from scoring. Single-annotator review establishes face validity for the labels but does not measure inter-rater reliability, which we note as a limitation.

Team	Test set 1			Test set 2 - tiebreak		
	$F1_T$	$F1_N$	$F1_M$	$F1_T$	$F1_N$	$F1_M$
URJC	**1.000**	**1.000**	**1.000**	0.810	0.770	**1.000**
GoBlueinformatics	**1.000**	**1.000**	**1.000**	0.626	0.758	**1.000**
LLATMU	**1.000**	**1.000**	**1.000**	0.697	0.783	0.617
MedMind	0.996	0.998	0.997	**0.849**	**0.865**	**1.000**
CUETDiagNLP	0.970	0.926	0.954	0.700	0.774	0.640
Blue	**1.000**	0.895	0.828	0.638	0.650	0.507
CaresAI	0.978	0.957	0.879	0.626	0.758	**1.000**
BBTEN (Baseline)	0.992	0.783	0.796	0.454	0.591	0.554

Table 1: F1 scores by team across tasks.

2.2 Baseline

The baseline system for this shared task is BB-TEN (Big Bird – TNM staging Extracted from Notes)(Kefeli et al., 2024). BB-TEN uses Clinical-BigBird, a BERT-variant with a sparse attention mechanism that extends the maximum input sequence from 512 to 4,096 tokens, enabling it to process lengthy pathology reports without truncation. The model was pre-trained on MIMIC-III clinical notes and then fine-tuned on approximately 7,000 publicly available TCGA pathology reports spanning 23 cancer types, with separate classification heads trained for each TNM component: tumor size (T1–4), regional lymph node involvement (N0–3), and distant metastasis (M0–1).

The authors demonstrated strong generalizability by applying the TCGA-trained models directly to nearly 8,000 independent pathology reports from Columbia University Medical Center without any institution-specific fine-tuning, achieving AU-ROC scores between 0.815 and 0.942. Notably, the Clinical-BigBird architecture outperformed both the shorter-context ClinicalBERT and a fine-tuned Llama 3 model on two of three classification tasks, while requiring substantially less training time and computational resources. The trained models are publicly available on HuggingFace.

2.3 Overview of Participant Approaches

A wide variety of approaches were used to tackle the extraction of TNM staging from clinical notes. All participants included some type of LLM, and their approaches can be grouped into three categories based on the underlying language model.

Domain-Adapted Encoders. Several teams fine-tuned encoder-only transformers pre-trained on clinical corpora. *LLATMU* (Hsiao et al., 2026) used BioClinical ModernBERT-Large, a long-context encoder supporting up to 8,192 tokens. *Blue* (Sharma et al., 2026) employed Clinical-BigBird, a sparse-attention model handling sequences up to 4,096 tokens. *CUETDiagNLP* (Dey et al., 2026) used GatorTron, and *GoBlueinformatics* (Wei, 2026) fine-tuned BioClinical ModernBERT-Large as part of a hybrid pipeline.

Open-Source Generative LLMs. Other teams adapted decoder-only LLMs via parameter-efficient fine-tuning. *URJC* (Madrueño et al., 2026) applied supervised fine-tuning to Qwen2.5-27B, while *GoBlueinformatics* used LoRA on OpenBioLLM-8B (a Llama-3-based biomedical model) alongside their encoder, combining generative and discriminative approaches.

Closed-Source API-Based Models. *MedMind* (Pradhan and Habersberger, 2026) used closed-source models, GPT-5.4-mini, requiring no local training but introducing external dependencies and reducing reproducibility.

Pipelines and Training Strategies. Participants combined model fine-tuning with task-specific post-processing. Encoder systems generally trained separate classifiers for T, N, and M, while generative systems produced structured outputs that were mapped back to component labels. Several teams addressed class imbalance with weighting, focal loss, label smoothing, or ensembling, and some added regex-based rules to capture explicit TNM strings before applying neural models to unresolved cases. Finally, *CaresAI*(Abubakar et al., 2026) explored a traditional machine learning pipeline using TF-IDF features, pretrained BERT embeddings, and stacked classifiers.

2.4 Evaluation

The evaluation consisted of two different test sets. The first test set was designed as a straightforward evaluation; most teams achieved F1 scores at or near 1.0 on each label. Excluding the baseline, the average scores across participating teams that submitted a manuscript were 0.993 ± 0.011, 0.972 ± 0.041, and 0.957 ± 0.062 for T, N, and M respectively. Four teams—*URJC*, *GoBlueinformatics*, *LLATMU*, and *Blue*—achieved a perfect 1.000 $F1_T$, and the first three of these also achieved perfect scores across all three axes.

Given the near-ceiling performance on the first test set, we developed a harder second test set with less explicit TNM staging information, requiring models to infer staging from contextual clinical descriptions rather than extracting verbatim mentions. This allowed for a much larger stratification of the participants' results. On this tiebreak set, the average scores dropped substantially to 0.725 ± 0.094, 0.783 ± 0.075, and 0.846 ± 0.194 for T, N, and M respectively.

Notably, the team using closed-source API-based models—*MedMind* (GPT-5.4-mini)—generalized best to the harder test set, achieving the highest T and N scores despite not leading on the first test set. *MedMind* achieved the overall best tiebreak performance with F1 scores of 0.849, 0.865, and 1.000 for T, N, and M. This suggests that larger general-purpose LLMs with zero-shot or few-shot prompting may be more robust to distribution shifts than fine-tuned domain-specific encoders, which showed larger performance drops between the two test sets. Among fine-tuned encoder systems, *URJC*'s hybrid regex-plus-LLM pipeline proved most resilient, ranking third overall on the tiebreak set. The M axis was generally the easiest to classify on the harder set, with five teams achieving a perfect 1.000 $F1_M$, likely due to its binary nature. A more detailed view of the results can be seen in Table 1.

3 Discussion

Six of seven teams reached or approached perfect F1 on test set 1, and every team dropped substantially on the tiebreaker. On TCGA-style pathology reports where explicit staging tokens are present in the text, TNM extraction is essentially solved by current systems. The harder problem, and the one closer to what a deployed system would face, is inferring stage from clinical findings when the report does not state it directly.

MedMind with GPT-5.4-mini, led the tiebreaker despite trailing on test set 1. The fine-tuned domain-specific encoders, which were built for clinical text, fell off more sharply when explicit staging tokens were removed. Large general-purpose LLMs may carry enough medical reasoning to handle stage inference under prompting, while smaller encoders fine-tuned on the surface form generalize less well to a setting where the surface form is gone. One caveat applies. The scored subset of test set 1 was generated with GPT-5.4 and the decoys with GPT-5.4-mini, which gave models in the GPT-5.x family a stylistic advantage on that set. The tiebreaker was generated the same way, but its target labels are the genuine T, N, and M values from TCGA, validated by a clinician, so the result is not a self-recognition artifact. Future iterations should generate notes with a model family disjoint from any expected participant systems to remove the question entirely.

Performance varied by component. M was the easiest axis on the tiebreaker, with five teams achieving perfect F1, because distant metastasis tends to be documented as discrete findings such as a hepatic lesion or a bone lytic deposit, and because M is binary. T was the hardest, since determining T1 through T4 requires reasoning over tumor size in centimeters and depth of invasion into specific anatomical layers. Every participating team beat the BB-TEN baseline on both test sets. On the tiebreaker, the gap between BB-TEN and the leading teams on T (0.45 versus 0.86) reflects two years of progress in long-context modeling and prompting strategies for clinical text.

Several limitations qualify these results. Label validation rested on a single clinician, which establishes face validity but does not measure inter-rater reliability. The test notes, while clinically reviewed, are synthetic and may not capture the full messiness of real clinical text, including incomplete reports, contradictory findings between sections, dictation artifacts, and institution-specific templates. A deployed system would face all of these. Future iterations of this task should add a small held-out set of real reports from a non-public source to test for that gap directly. They should also broaden cancer types and rebalance toward rare classes such as N3 and M1, which remain underrepresented.

References

Sajjad Abedian, Evan T. Sholle, Prakash M. Adekkanattu, Marika M. Cusick, Stephanie E. Weiner, Jonathan E. Shoag, Jim C. Hu, and Thomas R. Campion. 2021. Automated extraction of tumor staging and diagnosis information from surgical pathology reports. *JCO Clinical Cancer Informatics*, 5:1054–1061.

Joseph Itopa Abubakar, Jorge Creiann Jarme, Favour Igwezeke, and Mary Adewunmi. 2026. Caresai at smm4h-heard 2026: Predicting tnm staging. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Mahul B. Amin, Stephen B. Edge, Frederick L. Greene, David R. Byrd, Robert K. Brookland, Mary K. Washington, Jeffrey E. Gershenwald, Carolyn C. Compton, Kenneth R. Hess, Daniel C. Sullivan, J. Milburn Jessup, James D. Brierley, Lauri E. Gaspar, Richard L. Schilsky, Charles M. Balch, David P. Winchester, Elliot A. Asare, Martin Madera, Donna M. Gress, and Laura R. Meyer, editors. 2017. *AJCC Cancer Staging Manual*, 8 edition. Springer, Cham.

Shuva Dey, Priyangshu Barua, and Mohammad Ashfak Habib. 2026. Cuet diagnlp at #smm4h-heard 2026: Per-axis tnm staging from pathology reports and opioid impact span detection from social media. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Patrick Edwards, Amarilys Bernacet, Florence K. L. Tangka, Paran Pordell, Jenny Beizer, Reda Wilson, Wendy Blumenthal, Sandra F. Jones, Maggie Cole-Beebe, and Sujha Subramanian. 2022. Operational characteristics of central cancer registries that support the generation of high-quality surveillance data. *Journal of Registry Management*, 49(1):10–16.

Alexander P. Glaser, Brian J. Jordan, Jason Cohen, Anuj Desai, Philip Silberman, and Joshua J. Meeks. 2018. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. *JCO Clinical Cancer Informatics*, 2:1–8.

Eric Hsiao, Min-Hsuan Ku, and Hsuan-Lei Shao. 2026. Llatmu at #smm4h-heard 2026: Clinical text structuring across qlora-based generation and partial-label classification. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Jenna Kefeli, Jacob Berkowitz, Jose M. Acitores Cortina, Kevin K. Tsang, and Nicholas P. Tatonetti. 2024. Generalizable and automated classification of TNM stage from pathology reports with external validation. *Nature Communications*, 15(1):8916.

Jenna Kefeli and Nicholas Tatonetti. 2024. TCGA-Reports: A machine-readable pathology report resource for benchmarking text-based AI models. *Patterns*, 5(3):100933.

Natalia Madrueño, Jose Walter Hernández Pérez, Rubén R. Fernández, and Soto Montalvo. 2026. Urjcteam at #smm4h-heard 2026: Tnm stage extraction with a regex-llm workflow. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Aatish Pradhan and Brian Habersberger. 2026. Medmind ai at #smm4h-heard 2026: Data extraction and generation using prompt engineering and structured outputs (tasks 1–6). In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Sam Preston, Mu Wei, Rajesh Rao, Robert Tinn, Naoto Usuyama, Michael Lucas, Yu Gu, Roshanthi Weerasinghe, Soohee Lee, Brian Piening, Paul Tittel, Naveen Valluri, Tristan Naumann, Carlo Bifulco, and Hoifung Poon. 2023. Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision. *Patterns*, 4(4):100726.

Dana E. Rollison, Gary M. Levin, Jeremy L. Warner, Rich Pinder, Lori A. Havener, Madhusmita Behera, Andrew R. Post, Rajan Gopalakrishnan, and Eric B. Durbin. 2022. Current and emerging informatics initiatives impactful to cancer registries. *Journal of Registry Management*, 49(4):153–160.

Krish Sharma, Rhea Singhal, and Jatin Bedi. 2026. blue at smm4h-heard 2026: Class-weighted transformer ensembles with structured decoding and chain-of-thought blending across six health nlp shared tasks. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Shangqing Wei. 2026. Goblueinformatics at #smm4h-heard 2026: Long-context encoders and generative biomedical llms for pathological tnm stage prediction. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Mary C. White, Frances Babcock, Nikki S. Hayes, Angela B. Mariotto, Faye L. Wong, Betsy A. Kohler, and Hannah K. Weir. 2017. The history and use of cancer registry data by public health cancer control programs in the United States. *Cancer*, 123(S24):4969–4976.

Adam Yala, Regina Barzilay, Laura Salama, Molly Griffin, Grace Sollender, Aditya Bardia, Constance Lehman, Julliette M. Buckley, Suzanne B. Coopey, Fernanda Polubriaginof, Judy E. Garber, Barbara L.

Smith, Michele A. Gadd, Michelle C. Specht, Thomas M. Gudewicz, Anthony J. Guidi, Alphonse Taghian, and Kevin S. Hughes. 2017. Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment*, 161(2):203–211.