Sakarya Tıp Dergisi
Sakarya Medical Journal

e-ISSN : 2146-409X
Publisher : Sakarya University

Vol. 16, No. 3, 346-353, 2026
DOI: http://doi.org/10.31832/smj.1889486

Research Article

How Artificial, How Intelligent? LLMs and Medical Experts Face Off in TMA Diagnosis

Gizem Zorlu Görgülügil[1*], Ceyda Kaçar[1], Elif Şen[1], Ayça İnci[2], Volkan Karakuş[3]

[1] Antalya Training and Research Hospital, Department of Internal Medicine, Antalya, Türkiye
gizemzorlug@gmail.com,
ceydakacar0@gmail.com,
es.elifsen1998@gmail.com

[2] Akdeniz University, Faculty of Medicine, Department of Nephrology, Antalya,Türkiye
aycainci2004@hotmail.com

[3] Antalya Training and Research Hospital, Department of Hematology, Antalya, Türkiye
dr_v_karakus@yahoo.com

* Corresponding Author

Abstract

Objective: To compare physician specialists and large language models in the differential diagnosis and plasma exchange decision making of standardized thrombotic microangiopathy case vignettes.

Methods: This case vignette based comparative study included three standardized TMA scenarios representing atypical hemolytic uremic syndrome, malignant hypertension associated TMA, and acquired thrombotic thrombocytopenic purpura. Nine evaluators participated: two nephrologists, two hematologists, two internal medicine specialists, and three large language models (ChatGPT-5.2, Gemini Pro, and Microsoft Copilot). For each case, participants provided a most likely diagnosis, three differential diagnoses, supporting clinical reasoning, and a plasma exchange decision. Diagnostic accuracy, plasma exchange appropriateness, and critical management errors were analyzed descriptively and comparatively.

Results: A total of 27 independent evaluations were analyzed. Overall Top-1 diagnostic accuracy was 51.9%. Accuracy varied by case, with the lowest performance observed in the aHUS scenario (22.2%) and the highest in malignant hypertension–associated TMA (77.8%). Physicians achieved a Top-1 accuracy of 50.0% (9/18), whereas LLMs demonstrated 55.6% accuracy (5/9). Plasma exchange decision accuracy was 63.0% overall, with overtreatment more common in non-TTP cases. Ten critical management errors (37.0%) were identified. Agreement for plasma exchange decisions demonstrated fair concordance ($\kappa = 0.29$).

Conclusions: Substantial variability exists in both diagnostic classification and therapeutic decision-making in TMA scenarios. While LLMs demonstrated competitive diagnostic performance in selected cases, management variability persisted across evaluator groups. These findings highlight the diagnostic complexity of TMA and support the role of artificial intelligence as an adjunctive clinical decision-support tool rather than an autonomous decision-maker.

Keywords: Thrombotic microangiopathy, Differential diagnosis, Large language models, Clinical decision making

Received: 15.02.2026
Accepted: 01.05.2026
Published Online: 02.09.2026

1. INTRODUCTION

Thrombotic microangiopathy (TMA) is a life-threatening syndrome characterized by microangiopathic hemolytic anemia, thrombocytopenia, and organ injury. Early differentiation between thrombotic thrombocytopenic purpura (TTP), atypical hemolytic uremic syndrome (aHUS), and secondary forms such as malignant hypertension associated TMA is critical, as treatment strategies particularly plasma exchange differ substantially and directly affect survival.[1]

Although ADAMTS13 testing has improved diagnostic precision, treatment decisions are often made before laboratory confirmation.[2] Secondary TMA conditions, including malignant hypertension and complement mediated disorders, may closely mimic TTP, leading to diagnostic uncertainty and potential overtreatment.[3] Structured models such as the TMA-INSIGHT score aim to improve classification but variability persists in practice.[4]

Machine learning approaches have shown promise in TTP risk prediction and diagnostic triage,[5] and laboratory based decision tools incorporating PLASMIC related features have been proposed to support clinical workflows.[6] Advances in AI assisted schistocyte detection further highlight the integration of computational tools in TMA evaluation.[7]

More recently, large language models (LLMs) have demonstrated the ability to generate differential diagnoses from narrative case descriptions.[8] While performance in complex diagnostic challenges has been encouraging,[9] data specific to high-risk hematologic emergencies remain limited. The potential influence of LLM outputs on physician reasoning has also been suggested.[10]

Cite as: Zorlu Görgülügil G, Kaçar C, Şen E, İnci A, Karakuş V. How artificial, how intelligent? LLMs and medical experts face Off in TMA diagnosis. Sakarya Med J. 2026;16(3):346-53. doi:10.31832/smj.1889486.

This is an open access paper distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License.

Given the diagnostic complexity of TMA and the clinical consequences of inappropriate plasma exchange, evaluating both physician and LLM decision making in standardized TMA scenarios is warranted.

2. MATERIALS AND METHODS

2.1. Study design and objective

This study was designed as a cross-sectional, case-vignette–based comparative evaluation of diagnostic reasoning and clinical decision-making in thrombotic microangiopathy (TMA). The primary objective was to compare the diagnostic performance and management decisions of physician specialists and large language models (LLMs) when presented with standardized TMA scenarios. The study aimed to evaluate differences in diagnostic prioritization and plasma exchange decisions rather than real world clinical outcomes.

2.2. Case development

Three anonymized TMA case scenarios were constructed from previously encountered clinical cases and reformatted into standardized structured vignettes. All identifying patient information was removed. Each vignette included presenting symptoms, physical examination findings, laboratory parameters relevant to hemolysis and renal function, and pertinent follow-up information.

The first case described a 50 year old woman presenting with diarrhea and abdominal pain who subsequently developed thrombocytopenia, acute kidney injury, elevated lactate dehydrogenase (LDH), and increasing schistocytes on peripheral smear. Neurological imaging was unremarkable, although electroencephalography suggested possible focal epileptic activity. This scenario was constructed to represent atypical hemolytic uremic syndrome (aHUS) and to challenge differentiation from thrombotic thrombocytopenic purpura (TTP).

The second case involved a 44 year old woman with severe uncontrolled hypertension (222/168 mmHg), grade 3 hypertensive retinopathy, microangiopathic hemolytic anemia, thrombocytopenia, and acute kidney injury with normal coagulation parameters. This vignette was designed to represent malignant hypertension–associated TMA and to assess recognition of secondary TMA in contrast to primary TTP.

The third case described a 36 year old man with profound thrombocytopenia (10,000/mm^3), marked schistocytosis (12%), elevated LDH, hyperbilirubinemia, and recent pesticide exposure. Renal function was relatively preserved. This scenario was constructed to represent acquired TTP requiring urgent plasma exchange.

The vignette-based approach was chosen to ensure standardization across evaluators and to enable controlled comparison of diagnostic reasoning. However, this design does not fully capture the dynamic nature of real world clinical decision making.

2.3. Participants and evaluation procedure

Nine evaluators participated in the study, including two nephrologists, two hematologists, two internal medicine specialists, and three large language models (ChatGPT-5.2, Gemini Pro, and Microsoft Copilot) [Figure 1]. All physicians had a minimum of five years of clinical experience. Each evaluator independently assessed all three cases. Physicians were blinded to one another's responses. LLMs were prompted in separate sessions to avoid contextual carryover between cases.

For each vignette, evaluators were asked to provide their most likely diagnosis, three prioritized differential diagnoses, clinical and laboratory findings supporting or refuting their diagnostic reasoning, and whether plasma exchange therapy was indicated. A brief rationale for the treatment decision was also requested. Responses were recorded verbatim without modification, and missing responses were retained as missing values in the analysis.

All large language models were evaluated using a standardized zero-shot prompt. The same clinical information was provided to each model. Each model was accessed in an independent session without memory carryover to avoid contextual bias. Model versions (ChatGPT-5.2, Gemini Pro, and Microsoft Copilot) and access dates were recorded.

Figure 1.

Radial representation of evaluator groups assessing standardized TMA case vignettes

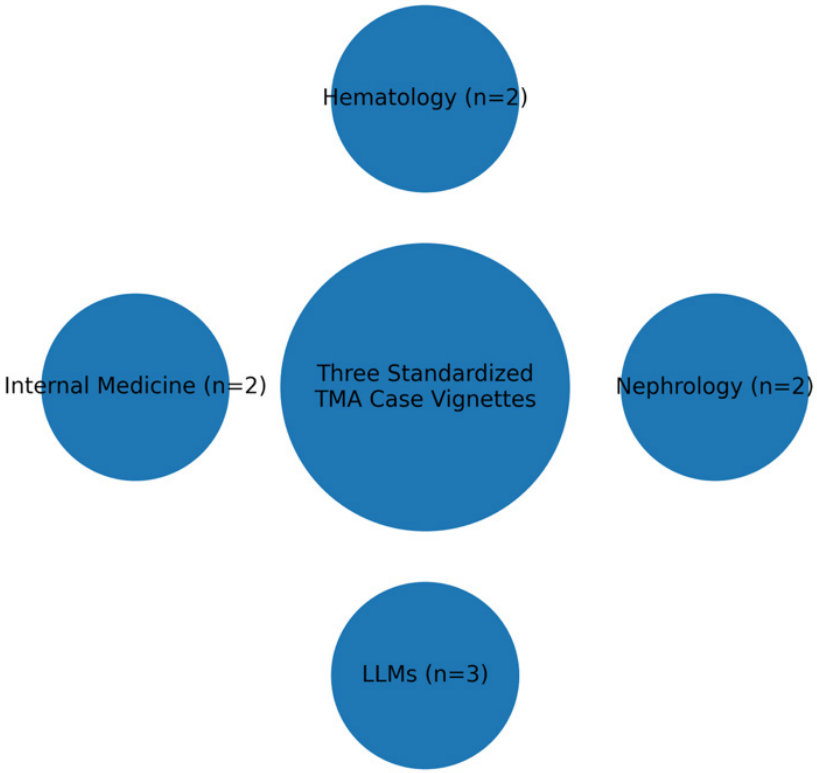


Table 1.

Structured evaluation components requested from each evaluator for standardized TMA case vignettes

Component	What Was Requested	Purpose in Analysis
Top-1 Diagnosis	Most likely diagnosis for each case	Primary diagnostic accuracy
Top-3 Differentials	Three prioritized differential diagnoses	Secondary diagnostic accuracy
Clinical & Lab Reasoning	Findings supporting/refuting diagnosis	Qualitative reasoning analysis
Plasma Exchange Decision	Yes / No / Uncertain	Management accuracy & error rate
Treatment Justification	Short rationale for decision	Critical error interpretation

2.4. LLM evaluation and prompt standardization

All large language models were evaluated using a standardized zero-shot prompt. The same clinical information was provided to each model. Each model was accessed in a separate session without memory carryover to avoid contextual bias.

Models were accessed through their standard user interfaces (ChatGPT-5.2, Gemini Pro, and Microsoft Copilot). System-level parameters such as temperature, system role, and web access were not externally controlled, and this limitation has been acknowledged.

No predefined word, token, or response-length limit was imposed; however, all models were instructed to follow the same structured response format.

2.5. Reference standard and outcome definitions

Reference diagnoses were determined by investigator consensus based on established clinical criteria and current literature. Case 1 was classified as atypical hemolytic uremic syndrome, Case 2 as malignant hypertension associated TMA, and Case 3 as acquired TTP.

Correct plasma exchange decisions were defined as not indicated in Cases 1 and 2 and indicated in Case 3.

The primary outcome was Top-1 diagnostic accuracy, defined as agreement between the evaluator's most likely diagnosis and the predefined reference diagnosis. Secondary outcomes included Top-3 diagnostic accuracy, defined as the presence of the reference diagnosis within the three listed differential diagnoses;

accuracy of plasma exchange decision; and the rate of critical management errors. A critical management error was defined as recommending plasma exchange in Cases 1 or 2 or failing to recommend plasma exchange in Case 3.

2.6. Statistical analysis

All analyses were conducted at the case evaluator level, yielding 27 total evaluations. Categorical variables were summarized as frequencies and percentages. Given the small sample size and pilot nature of the study, comparisons between physicians and LLMs were considered exploratory. Fisher's exact test was used where appropriate. A two sided p-value of less than 0.05 was considered statistically significant. Emphasis was placed on descriptive trends and effect sizes rather than inferential conclusions. Sensitivity and specificity analyses were not performed due to the limited sample size and the vignette-based design of the study. Given the exploratory nature of the study, effect size–based interpretation was prioritized over inferential statistical conclusions.

2.7. Ethics

Ethical approval was obtained from the Scientific Research Ethics Committee of Antalya Training and Research Hospital (Decision No: 3/28; Date: 13 February 2025). All physician participants provided informed consent prior to participation. The study adhered to the Declaration of Helsinki. The clinical scenarios were based on fully anonymized data.

3. RESULTS

3.1. Overall diagnostic performance

A total of 27 independent evaluations (9 evaluators × 3 cases) were included in the analysis. The overall Top-1 diagnostic accuracy across all evaluations was 51.9% (14/27).

Diagnostic accuracy varied substantially across cases (Table 2). Case 1 (atypical hemolytic uremic syndrome, aHUS) demonstrated the lowest accuracy, with only 2 of 9 evaluators (22.2%) correctly identifying the predefined reference diagnosis. In contrast, Case 2 (malignant hypertension associated TMA) showed the highest accuracy, with 7 of 9 evaluators (77.8%) providing the correct Top-1 diagnosis. Case 3 (acquired TTP) yielded an intermediate accuracy of 55.6% (5/9).

Table 2.

Top-1 diagnostic accuracy by case

Case	Correct / Total	Accuracy (%)
Case 1 (aHUS)	2 / 9	22.2
Case 2 (HT-TMA)	7 / 9	77.8
Case 3 (TTP)	5 / 9	55.6
Overall	14 / 27	51.9

The particularly low accuracy in Case 1 reflects the diagnostic difficulty in differentiating atypical HUS from TTP in overlapping microangiopathic presentations.

3.2. Physician versus LLM performance

When stratified by evaluator type, physicians (6 evaluators; 18 total case evaluations) achieved a Top-1 diagnostic accuracy of 50.0% (9/18), whereas large language models (3 evaluators; 9 total evaluations) achieved a Top-1 accuracy of 55.6% (5/9) (Table 3).

Table 3.

Top-1 diagnostic accuracy by evaluator type

Evaluator Group	Correct / Total	Accuracy (%)
Physicians	9 / 18	50.0
LLMs	5 / 9	55.6

Overall Top-1 diagnostic accuracy was numerically similar between physicians and LLMs (50.0% vs. 55.6%, respectively). In Case 2, all three LLMs correctly identified malignant hypertension–associated TMA, whereas two physician evaluations misclassified the case. However, given the limited sample size, Fisher's exact test did not demonstrate a statistically significant difference between evaluator groups ($p > 0.05$), and these comparisons should therefore be interpreted as exploratory.

3.3. Subgroup analysis by medical specialty

Diagnostic performance varied among physician specialties. Internal medicine specialists achieved the highest Top-1 accuracy at 66.7% (4/6 evaluations). Hematologists demonstrated an intermediate performance of 50.0% (3/6), while nephrologists showed a lower accuracy of 33.3% (2/6).

Notably, both hematologists correctly identified TTP in Case 3, whereas neither nephrologist selected TTP as the most likely diagnosis in that case. Internal medicine specialists consistently identified malignant hypertension associated TMA in Case 2.

These findings suggest variability in diagnostic prioritization across specialties, although small subgroup sizes preclude definitive conclusions.

3.4. Top-3 diagnostic accuracy

Top-3 diagnostic accuracy exceeded Top-1 accuracy across evaluator groups. In multiple instances where the Top-1 diagnosis was incorrect, the reference diagnosis was included within the differential list, indicating partial recognition of the underlying pathology. This pattern was particularly observed in Case 3, where TTP frequently appeared among differential diagnoses even when not selected as the most likely diagnosis.

3.5. Plasma exchange decisions

Overall plasma exchange decision accuracy was 63.0% (17/27 evaluations).

In Case 1 (aHUS; plasma exchange not indicated), 6 of 9 evaluators (66.7%) recommended plasma exchange, reflecting a substantial overtreatment tendency. In Case 2 (malignant hypertension–associated TMA; plasma exchange not indicated), inappropriate plasma exchange was recommended in 22.2% of evaluations (2/9). In Case 3 (TTP; plasma exchange indicated), 2 evaluators (22.2%) failed to recommend plasma exchange, representing potential undertreatment.

Plasma exchange recommendations varied by specialty. Hematologists and nephrologists more frequently recommended plasma exchange in ambiguous cases, whereas internal medicine specialists and LLMs demonstrated relatively more selective use.

3.6. Critical management errors

A total of 10 critical management errors (37.0%) were identified across 27 evaluations. Most errors resulted from unnecessary plasma exchange initiation in non-TTP cases rather than failure to initiate plasma exchange in confirmed TTP.

Physicians demonstrated a stronger inclination toward precautionary plasma exchange pending ADAMTS13 testing, reflecting a safety-oriented approach. In contrast, LLMs displayed comparatively more restrained plasma exchange recommendations, though occasional diagnostic imprecision persisted.

3.7. Agreement analysis

Agreement between evaluators and the predefined reference standard for plasma exchange decisions was assessed using Cohen's kappa. According to the Landis and Koch classification, $\kappa = 0.29$ corresponds to fair agreement, indicating modest concordance among evaluators.

Observed agreement was 63.0%, with expected agreement by chance estimated at 48.1%, yielding a Cohen's κ of 0.29, indicating fair agreement. Given the multi-rater structure and limited sample size, kappa values should be interpreted cautiously.

3.8. Model-specific performance

Diagnostic performance varied among individual LLMs. ChatGPT-5.2 and Gemini Pro each correctly identified 2 of 3 cases (66.7%), while Microsoft Copilot correctly classified 1 of 3 cases (33.3%). Gemini demonstrated the highest plasma exchange decision accuracy (3/3 correct). Inter-model variability was most evident in Case 1, where diagnostic classification diverged between models.

4. DISCUSSION

This study demonstrates marked variability in both diagnostic classification and therapeutic decision-making among physicians and large language models (LLMs) in thrombotic microangiopathy (TMA) scenarios. The overall Top-1 diagnostic accuracy was modest, and performance was particularly low in the atypical hemolytic uremic syndrome (aHUS) case, reflecting the well-recognized clinical overlap between TTP, aHUS, and secondary forms of TMA.[1] Differentiating these entities remains one of the most challenging areas in acute hematologic care.

Although ADAMTS13 activity testing has substantially improved diagnostic precision in TTP, treatment decisions are frequently made before laboratory confirmation becomes available.[2] As a result, clinicians often face the dilemma of initiating plasma exchange empirically. In our study, the high rate of plasma exchange recommendations in non-TTP cases likely reflects a safety-oriented clinical reflex. Secondary TMA conditions, including malignant hypertension and complement-mediated disorders, may closely mimic TTP and contribute to overtreatment.[3] Similarly, complement-mediated TMA and hypertensive emergencies can present with overlapping phenotypes, further increasing diagnostic complexity.[11]

Structured classification models such as the TMA-INSIGHT score have been developed to reduce diagnostic ambiguity;[4] however, variability persists in real-world decision-making. The fair level of agreement observed in plasma exchange decisions (Cohen's $\kappa = 0.29$) underscores the limited standardization of management strategies in TMA.

Machine learning–based approaches have shown promise in TTP risk stratification and mortality prediction.[5] Laboratory-driven decision-support

algorithms incorporating PLASMIC-related features have also been proposed to optimize testing utilization and clinical workflows.[6] In addition, inflammatory index–guided models have demonstrated feasibility for rapid TTP recognition.[4] However, most of these systems rely on structured datasets and may not fully capture the contextual complexity of real-time clinical reasoning.

From a laboratory perspective, artificial intelligence–assisted schistocyte detection has emerged as an important development in TMA diagnostics.[7] Imaging flow cytometry combined with machine learning classifiers has demonstrated promising performance in high-throughput schistocyte quantification.[12] These advances illustrate the expanding integration of computational tools into hematologic workflows.

Regarding LLM performance, our findings indicate that these models can achieve competitive diagnostic accuracy in selected scenarios. Previous studies have shown that LLMs may perform meaningfully in complex diagnostic challenges,[9] and vignette-based evaluations have demonstrated partial but variable differential diagnosis accuracy.[8] Furthermore, LLM outputs in rare hematologic diseases have been shown to influence physician reasoning, although performance consistency remains heterogeneous.[10] In our study, LLMs performed particularly well in the malignant hypertension–associated TMA case but demonstrated greater diagnostic uncertainty in the more ambiguous aHUS scenario.

Specialty-based analysis further highlights the multidisciplinary nature of TMA management. While hematologists demonstrated higher accuracy in identifying TTP, variability across nephrology and internal medicine specialists suggests that clinical background and experiential factors influence diagnostic prioritization. These findings reinforce that TMA decision-making is shaped not only by laboratory data but also by cognitive and specialty-specific heuristics.

Taken together, these results highlight marked heterogeneity in both diagnostic classification and therapeutic decision-making in TMA scenarios. Although large language models demonstrated competitive diagnostic performance in certain cases, critical management variability persisted across all evaluator groups.[13] These findings emphasize the diagnostic complexity of TMA and suggest that both human and AI-based assessments remain susceptible to error in clinically ambiguous contexts.

Importantly, current LLMs function largely as static text-generation systems without dynamic evidence retrieval. Emerging Retrieval-Augmented Generation (RAG) architectures and agentic AI systems offer the potential to integrate real-time guideline retrieval, structured laboratory inputs, and stepwise decision planning.[14] In high-risk hematologic emergencies such as TMA, such hybrid systems may enhance contextual reasoning and reduce variability. However, issues of safety, interpretability, accountability, and clinical validation must be rigorously addressed before implementation. Future studies integrating retrieval-augmented and agentic AI systems may further improve contextual diagnostic reasoning.

5. CONCLUSION

In this comparative evaluation of standardized TMA case vignettes, both physician specialists and large language models demonstrated variability in diagnostic classification and plasma exchange decision-making. While LLMs achieved competitive diagnostic accuracy in selected scenarios, critical management variability persisted across evaluator groups.

These findings underscore the diagnostic complexity of TMA and suggest that artificial intelligence tools may serve as adjunctive decision-support systems rather than independent clinical decision-makers. Further studies incorporating larger case sets and integrated AI architectures are warranted to clarify the safe implementation of such systems in high risk hematologic settings.

Limitations

This study has several limitations. First, the use of three standardized case vignettes may not fully capture the complexity of real-world TMA presentations. Clinical decision-making in practice is influenced by dynamic patient evolution, serial laboratory data, multidisciplinary consultation, and contextual factors that cannot be entirely simulated in vignette-based designs.

Second, the relatively small sample size (27 total evaluations, including 9 LLM evaluations and 6 evaluations per specialty subgroup) limits statistical power and the generalizability of subgroup analyses. Therefore, the findings should be interpreted as exploratory.

Third, reference diagnoses were determined by investigator consensus rather than by an independent blinded adjudication panel, which may introduce potential bias. Although grounded in established diagnostic criteria, this approach may not fully account for diagnostic uncertainty inherent in TMA cases.

Additionally, the vignette-based design does not capture dynamic clinical processes such as serial laboratory changes, delayed ADAMTS13 results, or multidisciplinary decision-making.

Finally, LLM performance was assessed as standalone outputs without integration into structured clinical workflows or evidence retrieval systems. The results therefore reflect isolated model reasoning rather than performance within fully integrated decision-support environments.

Article Information Form

Author Contributions

Conceptualization: GZG; Methodology: GZG, Aİ, VK; Validation: GZG, CK, Aİ, VK; Formal analysis: GZG; Investigation: GZG, CK, EŞ, Aİ, VK; Resources: GZG, CK, EŞ, Aİ, VK; Data curation: GZG; Writing – original draft: GZG; Writing – review & editing: GZG, CK, EŞ, Aİ, VK; Visualization: GZG; Supervision: GZG, Aİ, VK; Project administration: GZG.

The Declaration of Conflict of Interest/ Common Interest

No conflict of interest or common interest has been declared by authors.

The Declaration of Ethics Committee Approval

This study was approved by the Scientific Research Ethics Committee of Antalya Training and Research Hospital (Decision No: 3/28; Date: February 13, 2025). The study was conducted in accordance with the principles of the Declaration of Helsinki.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Artificial Intelligence Statement

Artificial intelligence-based tools were used solely for language editing, translation, and the preparation of selected visual materials during manuscript preparation. All scientific content, data analysis, interpretation of results, and conclusions were developed and verified by the authors.

Copyright Statement

Authors own the copyright of their work published in the journal and their work is published under the CC BY-NC 4.0 license.

REFERENCES

1. Story CM, Gerber GF, Chaturvedi S. Medical consult: aHUS, TTP? How to distinguish and what to do. Hematology Am Soc Hematol Educ Program. 2023;2023(1):745-53. doi:10.1182/hematology.2023000501.

2. Papakonstantinou A, Kalmoukos P, Mpalaska A, Koravou EE, Gavriilaki E. ADAMTS13 in the new era of TTP. Int J Mol Sci. 2024;25(15):8137. doi:10.3390/ijms25158137.

3. Khanal N, Dahal S, Upadhyay S, Bhatt VR, Bierman PJ. Differentiating malignant hypertension-induced thrombotic microangiopathy from thrombotic thrombocytopenic purpura. Ther Adv Hematol. 2015;6(3):97-102. doi:10.1177/2040620715571076.

4. Addad VV, Palma LMP, Vaisbich MH, Pacheco Barbosa AM, da Rocha NC, de Almeida Cardoso MM, et al. A comprehensive model for assessing and classifying patients with thrombotic microangiopathy: the TMA-INSIGHT score. Thromb J. 2023;21:119. doi:10.1186/s12959-023-00564-6.

5. Abou-Ismail MY, Zhang C, Presson AP, Chaturvedi S, Antun AG, Farland AM, et al. A machine learning approach to predict mortality due to immune-mediated thrombotic thrombocytopenic purpura. Res Pract Thromb Haemost. 2024;8(3):102388. doi:10.1016/j.rpth.2024.102388.

6. Herrera Rivera N, McClintock DS, Alterman MA, Alterman TAL, Pruitt HD, Olsen GM, et al. A clinical laboratorian's journey in developing a machine learning algorithm to assist in testing utilization and stewardship. J Lab Precis Med. 2023;8:22. doi:10.21037/jlpm-23-11.

7. Zhang Z, Yang S, Wang X. Schistocyte detection in artificial intelligence age. Int J Lab Hematol. 2024;46(3):427-33. doi:10.1111/ijlh.14260.

8. Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Med Inform. 2023;11:e48808. doi:10.2196/48808.

9. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA. 2023;330(1):78-80. doi:10.1001/jama.2023.8288.

10. Yu H, Chen T, Zhang X, Yang Y, Liu Q, Yang C, et al. Performance of large language models in diagnosing rare hematologic diseases and the impact of their diagnostic outputs on physicians: combined retrospective and prospective study. J Med Internet Res. 2025;27:e77334. doi:10.2196/77334.

11. Cavero T, Arjona E, Soto K, Caravaca-Fontán F, Rabasco C, Bravo L, et al. Severe and malignant hypertension are common in primary atypical hemolytic uremic syndrome. Kidney Int. 2019;96(4):995-1004. doi:10.1016/j.kint.2019.05.014.

12. Demagny J, Roussel C, Le Guyader M, Guiheneuf E, Harrivel V, Boyer T, et al. Combining imaging flow cytometry and machine learning for high-throughput schistocyte quantification: a SVM classifier development and external validation cohort. EBioMedicine. 2022;83:104209. doi:10.1016/j.ebiom.2022.104209.

13. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. doi:10.1038/s41586-023-06291-2.

14. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459-74.