

Can ChatGPT Function as a Virtual Multidisciplinary Team? A Proof-of-Concept Study in Vascular Malformation Syndromes

Yingying Dai, Xu Liu, Chenglong Han, Qingfeng Liu, Yuwei Huang, Long Lin,
Min-kyu Kim, Linghong Guo, Xian Jiang

Submitted to: JMIR Formative Research
on: March 15, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 15

 Figures 16

 Figure 1..... 17

 Multimedia Appendixes 18

 Multimedia Appendix 1..... 19

Can ChatGPT Function as a Virtual Multidisciplinary Team? A Proof-of-Concept Study in Vascular Malformation Syndromes

Yingying Dai^{1, 2*}; Xu Liu^{1, 2*}; Chenglong Han^{1*}; Qingfeng Liu¹; Yuwei Huang¹; Long Lin¹; Min-kyu Kim¹; Linghong Guo^{1, 2}; Xian Jiang^{1, 2} Prof Dr Med

¹Department of Dermatology West China Hospital Sichuan University Chengdu CN

²Clinical Institute of Inflammation and Immunology, Laboratory of Dermatology Frontiers Science Center for Disease-related Molecular Network West China Hospital of Sichuan University Chengdu CN

*these authors contributed equally

Corresponding Author:

Xian Jiang Prof Dr Med
Department of Dermatology
West China Hospital
Sichuan University
No. 37 Guoxue Alley, Wuhou District
Chengdu
CN

Abstract

Background: Sturge-Weber syndrome (SWS) and Klippel-Trenaunay syndrome (KTS) are complex vascular malformation syndromes that require multidisciplinary team (MDT) management. However, the traditional MDT approach faces challenges such as time coordination, geographical barriers, and inefficiencies in cross-disciplinary communication.

Objective: This study aims to evaluate ChatGPT's potential in simulating MDT decision-making for SWS and KTS by comparing its diagnostic and treatment recommendations with traditional MDT conclusions.

Methods: A case-based proof-of-concept design was employed, retrospectively analyzing MDT records of two SWS and two KTS patients. Clinical data, imaging, and laboratory results were input into ChatGPT, and its outputs were evaluated by two dermatology experts using a 1-5 Likert scale across five dimensions: accuracy, completeness, appropriateness, insight, and safety.

Results: ChatGPT performed well in most dimensions, particularly in appropriateness, but showed occasional uncertainty in handling complex or rare cases, such as in gene-phenotype associations. Inter-rater reliability ranged from negligible to moderate (ICC -0.00 to 0.63), with no significant differences observed between the experts' ratings ($p > 0.05$).

Conclusions: ChatGPT shows strong potential as a tool for simulating MDT decision-making, particularly in diagnostic accuracy, completeness of recommendations, and cross-disciplinary insights. However, it has limitations in managing complex cases and ensuring the feasibility of its recommendations. Future studies with larger sample sizes and multi-center validation are needed to fully assess its clinical value.

(JMIR Preprints 15/03/2025:73989)

DOI: <https://doi.org/10.2196/preprints.73989>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all.
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#), my accepted manuscript PDF will be available to all.
No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Can ChatGPT Function as a Virtual Multidisciplinary Team? A Proof-of-Concept Study in Vascular Malformation Syndromes

Yingying Dai^{1,2†}, Liu Xu^{1,2†}, Han Chenglong^{1†}, Liu Qingfeng¹, Huang Yuwei¹, Long Lin¹, Min-kyu Kim¹, Linghong Guo^{1,2*}, Xian Jiang^{1,2*}

1 Department of Dermatology, West China Hospital, Sichuan University, Chengdu 610041, China

2 Laboratory of Dermatology, Clinical Institute of Inflammation and Immunology, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, Chengdu 610041, China

† These authors contributed equally to this work and are co-first authors.

* Corresponding authors: PhD. Linghong Guo (weidanfeng2002@126.com) and Prof. Xian Jiang (jiangxian@scu.edu.cn)

Abstract

Background: Sturge-Weber syndrome (SWS) and Klippel-Trenaunay syndrome (KTS) are complex vascular malformation syndromes that require multidisciplinary team (MDT) management. However, the traditional MDT approach faces challenges such as time coordination, geographical barriers, and inefficiencies in cross-disciplinary communication.

Objectives: This study aims to evaluate ChatGPT's potential in simulating MDT decision-making for SWS and KTS by comparing its diagnostic and treatment recommendations with traditional MDT conclusions.

Methods: A case-based proof-of-concept design was employed, retrospectively analyzing MDT records of two SWS and two KTS patients. Clinical data, imaging, and laboratory results were input into ChatGPT, and its outputs were evaluated by two dermatology experts using a 1-5 Likert scale across five dimensions: accuracy, completeness, appropriateness, insight, and safety.

Results: ChatGPT performed well in most dimensions, particularly in appropriateness, but showed occasional uncertainty in handling complex or rare cases, such as in gene-phenotype associations. Inter-rater reliability ranged from negligible to moderate (ICC -0.00 to 0.63), with no significant differences observed between the experts' ratings ($p > 0.05$).

Conclusions: ChatGPT shows strong potential as a tool for simulating MDT decision-making, particularly in diagnostic accuracy, completeness of recommendations, and cross-disciplinary insights. However, it has limitations in managing complex cases and ensuring the feasibility of its recommendations. Future studies with larger sample sizes and multi-center validation are needed to fully assess its clinical value.

Keyword: Sturge-Weber syndrome; Klippel-Trenaunay syndrome; ChatGPT; Artificial Intelligence; multidisciplinary team

Introduction

Background: Vascular Malformation Syndromes and the Role of MDTs

Sturge-Weber syndrome (SWS) and Klippel-Trenaunay syndrome (KTS) are typical multi-system vascular malformation syndromes (VMS), involving various organ systems such as the skin vasculature, nervous system, and musculoskeletal structures[1]. SWS is mainly characterized by facial port-wine stains, meningeal vascular malformations, and glaucoma, while KTS typically presents with cutaneous vascular malformations, varicose veins, and limb hypertrophy[2,3]. Due to the complex and highly heterogeneous clinical features of these syndromes, diagnosis and treatment often require a multidisciplinary team (MDT) approach, which integrates expertise and therapeutic strategies from multiple specialties[4–6]. Although the MDT approach aims to improve diagnostic and treatment outcomes through cross-disciplinary collaboration, its implementation faces significant

challenges. The traditional MDT mode usually requires fixed schedules, locations, and involves experts from diverse fields, such as dermatology, neurology, ophthalmology, and pediatric surgery[7]. However, practical issues—such as scheduling difficulties, geographical barriers, high resource consumption, and inefficient communication across disciplines—often lead to diagnostic delays or insufficiently personalized treatment plans, ultimately affecting patient prognosis[8,9]. These challenges are especially pronounced in rural and remote areas with limited healthcare resources. Therefore, there is an urgent need to explore innovative alternatives to the traditional MDT mode[10].

Artificial Intelligence and Its Potential in MDT Simulations

Recent advances in artificial intelligence (AI) have significantly influenced various fields of medicine, particularly in clinical decision-making[11]. One of the most promising AI tools is OpenAI's ChatGPT, a natural language processing model capable of integrating heterogeneous data, performing logical reasoning, and generating human-like responses. In the context of multidisciplinary clinical decision-making, GPT-based models have been validated for their ability to integrate genomic, transcriptomic, and imaging data, thereby supporting cancer diagnosis and prognosis prediction effectively[12,13]. This cross-modal integration capability lays the foundation for applying GPT models in the management of complex diseases, such as VMS, which often requires comprehensive interpretation of multi-omics data in the MDT setting. Notably, the recently introduced ChatGPTo1 model is specifically designed to handle complex reasoning tasks and effectively address challenges requiring advanced inference[14]. Although ChatGPT has shown considerable promise in clinical decision-making, its potential application in VMS, particularly in simulating the MDT process, remains underexplored.

Objectives of This Study

Given the complexity of managing SWS and KTS and the high demand for multidisciplinary collaboration, this study addresses two main questions: First, can ChatGPT generate diagnostic and treatment recommendations similar to those of traditional MDTs? Second, what are the feasibility and limitations of using ChatGPT to simulate MDTs, particularly in integrating cross-disciplinary information and offering personalized treatment plans? To address these questions, we employ a proof-of-concept approach, retrospectively analyzing MDT records of two KTS and two SWS patients. We simulate the MDT process with ChatGPT and compare ChatGPT-generated recommendations to actual MDT conclusions. This study provides initial evidence for AI-assisted management of complex VMS and lays the foundation for future large-scale, multi-center studies. Additionally, it offers insights into the development of innovative cross-disciplinary AI collaboration in healthcare.

Method

Study Design and Case Selection

This study adopts a case-based proof-of-concept design to retrospectively analyze the MDT records of two patients with SWS and two patients with KTS. Figure 1 presents a flowchart outlining the research process. Four cases were randomly selected from a pool of available records with complete clinical data, including patient history, imaging results, and necessary genetic testing

ChatGPT Simulation of MDT Consultation

The clinical data, imaging results, and laboratory findings were fed into ChatGPT (ChatGPTo1, accessed on December 28, 2024) following predefined input instructions to generate diagnostic and treatment recommendations (supplement file 1). All data were anonymized prior to input in accordance with patient privacy standards[15].

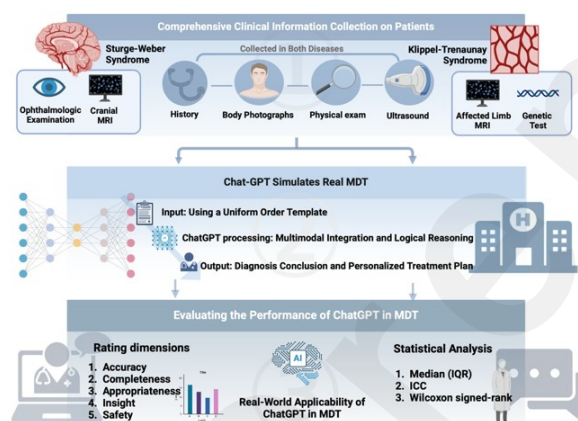
Expert Evaluation Criteria

Two dermatology experts (LQF, who participated in the MDT, and LX, who did not) independently rated the outputs from ChatGPT. The evaluation criteria included accuracy, completeness, appropriateness, insight, and safety[16], using a 1-5 Likert scale[17]. Experts also provided personal insights and feedback whenever discrepancies were found between ChatGPT's outputs and actual MDT conclusions, to help analyze the potential reasons for these discrepancies.

Statistical Analysis

The expert ratings will be summarized using descriptive statistics, including median and interquartile range (IQR), to demonstrate the distribution of scores. To assess inter-rater reliability, the Intraclass Correlation Coefficient (ICC) will be used to analyze the consistency between the two experts[18]. Additionally, the Wilcoxon signed-rank test will be employed to examine significant differences in ratings between the experts on the same dimensions[19]. Exact p-values were attempted for the Wilcoxon signed-rank test; however, due to the presence of tied ranks and zero differences in the small sample ($n=4$), normal approximation with continuity correction was applied. Statistical significance will be interpreted with caution, considering the reduced power in detecting differences due to the limited sample size.

Figure 1. Research framework for evaluating ChatGPT in multidisciplinary team consultations



Result

Expert Ratings of ChatGPT's Performance

This study conducted a multi-dimensional evaluation of the diagnostic and treatment suggestions generated by ChatGPT in MDT consultation, based on assessments by two clinical experts. A 1-5 Likert scale was used for each dimension, with the results presented in Table 1. Descriptive analysis indicated that the median score for Expert 1 (LX) ranged from 4.0 to 5.0 (Interquartile Range [IQR] = 0.0–0.3), with the appropriateness (Median = 5.0, IQR = 0.3) and safety (Median = 5.0, IQR = 0.0) dimensions both receiving the highest possible score. For Expert 2 (LQF), the median score ranged from 3.5 to 4.5 (IQR = 0.3–1.3), with the completeness dimension receiving the highest score (Median = 4.5, IQR = 1.3). Overall, ChatGPT's suggestions performed best in the appropriateness and safety dimensions, but there was considerable variability in expert ratings for the completeness and insight dimensions.

Inter-Rater Reliability

Intraclass correlation coefficient (ICC) analysis revealed that the consistency levels across dimensions ranged from negligible to moderate (ICC range: -0.00 to 0.63). According to Cicchetti's

standards (ICC < 0.40 is poor, 0.40–0.59 is moderate)[20], the highest consistency was observed in the accuracy dimension (ICC = 0.63, 95% CI: -0.20 to 0.97), while the completeness dimension showed negligible consistency (ICC = -0.00, 95% CI: -1.33 to 0.90). Notably, the lower bound of the ICC confidence interval with a negative value may reflect the instability of the statistical model due to the small sample size (n = 4) and should be interpreted with caution.

Statistical Comparison of Expert Ratings

Results from the Wilcoxon signed-rank test showed no statistically significant differences in ratings between the experts across all dimensions ($p > 0.05$). In the appropriateness dimension, despite some minor discrepancies in the experts' ratings (Expert 1: 4/5/5/5; Expert 2: 4/5/5/4), the p-value was 1.000. This result may be attributed to the small sample size (n = 4) and the presence of three pairs of zero differences, leading to insufficient statistical power. It suggests that there might be subtle subjective differences in the rating criteria between the experts, although these differences were not detected due to the sample size limitations. For other dimensions, the difference in the safety dimension was close to the significance threshold ($p = 0.174$), but the low sample size limited the test's power.

Table 1. Expert evaluation of ChatGPT outputs in MDT consultations

Dimension	Expert 1 score (Median, IQR)	Expert 2 score (Median, IQR)	ICC (95% CI)	Wilcoxon p-value
Accuracy	4.0 (0.3)	3.5 (1.3)	0.63 (-0.20, 0.97) ^a	0.346
Completeness	4.0 (0.0)	4.5 (1.3)	-0.00 (-1.33, 0.90) ^a	0.773
Appropriateness	5.0 (0.3)	4.5 (1.0)	0.57 (-0.47, 0.96) ^a	1.000 ^b
Insight	5.0 (0.3)	4.0 (0.3)	0.25 (-0.36, 0.91) ^a	0.346
Safety	5.0 (0.0)	4.0 (0.5)	0.00 (-0.26, 0.76) ^a	0.174

a. ICC estimates with negative lower bounds suggest negligible agreement beyond chance in this small sample.

b. p-value of 1.000 reflects insufficient statistical power to detect differences due to tied ranks and zero differences in small samples (n=4).

ICC calculated using a two-way mixed-effects model for absolute agreement.

Wilcoxon signed-rank tests performed with normal approximation and continuity correction due to tied ranks and zero differences.

Discussion

Strengths and Weaknesses of ChatGPT in MDT Simulation

Early studies with small sample sizes in cancer and cardiovascular diseases have shown promising results regarding ChatGPT's performance in simulating MDT consultations[21,22]. To our knowledge, this is the first study to evaluate ChatGPT's ability to simulate MDT decision-making for complex VMS, such as SWS and KTS.

ChatGPT demonstrated strong diagnostic capabilities in VMS cases, aligning with traditional MDT conclusions in the primary diagnoses. Each diagnosis was supported by clinical signs, imaging results, and differential diagnoses, reducing the risk of misdiagnosis. For instance, in Case 2, ChatGPT accurately identified the classic triad of SWS — facial port-wine stain, seizures, and leptomeningeal angiomas. While ChatGPT's diagnostic reasoning was comprehensive, it occasionally

introduced uncertainties. In Case 4, for example, ChatGPT did not fully explore the clinical significance of the TYR gene mutation, whereas the traditional MDT provided a more detailed interpretation based on expert experience. This highlights a challenge for AI models: although AI can synthesize broad data, it may struggle with rare gene-phenotype associations. The static nature of ChatGPT's knowledge base (last updated in September 2021) further exacerbates this issue, limiting its ability to incorporate the most recent research. In contrast, MDT experts continuously update their knowledge and can apply the latest advancements in medical science. To address this, emerging technologies such as Grok2, which enable real-time data retrieval, could bridge this gap by enabling AI models to access up-to-date research dynamically[23].

Clinical Implications and Future Research

While ChatGPT provided detailed and structured recommendations, covering medications, surgical interventions, imaging, genetic counseling, and rehabilitation, some suggestions were impractical or excessive. For example, in Case 3, ChatGPT proposed lymph node transplantation, which, although theoretically valid, is not immediately feasible in many healthcare settings. In contrast, traditional MDTs are more likely to propose validated, actionable solutions that consider practical constraints. A notable strength of ChatGPT is its ability to provide transparent, logical explanations for its recommendations, which is valuable for clinical education and patient communication[24,25]. In contrast, traditional MDT decisions often rely on tacit knowledge and intuition, which are effective but difficult to articulate or replicate[26]. This underscores the potential of ChatGPT as an MDT support tool, especially in contexts requiring transparency and structured reasoning. Regarding patient safety, ChatGPT identified potential risks (e.g., glaucoma progression, seizure recurrence, vascular complications) and recommended preventive measures. It emphasized further evaluations (e.g., MRI, genetic testing) before confirming irreversible decisions. While this cautious approach is generally beneficial, it could lead to unnecessary costs or patient anxiety. Traditional MDTs, by contrast, balance risks and benefits based on experience, which may result in a more pragmatic approach. Developing standardized safety assessment tools could help address this discrepancy[27].

Limitations and Future Research Directions

This study assessed ChatGPT's performance in real clinical scenarios, providing a more accurate reflection of its potential clinical value[28]. However, the study has several limitations, including the small sample size, retrospective design, lack of external validation, and absence of patient outcome assessments. Furthermore, the model's capacity for long-term learning was not considered. Future research should involve larger sample sizes, multi-center data, and validation of patient outcomes to better evaluate ChatGPT's clinical value.

Conclusion

ChatGPT shows strong potential as an MDT support tool, particularly in diagnostic appropriateness, completeness of recommendations, and cross-disciplinary insights. However, due to limitations in handling complex and rare cases, ensuring feasibility, and updating its knowledge base, its role should remain complementary to expert decision-making. The development of real-time learning capabilities and hybrid augmented intelligence systems could bridge the gap between AI and human expertise, offering more robust solutions for complex disease management[29,30]. This study lays the foundation for future large-scale research.

Conflicts of interest

None of the authors have any relevant conflicts to disclose.

Multimedia Appendix 1

See supplement file 1

Abbreviations

SWS: Sturge-Weber syndrome

KTS: Klippel-Trenaunay syndrome

VMS: vascular malformation syndromes

MDT: multidisciplinary team

AI: artificial intelligence

IQR: interquartile range

ICC: Intraclass Correlation Coefficient

MRI: magnetic resonance imaging



References

1. Wassef M, Blei F, Adams D, et al. Vascular Anomalies Classification: Recommendations From the International Society for the Study of Vascular Anomalies. *Pediatrics*. 2015;136(1):e203-214. doi:10.1542/peds.2014-3673
2. Yeom S, Comi AM. Updates on Sturge-Weber Syndrome. *Stroke*. 2022;53(12):3769-3779. doi:10.1161/STROKEAHA.122.038585
3. Turner VL, Kearns C, Wattamwar K, McKenney AS. Klippel-Trenaunay Syndrome. *Radiogr Rev Publ Radiol Soc N Am Inc*. 2022;42(6):E167-E168. doi:10.1148/rg.220150
4. Mazhar A, Moosa S, Abbas A, Mallick Y, Samad L. A multi-disciplinary, multimodal approach for the management of vascular anomalies. *Pak J Med Sci*. 2020;36(1):S14-S19. doi:10.12669/pjms.36.ICON-Suppl.1710
5. De la Torre AJ, Luat AF, Juhász C, et al. A Multidisciplinary Consensus for Clinical Care and Research Needs for Sturge-Weber Syndrome. *Pediatr Neurol*. 2018;84:11-20. doi:10.1016/j.pediatrneurol.2018.04.005
6. Karmacharya RM, Vaidya S, Bhatt S, et al. Klippel-Trenaunay Syndrome: Case series from a university hospital of Nepal. *Ann Med Surg* 2012. 2022;78:103732. doi:10.1016/j.amsu.2022.103732
7. Brown GTF, Bekker HL, Young AL. Quality and efficacy of Multidisciplinary Team (MDT) quality assessment tools and discussion checklists: a systematic review. *BMC Cancer*. 2022;22(1):286. doi:10.1186/s12885-022-09369-8
8. Sires JD, Williams N, Huilgol SC, Harvey I, Antoniou G, Dawson J. An integrated multidisciplinary team approach to the management of vascular anomalies: challenges and benefits. *Pediatr Surg Int*. 2020;36(10):1149-1156. doi:10.1007/s00383-020-04727-8
9. Lidsky ME, Markovic JN, Miller MJ, Shortell CK. Analysis of the treatment of congenital vascular malformations using a multidisciplinary approach. *J Vasc Surg*. 2012;56(5):1355-1362; discussion 1362. doi:10.1016/j.jvs.2012.04.066
10. Lammila-Escalera E, Greenfield G, Barber S, Nicholls D, Majeed A, Hayhoe BWJ. A Systematic Review of Interventions that Use Multidisciplinary Team Meetings to Manage Multimorbidity in Primary Care. *Int J Integr Care*. 22(4):6. doi:10.5334/ijic.6473
11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
12. Zhang C, Xu J, Tang R, et al. Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment. *J Hematol Oncol J Hematol Oncol*. 2023;16(1):114. doi:10.1186/s13045-023-01514-5
13. Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA Cancer J Clin*. 2019;69(2):127-157. doi:10.3322/caac.21552
14. Introducing OpenAI o1. September 12, 2024. Accessed March 1, 2025. <https://openai.com/o1/>

15. Kayaalp M. Patient Privacy in the Era of Big Data. *Balk Med J.* 2018;35(1):8-17. doi:10.4274/balkanmedj.2017.0966
16. Ho CN, Tian T, Ayers AT, et al. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. *BMC Med Inform Decis Mak.* 2024;24(1):357. doi:10.1186/s12911-024-02757-z
17. Sullivan GM, Artino AR. Analyzing and interpreting data from likert-type scales. *J Grad Med Educ.* 2013;5(4):541-542. doi:10.4300/JGME-5-4-18
18. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420-428. doi:10.1037//0033-2909.86.2.420
19. Whitley E, Ball J. Statistics review 6: Nonparametric methods. *Crit Care Lond Engl.* 2002;6(6):509-513. doi:10.1186/cc1820
20. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess.* 1994;6(4):284-290. doi:10.1037/1040-3590.6.4.284
21. Sudri K, Motro-Feingold I, Ramon-Gonen R, et al. Enhancing Coronary Revascularization Decisions: The Promising Role of Large Language Models as a Decision-Support Tool for Multidisciplinary Heart Team. *Circ Cardiovasc Interv.* 2024;17(11):e014201. doi:10.1161/CIRCINTERVENTIONS.124.014201
22. Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *Npj Breast Cancer.* 2023;9(1):44. doi:10.1038/s41523-023-00557-8
23. Grok-2 Beta Release. Accessed February 27, 2025. <https://x.ai/blog/grok-2>
24. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, et al. AI in Surgical Curriculum Design and Unintended Outcomes for Technical Competencies in Simulation Training. *JAMA Netw Open.* 2023;6(9):e2334658. doi:10.1001/jamanetworkopen.2023.34658
25. Finkelstein J, Gabriel A, Schmer S, Truong TT, Dunn A. Identifying Facilitators and Barriers to Implementation of AI-Assisted Clinical Decision Support in an Electronic Health Record System. *J Med Syst.* 2024;48(1):89. doi:10.1007/s10916-024-02104-9
26. Greenhalgh J, Flynn R, Long AF, Tyson S. Tacit and encoded knowledge in the use of standardised outcome measures in multidisciplinary team decision making: a case study of in-patient neurorehabilitation. *Soc Sci Med* 1982. 2008;67(1):183-194. doi:10.1016/j.socscimed.2008.03.006
27. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. *JAMA Netw Open.* 2023;6(9):e2335377. doi:10.1001/jamanetworkopen.2023.35377
28. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology.* 2018;286(3):800-809. doi:10.1148/radiol.2017171920
29. Burton JW, Lopez-Lopez E, Hechtlinger S, et al. How large language models can reshape

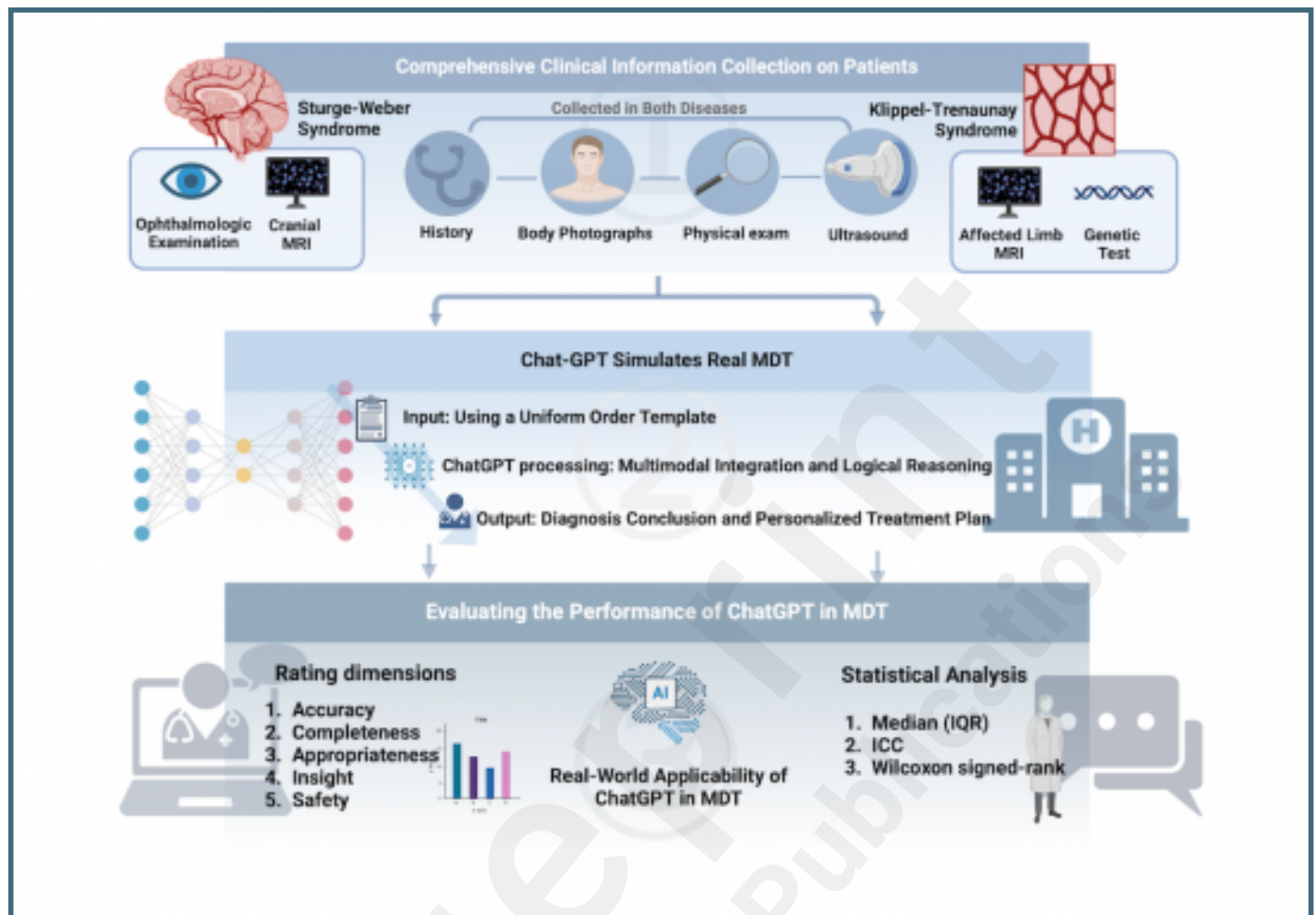
- collective intelligence. *Nat Hum Behav.* 2024;8(9):1643-1655. doi:10.1038/s41562-024-01959-9
30. Hirosawa T, Suzuki T, Shiraishi T, et al. Adapting Artificial Intelligence Concepts to Enhance Clinical Decision-Making: A Hybrid Intelligence Framework. *Int J Gen Med.* 2024;17:5417-5422. doi:10.2147/IJGM.S497753



Supplementary Files

Figures

Research framework for evaluating ChatGPT in multidisciplinary team consultations.



Multimedia Appendixes

Predefined instructions.

URL: <http://asset.jmir.pub/assets/7e7083a27de233f971f304196848ca3a.docx>

