CAS 2026

Critical Care Medicine Abstracts

Contents

Accuracy of large language models in performing critical care and perioperative clinical calculations 3

Are AI-based prediction tools for clinical deterioration ready for clinical use? A systematic review of implementation readiness in hospitalized adults. 6

Stroke location and cardiovascular-autonomic disturbances: a scoping review of cortical autonomic network involvement after ischemic stroke 8

Accuracy of large language models in performing critical care and perioperative clinical calculations

Submission ID

229

AUTHORS

Rozman, Robert C.;[1] Barr, Austin A.;[2] Doyle, Marcus A.[2]

[1]Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada; [2]Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

INTRODUCTION

Clinical calculators are evidence-based tools that integrate patient data into scoring systems to support clinical decision-making. These calculators are particularly useful in critical care and perioperative medicine for risk stratification, prognosis, and treatment selection [1]. Although online platforms such as MDCalc broaden access to clinical calculators, manual data entry can be time-consuming and prone to transcription errors, especially when numerous input parameters are required [2]. As a result, large language models (LLMs) have been proposed as tools to automate clinical calculations [3]; however, accuracy and comparative performance between LLMs have not been well-characterized. In this study, we evaluate three widely used LLMs: ChatGPT, OpenEvidence, and Gemini, across 30 clinical calculators relevant to critical care and perioperative medicine.

METHODS

Thirty validated and commonly used clinical calculators in critical care and perioperative medicine were selected to represent a wide array of computational structures. These included: classification-based (e.g., ASA Physical Status), simple additive (e.g., STOP-BANG Score), weighted additive (e.g., APACHE II Score), and equation-based (e.g., Shock Index) calculators. Three LLMs were studied: ChatGPT 5.2, OpenEvidence, and Gemini 3. Standardized prompts were developed for each calculator that included explicit instructions and relevant patient data required to perform calculations. Five different trials were performed per calculator per LLM, and the same prompts were used for all LLMs. LLM-generated outputs were recorded and compared against reference values calculated using MDCalc, with accuracy defined as the proportion of correct outputs across the five trials. Fast-response modes were used for ChatGPT 5.2 and Gemini 3 (not available for OpenEvidence) to reflect the need for timely responses in clinical settings. Each prompt was submitted in a separate chat with memory disabled. Overall performance for each LLM was calculated by aggregating the accuracy across all 30 calculators. Pairwise comparisons

between each LLM's overall performance were conducted using a two-sided Wilcoxon rank-sum test.

RESULTS

Overall, ChatGPT 5.2 outputted correct calculations in 143/150 trials (95.3%), compared to OpenEvidence (108/150; 72.0%) and Gemini 3 (100/150; 66.7%). ChatGPT 5.2 significantly outperformed both OpenEvidence (p = <0.001) and Gemini 3 (p = <0.001), while no significant difference was observed between OpenEvidence and Gemini 3 (p = 0.64). The calculators that all three LLMs achieved perfect accuracy on included the: Apfel Score for PONV, and CHA_2DS_2-VASc Score, Creatinine Clearance (Cockcroft-Gault Equation), ROX Index, Shock Index, and qSOFA Score. For ChatGPT 5.2, its lowest accuracy (60%) was observed on the 2021 CKD-EPI Creatinine-Cystatin C eGFR calculator. OpenEvidence demonstrated 0% accuracy on the Estimated Blood Volume (Nadler's Equation) and NEWS2. Gemini 3 achieved 0% accuracy on the ARISCAT Score, Caprini Score for VTE, NEWS2, and SOFA Score. Notably, three of ChatGPT's seven total errors occurred in classification-based calculators. Across models, equation-based calculators demonstrated the highest accuracy.

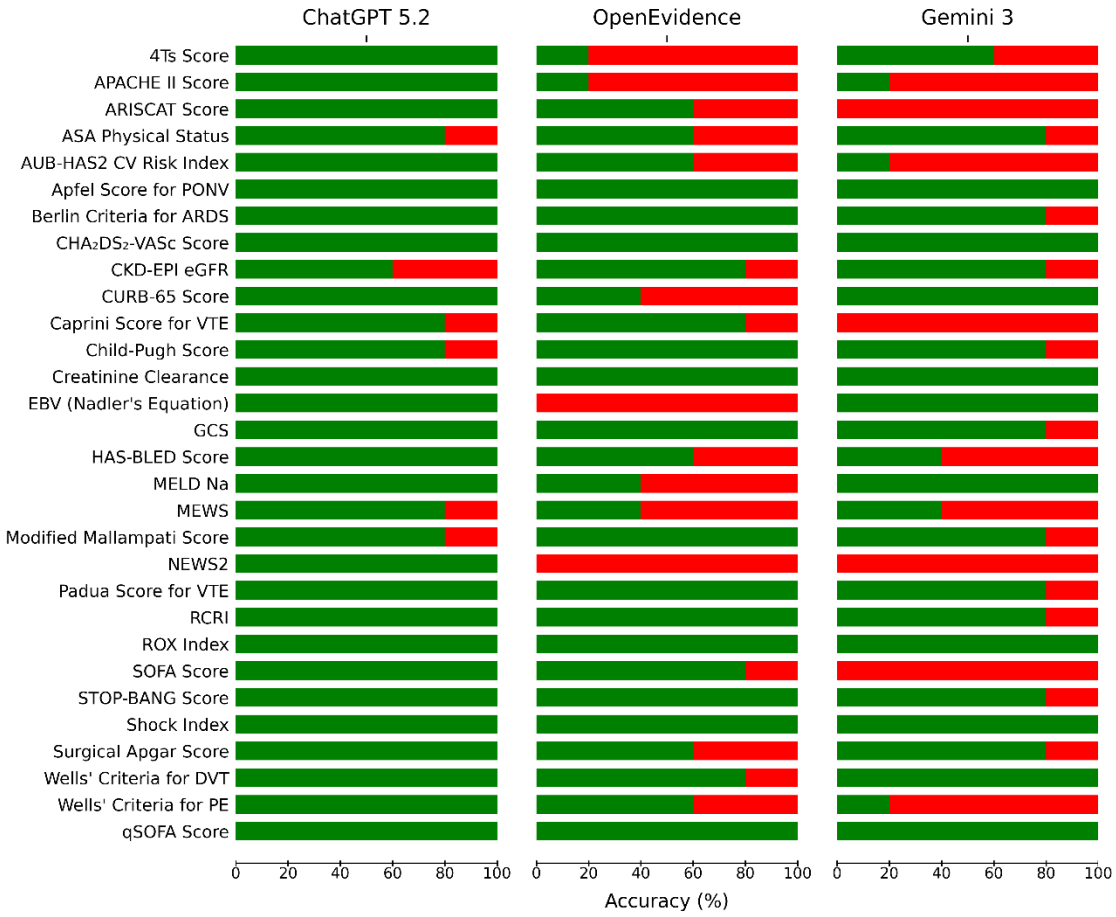
DISCUSSION

Here we assessed a novel approach to conducting clinical calculations in critical care and perioperative medicine. ChatGPT 5.2 demonstrated promising accuracy across a diverse set of clinical calculators; however, performance was not reliably accurate across all three LLMs. Despite the broad clinical use of LLMs [3], this presents an important limitation to their utility. Further development is required to improve the accuracy of these LLMs in clinical calculations before application to patient care. Future work should also assess additional LLMs and clinical calculators relevant to other medical disciplines.

REFERENCES

1. Smilowitz, N. R., & Berger, J. S. (2020). Perioperative Cardiovascular Risk Assessment and Management for Noncardiac Surgery: A Review. *JAMA, 324*(3), 279–290. https://doi.org/10.1001/jama.2020.7840
2. Mickelsson, M., Ekblom, K., Stefansson, K., Själander, A., Näslund, U., & Hultdin, J. (2025). Exploring the extent of post-analytical errors, with a focus on transcription errors - an intervention within the VIPVIZA study. *Clinical chemistry and laboratory medicine, 63*(8), 1555–1560. https://doi.org/10.1515/cclm-2025-0009
3. Pressman, S. M., Borna, S., Gomez-Cabello, C. A., Haider, S. A., Haider, C. R., & Forte, A. J. (2024). Clinical and surgical applications of large language models: a systematic review. *Journal of Clinical Medicine*, *13*(11), 3041. https://doi.org/10.3390/jcm13113041

Figure 1. Accuracy of ChatGPT 5.2, OpenEvidence, and Gemini 3 across 30 critical care and perioperative clinical calculators. Horizontal bars represent the proportion of correct calculations (green) and incorrect calculations (red) across five trials per calculator



Are AI-based prediction tools for clinical deterioration ready for clinical use? A systematic review of implementation readiness in hospitalized adults.

Submission ID

187

AUTHORS

Zhang, Eric;[1] Daniel Kwan;[1] Jill Shah;[1] Sara Pollanen;[1] Kalia Kamini;[2]

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, Canada; [2]Department of Clinical Systems Informatics, Trillium Health Partners, Mississauga, Canada

INTRODUCTION

Artificial intelligence (AI)-based tools have been increasingly used to predict clinical deterioration, defined as an acute worsening of a hospitalized patient's physiological status that may lead to ICU transfer, cardiac arrest, or death.[1] Early identification of deterioration is closely linked to perioperative risk stratification, rapid response activation, and ICU triage, domains directly relevant to anesthesiology.[2] While many AI models report strong discriminative performance for outcomes such as ICU transfer, cardiac arrest, and in-hospital mortality, performance alone does not ensure clinical usefulness. Less is known about the readiness of these tools for real-world implementation, including prospective validation, integration into clinical workflows, and reproducibility.

METHODS

We conducted a systematic review registered in PROSPERO (CRD420251091119). MEDLINE, Embase, Scopus, and IEEE Xplore were searched for studies evaluating artificial intelligence or machine learning-based models designed to predict clinical deterioration in adult inpatients, defined by outcomes including ICU transfer, cardiac arrest, or in-hospital mortality. Two reviewers independently screened titles and abstracts, assessed full-text eligibility, and extracted data using a standardized form. Implementation readiness was operationalized using predefined, study-level criteria, including validation strategy (internal vs external), prospective or real-time evaluation, deployment within live clinical environments, integration into electronic health records, presence of clinician-facing outputs, prediction horizon, and reporting of operational and workflow considerations. Disagreements were resolved by consensus. Methodological quality and reproducibility were assessed secondarily using selected domains of the APPRAISE-AI framework, focusing on data quality, robustness of results, reporting of model inputs and thresholds, and reproducibility.[3]

RESULTS

Thirty-seven studies met inclusion criteria, representing over 3.2 million inpatient encounters across North America, Europe, and East Asia. Most studies were conducted on general medical or mixed medical-surgical wards and were retrospective in design. Only 8/37 (21.6%) studies evaluated models in prospective or real-time clinical settings, and 10/37 (27.0%) reported prospective validation, while 22/37 (59.5%) relied solely on retrospective external validation. Prediction horizons varied widely, from under 6 hours to more than 24 hours prior to deterioration, with only 5/37 (13.5%) aligned with real-time clinical decision-making. Commonly evaluated tools included eCART, CHARTwatch, and the Epic Deterioration Index. Although reported discrimination was high (AUROC range 0.72-0.97), few studies described EHR integration, clinician-facing alerting strategies, or mitigation of alert fatigue. Secondary APPRAISE-AI assessment identified strengths in clinical relevance and data quality but frequent limitations in methodological robustness, reporting of model inputs and thresholds, and reproducibility that may limit real-world deployability.

DISCUSSION

Despite strong predictive performance, most AI-based deterioration tools lack key features required for real-world hospital deployment, including prospective validation, workflow integration, and reproducibility. Secondary APPRAISE-AI assessment suggests that gaps in reporting of model inputs, thresholds, and robustness further limit scalability. These findings indicate that many tools remain research instruments rather than deployable clinical systems. Evaluating implementation readiness alongside traditional performance metrics is essential to guide safe, effective, and scalable integration of AI into acute care practice.

REFERENCES

1. Churpek MM, Yuen TC, Edelson DP. Predicting clinical deterioration in the hospital: the impact of outcome selection. Resuscitation. 2013 May;84(5):564–8. DOI: 10.1016/j.resuscitation.2012.09.024
2. Bakkes THGF, Mestrom EHJ, Ourahou N, Kaymak U, de Andrade Serra PJ, Mischi M, et al. Predictive modeling of perioperative patient deterioration: combining unanticipated ICU admissions and mortality for improved risk prediction. Perioper Med (Lond). 2024 July 3;13:66. DOI: https://doi.org/10.1186/s13741-024-00420-9
3. Kwong JCC, Khondker A, Lajkosz K, McDermott MBA, Frigola XB, McCradden MD, et al. APPRAISE-AI Tool for Quantitative Evaluation of AI Studies for Clinical Decision Support. JAMA Netw Open. 2023 Sept 25;6(9):e2335377. DOI: https://doi.org/10.1001/jamanetworkopen.2023.35377

Stroke location and cardiovascular-autonomic disturbances: a scoping review of cortical autonomic network involvement after ischemic stroke

Submission ID

157

AUTHORS

Elganga, Mouad;[1] Voznyy, Vitaliy;[1] Abu Al-Burak, Salem;[2] Elsherbini, Adham;[1] Chowdhury, Tumul[3]

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; [2]Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada; [3]Department of Anesthesiology and Perioperative Medicine, Marnix E. Heersink School of Medicine, The University of Alabama at Birmingham, Alabama, USA

INTRODUCTION

Acute ischemic stroke is commonly associated with secondary cardiovascular and autonomic abnormalities that can affect early morbidity, neurological recovery, and long-term outcomes.[1] Injury to key nodes within the cortical autonomic network represents a proposed pathophysiological link between focal cerebral ischemia and subsequent dysregulation of cardiac and autonomic function.[2] Prior investigations suggest that infarct topography and hemispheric dominance may influence autonomic balance and susceptibility to adverse cardiac events; however, findings across studies have been inconsistent.[3,4] Variability in study design, autonomic outcome measures, and imaging-based localization has limited synthesis of this literature. As a result, the relationship between stroke location, laterality, and autonomic-cardiac consequences has yet to be systematically characterized, underscoring the need for a comprehensive mapping of existing evidence.

METHODS

A comprehensive literature search was conducted in MEDLINE, EMBASE, and the Cochrane Central Register of Controlled Trials from database inception through December 9, 2025. Studies were eligible if they enrolled adult patients with radiologically confirmed ischemic stroke and examined cardiovascular or autonomic outcomes stratified by infarct location or laterality. Title and abstract screening, followed by full-text screening, were conducted in duplicate by two independent reviewers. Key outcomes extracted from included studies were study design, methods of lesion localization and classification, and reported autonomic and cardiovascular outcomes. Given anticipated heterogeneity in study

populations and outcome measures, findings were synthesized using a narrative approach rather than quantitative pooling.

RESULTS

A total of 34 studies were included, comprising 6,758 patients with ischemic stroke from prospective and observational cohorts. Stroke location across studies was classified by hemispheric laterality, insular involvement, lobar or cortical–subcortical regions, and voxel-based lesion–symptom mapping, with several studies using multiple approaches. Outcomes included autonomic measures, ECG abnormalities, blood pressure/catecholamine responses, cardiac biomarkers/echocardiography, and clinical cardiac events. Of the 23 studies examining insular involvement, 18 reported associations with autonomic/cardiac abnormalities, while 5 did not. Right insular lesions were more often linked to sympathetic predominance and arrhythmias, whereas left insular lesions were more often associated with myocardial injury biomarkers, reduced ejection fraction, or longer-term cardiac events. Voxel-based studies localized effects to distributed right-hemisphere networks involving the dorsal anterior insula, fronto-parietal operculum, basal ganglia, thalamus, and amygdala. Seven studies examined posterior circulation or brainstem strokes, with lateral medullary cohorts showing marked autonomic dysfunction, particularly in ventral and right-sided lesions.

DISCUSSION

Existing evidence indicates that the anatomical distribution of ischemic stroke, particularly involvement of the right hemisphere and regions within the cortical autonomic network, is associated with clinically relevant cardiovascular and autonomic abnormalities. Interpretation of these associations is constrained by considerable variability in lesion localization approaches, outcome definitions, and methods used to assess autonomic function across studies. Future investigations would benefit from anatomically precise and standardized frameworks that combine advanced neuroimaging techniques with detailed physiological and cardiac monitoring. Such approaches are essential to better define brain heart interactions and to inform risk stratification and targeted management strategies following ischemic stroke.

REFERENCES

1. Kumar S, Chou SH y, Smith CJ, Nallaparaju A, Laurido-Soto OJ, Leonard AD, et al. Addressing systemic complications of acute stroke: A scientific statement from the American Heart Association. Stroke [Internet]. 2024 Dec 5;56(1):e15–29. Available from: https://www.ahajournals.org/doi/full/10.1161/STR.0000000000000477?rfr_dat=cr_pub++0pubmed&url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org
2. Benarroch EE. The central autonomic network: functional organization, dysfunction, and perspective. Mayo Clinic Proceedings [Internet]. 1993 Oct 1;68(10):988–1001. Available from: https://pubmed.ncbi.nlm.nih.gov/8412366/

3. Sposato LA, Hilz MJ, Aspberg S, Murthy SB, Bahit MC, Hsieh CY, et al. Post-Stroke cardiovascular complications and neurogenic cardiac injury. Journal of the American College of Cardiology [Internet]. 2020 Nov 30;76(23):2768–85. Available from: https://www.jacc.org/doi/10.1016/j.jacc.2020.10.009
4. Meyer S, Strittmatter M, Fischer C, Georg T, Schmitz B. Lateralization in autonomic dysfunction in ischemic stroke involving the insular cortex. Neuroreport [Internet]. 2004 Jan 22;15(2):357–61. Available from: https://pubmed.ncbi.nlm.nih.gov/15076768/

Table Study characteristics of included investigations evaluating stroke location and cardiovascular or autonomic outcomes

Study ID/Reference	Year of Publication	Study Design	Country of Study	N patients with stroke	Age (mean ± SD)	Male n (%)
Abboud 2006	2006	PC	France	493	N/A	N/A
Algra 2003	2003	PC	Netherlands	1483	66.6 ± 14.1	1065 (71.8)
Ay 2006	2006	PC	US	100	74.8 ± 12.3	50 (50)
Barron 1994	1994	PC	Isreal	40	68.8	21 (52.5)
Chen 2013	2013	PC	Taiwan	75	59.6 ± 11.7	46 (61.3)
Christensen 2005	2005	PC	Denmark	179	73 ± 12.6	N/A
Colivicchi 2004	2024	PC	Italy	103	N/A	N/A
Constantinescu 2019	2019	PC	Romania	71	N/A	34 (47.9)
Constantinescu 2016	2016	PC	Romania	40	64.3 ± 8.9	N/A
Constantinescu 2018	2018	PC	Romania	30	59.6	15 (50)
DeVos 2017	2017	PC	Belgium	135	N/A	70 (51.9)
Diserens 2006	2006	PC	Switzerland	100	70.8 ± 15.2	51 (51)
Dutsch 2007	2007	PC	Germany	15	N/A	13 (46)
Fink 2005	2005	Case Series	New Zealand	150	71	N/A
Hong 2013	2013	PC	South Korea	25	50.5 ± 8.3	20 (80.0)
Kitamura 2018	2018	PC	Japan	90	70 ± 46.7	45 (50.0)
Krause 2017	2017	RC	Germany	299	80	112 (49)
Laowattana 2006	2005	PC	US	116	65.9	56
Meyer 2004	2004	PC	Germany	29	N/A	N/A
Min 2022	2022	RC	US	415	72.3	219 (52.8)
Naver 1996	1996	PC	Sweden	23	59 ± 13	N/A
Nayani 2016	2016	PC	India	101	63	72 (72)
Raphaely-Beer 2020	2020	PC	Isreal	25	N/A	15 (60)
Rincon 2008	2008	PC	US	655	69.7 ± 12.7	292 (44.6)
Sahin 2025	2025	PC	Turkey	60	N/A	N/A
Sander 1995	1995	PC	Germany	35	66	22 (62.9)
Seifert 2015	2015	PC	Germany	150	N/A	85 (56.7)
Strittmatter 2003	2003	PC	Germany	39	61.3 ± 3.7	N/A
Sykora 2009	2009	PC	Slovakia	52	N/A	N/A
Tatschl 2006	2006	PC	Austria	109	N/A	N/A
Tokgozoglu 1999	1999	RC	Turkey	62	N/A	N/A
Vassilopoulou 2020	2020	PC	Greece	1212	71.4 ± 12.5	653 (53.9)
Wang 2022	2022	PC	Germany	61	66.5 ± 9.7	42 (68.9)
Zhao 2020	2020	PC	China	186	59.7 ± 9.7	150 (80.6)

prospective cohort (PC); retrospective cohort (RC); standard deviation (SD); not available (N/A).