ISSN 2524 – 0684 **e-ISSN 2524 – 0692**

UDC 616-083.98:004.8
IRSTI 76.29.44
DOI: 10.53065/kaznmu.2026.77.2.005

Received: 21.04.2026
Accepted for publication: 11.06.2026

DIAGNOSTIC ACCURACY AND THERAPEUTIC RATIONALITY IN ACID-BASE IMBALANCES: A COMPARATIVE ANALYSIS BETWEEN PHYSICIANS AND ARTIFICIAL INTELLIGENCE

S.T. AMRIN [1], A.S. SATBAYEVA [2], A.A. ABDUSSEMYATOVA [3], Y.A. DUISEN [4]

[1] LLP Kazakhstan Medical University "KSPH", Almaty, Kazakhstan
[2] S.D. Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan
[3] Scientific Research Institute of Cardiology and Internal Diseases, Almaty, Kazakhstan
[4] A.N. Syzganov National Scientific Center of Surgery, Almaty, Kazakhstan

Abstract

Introduction. Acid-base balance disorders are critical conditions in intensive care units requiring rapid and accurate management. The study explores the potential of large language models to serve as accessible clinical decision support systems to reduce iatrogenic errors.

Aim. To compare the accuracy of diagnosing acid-base balance disorders and the rationality of therapeutic recommendations proposed by artificial intelligence based on ChatGPT and by intensive care physicians.

Materials and Methods. Study design: retrospective, single-center, comparative study. The analysis included 302 clinical and laboratory cases of patients treated in an intensive care unit between 2024 and 2025. Using prompt engineering techniques, an adapted ChatGPT model named "ReanimatorKZ" was developed. A comparative expert evaluation was conducted to assess the conclusions of ChatGPT and intensive care physicians regarding acid–base disorders. Statistical analysis was performed using StatTech v.4.12.7 and SPSS Statistics 27.0.1.

Results. In the group of doctors, diagnostic accuracy was 71.2% correct, 24.8% partially correct, and 4.0% incorrect conclusions. The AI demonstrated a lower rate of completely incorrect diagnostic conclusions (2.0%), while completely correct diagnoses accounted for 64.9%. The physicians' treatment strategies were completely correct in 60.9% of cases, whereas the AI's recommendations were completely correct in 89.7% of cases, with no completely incorrect therapeutic recommendations classified for the AI. Statistically significant differences were confirmed using paired tests ($p < 0.05$).

Conclusion. An adapted version of ChatGPT demonstrated a high level of diagnostic accuracy in identifying acid–base disorders, comparable to that of intensive care physicians, and superior accuracy in formulating therapeutic recommendations for these conditions. Our study supports the potential for developing effective and readily scalable clinical decision support systems based on widely available artificial intelligence models. However, additional prospective validation is required before such systems can be implemented in routine clinical practice.

Keywords: artificial intelligence, intensive care, clinical decision support systems, ChatGPT, diagnostic accuracy.

Introduction. Artificial Intelligence (AI) is increasingly playing a pivotal role in modern medicine, particularly in intensive care, where the volume and complexity of data necessitate new approaches to analysis and clinical decision-making [1,2,3]. In intensive care units (ICUs), AI has the potential to significantly enhance the quality and safety of medical care through the following avenues:

Early diagnosis and prediction of clinical deterioration: Specialized AI models demonstrate high accuracy in predicting mortality, sepsis, and other adverse outcomes in critically ill patients [4]. The implementation of such systems enables the timely identification of patients at risk of deterioration and facilitates a prompt clinical response [5, 6].

Optimization of clinical processes: AI automates the collection, analysis, and interpretation of data (including vital sign monitoring, laboratory results, and medical imaging), thereby reducing the cognitive load on physicians and standardizing decision-making. There are "Closed-loop" systems, which independently adjust drug dosages and mechanical ventilation parameters, ensure precise and safe patient management during anesthesia. This reduces the incidence of complications (such as hypotension) and accelerates recovery [7, 8]. AI allows for the individualization of treatment, including precise dosage titration, complication forecasting, and risk stratification. This is critically important in cases of multi-organ failure and complex comorbid backgrounds [9].

Despite the rapid surge in publications, most research in the field of AI for intensive care remains in its early stages. Only a small fraction of algorithms has undergone external validation or integration into real-world clinical practice. Key barriers include insufficient data quality, the risk of bias, limited model transparency ("black box" effect), difficulties in workflow integration, and a lack of trust among clinicians. In the United States, only a few AI tools have received FDA clearance for ICU applications, yet their widespread implementation remains limited [10,11].

Successful integration of AI into intensive care requires multidisciplinary collaboration, standardization of reporting, transparent and interpretable models, and continuous post-implementation monitoring of efficacy and safety. Further advancement in AI is expected to transform management approaches for critically ill patients [12].

Acid-base balance (ABB) disturbances are among the most common conditions in patients in critical or unstable states. Since systemic metabolism depends on pH levels, appropriate therapeutic correction should be initiated as early as possible to improve patient outcomes [13,14].

According to major international cohort studies, the prevalence of acidosis (pH < 7.35) in patients within the first 24 hours of ICU admission is approximately 57.8%. Among these cases, metabolic acidosis accounted for 42.9%, mixed acidosis for 30.3%, and respiratory acidosis for 25.9% [15,16,17]. In specific studies of mechanically ventilated patients at the time of ICU admission, acidemia was identified in 32%, alkalemia in 17%, and normal pH in 51% [18].

Direct interpretation of ABB analysis requires the intensivist to possess not only knowledge of reference values but also an understanding of the pathophysiological causes and the most effective corrective pathways.

Modern AI technologies have reached a level that allows for both the diagnosis of disturbances and the recommendation of corrective algorithms. Consequently, the objective of this study is to compare the diagnostic accuracy and efficacy of therapeutic recommendations between AI and intensive care physicians.

Research question (PICO):

P (Population): Patients in ICU.

I (Intervention): Use of the adapted AI language model "Reanimator KZ".

C (Comparison): Independent clinical decisions by physicians.

O (Outcome): Improvement in the accuracy of ABB diagnosis and the correctness of therapeutic recommendations.

Materials and methods

Ethical considerations

ISSN 2524 – 0684 **e-ISSN 2524 – 0692**

The study was conducted in accordance with the ethical principles of the Declaration of Helsinki of the World Medical Association. Approval was obtained from the Local Ethics Committee of the Kazakhstan Medical University "KSPH" (Protocol No. IRB-433-2025 of November 25, 2025). Due to the retrospective design of the study and the use of existing medical records, informed consent from patients was not required. All data were thoroughly anonymized and presented in an aggregated format, ensuring full confidentiality and non-disclosure of personal medical information in compliance with the legislation of the Republic of Kazakhstan.

Study Design

A retrospective, single-center, observational comparative study aimed at evaluating the diagnostic accuracy and completeness of therapeutic recommendations for ABB disorders in ICU patients. The study compared the diagnostic and treatment decisions made by intensive care physicians with those made by an AI model.

Data collection

The collection of clinical and laboratory data was carried out from the medical records of patients who received treatment in the anesthesiology and intensive care department of the Talgar Central District Hospital in the period from 2024 to 2025.

Inclusion Criteria: Clinical cases of patients treated in the ICU, for whom ABB analyses were recorded during the course of treatment. No restrictions were placed on sex, age, or diagnosis.

Exclusion Criteria: Incomplete clinical or laboratory datasets insufficient for accurate ABB interpretation and formulation of recommendations; Cases where blood gas parameters were obtained with significant technical errors or without a specified sample source (arterial or venous blood).

For each clinical case, systematic collection of the following parameters were performed:

- Demographic and Clinical Data: Age, sex, body weight, primary and secondary diagnoses, and complications.
- Laboratory Parameters and ABB: Blood sample source (arterial/venous), a full spectrum of gas exchange parameters (pH, pCO_2, pO_2, HCO_3^-, BE), electrolytes (Na^+, K^+, Cl^-, Ca^{2+}), lactate, glucose, hematocrit, and hemoglobin levels.
- Vital Signs: Oxygenation level (SpO_2), hemodynamic status (normotension, hypertension, shock), renal function (creatinine, urine output), and body temperature.
- Clinical Status: Respiratory support (ranging from spontaneous breathing to mechanical ventilation) and neurological status (Glasgow Coma Scale score and presence of psychomotor agitation).

Preparation of the AI model based on ChatGPT

ChatGPT was selected as the experimental model for the application of AI in medical practice, specifically for the interpretation of ABB analysis and the subsequent provision of clinical recommendations. The choice of this software over other alternatives was based on its widespread popularity and superior reasoning capabilities at the time of the study. For the purpose of this research, a customized version of ChatGPT named "Reanimator KZ" was developed.

To configure the model, a detailed system prompt was utilized. The original interaction with the AI model, including the system prompt and clinical queries, was conducted in Russian to ensure precise alignment with local clinical guidelines and terminology. The English translation of the system prompt used to configure the "Reanimator KZ" model is as follows:

"ROLE AND PURPOSE" you are an expert in anesthesiology, resuscitation, and intensive care. Your primary goal is to assist the anesthesiologist-intensivist in making the most effective, evidence-based, and safe clinical decisions for critically ill patients.

ISSN 2524 – 0684 **e-ISSN 2524 – 0692**

CORE OPERATING PRINCIPLES. Evidence-Based Medicine: Use modern international clinical guidelines and consensus statements (ESICM, SCCM, Surviving Sepsis Campaign, ARDSNet, Neurocritical Care Society, ERC, etc.). Avoid outdated, empirical, and unproven approaches. If a recommendation is based on a weak level of evidence, explicitly state this. Clinical Applicability: Provide specific answers: what to do, what to change, within what limits, and what the risks are. Patient Safety: Always specify exact drug dosages, mechanical ventilation parameters, and target parameters for blood pressure, gas exchange, urine output, etc. Warn about potential complications and iatrogenic risks. When making any calculations, if data is missing, always clarify: height, weight, age, gender, and comorbidities. If there is insufficient data for a safe response, state this directly and request the missing information. Honesty and Limitations: If evidence-based information is lacking or contradictory, report it. Do not present hypotheses, personal opinions, or traditional "schools of thought" as clinical guidelines. Do not fabricate data. If information is insufficient, you are required to directly state that there is not enough data for a safe decision and list exactly what data is needed. Strictly prohibited: Giving advice outside the framework of evidence-based medicine; ignoring the individual characteristics of the patient; using vague or evasive phrasing. Communication Style: Professional.

To ensure the localization of solutions for the Republic of Kazakhstan, the following regulatory acts and specialized medical literature were integrated into the model's knowledge base:

- Regulatory Framework of the Republic of Kazakhstan: The Code "On Public Health and the Healthcare System", the Standards for the Provision of Anesthesiology and Intensive Care (Order No. 78), and the Transfusiology Rules (Order No. 140).
- Clinical Guidelines of the Ministry of Health of the Republic of Kazakhstan (cardiology, pulmonology, neurology, etc.).
- Fundamental guides on fluid and electrolyte imbalances, principles of mechanical ventilation, and neuro-intensive care.

Description of Interaction with the AI Model

Based on the collected clinical and laboratory data, a unified text query was generated for each case and uploaded into the "Reanimator KZ" model.

The prompt included patient demographics, body weight, primary diagnosis, blood sample source, acid–base balance parameters (pH, pCO_2, pO_2, HCO_3^-, and base excess), electrolyte and metabolic variables (sodium, potassium, chloride, ionized calcium, lactate, and glucose), hematological parameters (hematocrit and hemoglobin), oxygen saturation, hemodynamic status, creatinine level, urine output, body temperature, respiratory status, Glasgow Coma Scale score, and the presence of psychomotor agitation. Based on these data, the model was requested to identify the type of acid–base balance disorder, provide a brief rationale for its classification, and suggest therapeutic management, including specific drug dosages when indicated.

Identical queries (without AI-generated conclusions) were presented to an independent group of physicians. To ensure the homogeneity of the control group and to exclude the influence of insufficient clinical experience, residents and physicians from other specialties were excluded from the analysis. The final control group consisted exclusively of 20 board-certified intensivists.

The final analysis included 302 independent, unique clinical cases. The cases were evenly distributed among the participants: on average, each physician independently evaluated approximately 15 unique clinical cases. This balanced distribution and independent evaluation scheme reduced the risk of single-evaluator bias and clustering effects.

Expert evaluation

ISSN 2524 – 0684 **e-ISSN 2524 – 0692**

A board-certified anesthesiologist and intensivist with 20 years of experience, serving as the head of the ICU, performed the expert evaluation of the AI and doctors' diagnoses and therapies. Initially, a blinded assessment method was intended in order to prevent the expert from identifying whether the evaluated responses belonged to the physicians or the AI. However, during the study, the distinction became apparent because all responses were presented without modification: AI-generated answers were consistently more detailed, extensive, and comprehensive, whereas physicians' responses were generally brief and concise. The accuracy of ABB disorder identification and the validity of the therapeutic recommendations (for both the physicians and the AI) were evaluated using a three-level scale:

- Correct: Full compliance with the reference standard or complete agreement between conclusions.
- Partially Correct: partial compliance with the reference standard (e.g., correct identification of the primary disorder, but an error in the wording; correct therapeutic approach, but requiring clarification).
- Incorrect: Complete non-compliance with the reference standard or a gross clinical error.

Reference criteria for ABB disorders were strictly regulated:

- Physiological norm: pH 7.35–7.45
- Subcompensated acidosis: pH 7.30–7.35
- Decompensated acidosis: pH < 7.30
- Subcompensated alkalosis: pH 7.45–7.50
- Decompensated alkalosis: pH > 7.50

Differential diagnosis of respiratory and metabolic components, as well as the interpretation of hemoglobin, electrolytes, lactate, and glucose levels, were carried out in strict accordance with generally accepted physiological norms.

Statistical analysis

Statistical analysis was performed using StatTech v. 4.12.7. Quantitative indicators were assessed for normal distribution using the Kolmogorov-Smirnov test. In the absence of a normal distribution, quantitative data were described using the median (Me) and the lower and upper quartiles (Q1–Q3). Categorical data were described using absolute values, percentages, and 95% confidence intervals (95% CI). To compare categorical variables between dependent samples (evaluating the accuracy of AI and physician responses on the identical clinical cases), tests for paired nominal data were applied, specifically the McNemar-Bowker test and the Marginal Homogeneity test. These specific paired tests were conducted using IBM SPSS Statistics v. 27.0.1. Differences were considered statistically significant at $p < 0.05$.

Results. The distribution of the 302 clinical cases analyzed across the main specialties proportionally reflected the overall statistics for admissions to the hospital's intensive care units. The largest group consisted of patients with cardiac and pulmonary conditions (falling under the general medicine specialty) (n = 126; 41.7%), followed by patients with neurological and stroke-related conditions (n = 54; 17.9%). Trauma and general surgery accounted for 13.2% (n = 40) and 12.3% (n = 37) of cases, respectively. The remainder of the cohort consisted of pediatric (n = 34; 11.3%) and infectious disease (n = 11; 3.6%) cases. This diverse clinical structure confirms that the artificial intelligence model was tested on a wide range of complex pathophysiological conditions. Descriptive statistics for the quantitative and categorical variables in the analyzed data are presented in Tables 1 and 2, respectively.

Table 1. Descriptive statistics for quantitative variables.

Variable	Me	Q1–Q3	n	min	max
Age, years	64	48 – 72	302	0	85

ISSN 2524 – 0684 e-ISSN 2524 – 0692

pH	7.28	7.18 – 7.35	302	6.65	7.74
pCO_2, mmHg	52.5	41.0 – 64.5	302	15	110
pO_2 (arterial), mmHg	78.0	66.0 – 92.0	302	42	252
HCO_3^-, mmol/L	18.5	14.2 – 23.5	302	2	35
Base Excess (BE), mmol/L	-6.5	-11.0 – -2.5	302	-25	+10
Lactate, mmol/L	3.2	1.8 – 5.4	302	0.9	28
Anion gap, mmol/L	16.5	12.4 – 22.0	302	10	26

Table 2. Descriptive statistics for categorical variables.

Variables	**Categories**	**Abs.**	**%**	**95% CI**
Doctor's diagnosis	incorrect	12	4.0	2.1 – 6.8
	correct	215	71.2	65.7 – 76.2
	partially correct	75	24.8	20.1 – 30.1
Doctor's therapy	incorrect	15	5.0	2.8 – 8.1
	correct	184	60.9	55.2 – 66.5
	partially correct	103	34.1	28.8 – 39.8
AI’s diagnosis	incorrect	6	2.0	0.7 – 4.3
	correct	196	64.9	59.2 – 70.3
	partially correct	100	33.1	27.8 – 38.7
AI’s therapy	correct	271	89.7	85.7 – 92.9
	partially correct	31	10.3	7.1 – 14.3
Coincidence of diagnosis	No	18	6.0	3.6 – 9.3
	Yes	225	74.5	69.2 – 79.3
	Partially yes	59	19.5	15.2 – 24.5
Diagnostic accuracy	Doctor	30	9.9	6.8 – 13.9
	AI	43	14.2	10.5 – 18.7
	Both are correct	229	75.8	70.6 – 80.5
Coincidence of treatment	No	3	1.0	0.2 – 2.9
	Yes	152	50.3	44.5 – 56.1
	Partially yes	147	48.7	42.9 – 54.5
Appropriateness of treatment	Doctor	24	7.9	5.2 – 11.6
	AI	128	42.4	36.7 – 48.2
	Both are correct	150	49.7	43.9 – 55.5
Doctor's specialty	ICU doctor	302	100.0	98.8 – 100.0
Doctor’s years of experience	1–3 years	6	2.0	0.7 – 4.3
	4-5 years	245	81.1	76.2 – 85.4
	6-10 years	27	8.9	6.0 – 12.7
	10 years +	24	7.9	5.2 – 11.6
Gender (patients)	Male	173	57.3	51.6–62.7
	Female	129	42.7	37.3–48.4
Mechanical ventilation	No	187	62.0	56.3–67.2
	Yes	115	38.0	32.8–43.7
Hemodynamic status (Shock / Vasopressors)	No	168	55.5	50.0–61.1
	Yes	134	44.5	38.9–50.0
Type of Acid–Base Disorder	Metabolic acidosis	86	28.5	23.7–33.8
	Respiratory acidosis	94	31.1	26.2–36.6
	Mixed acidosis	89	29.5	24.6–34.8
	Metabolic alkalosis	14	4.6	2.8–7.6
	Respiratory alkalosis	5	1.7	0.7–3.8

Mixed alkalosis	0	0.0	0.0–1.3
Normal / compensated state	14	4.6	2.8–7.6

We analyzed the differences in diagnostic accuracy between doctors and AI (Table 3)

Table 3. Analysis of the relationship between AI diagnosis and the doctor's diagnosis.

Variable	Categories	Doctor's diagnosis			χ^2	df	p
		incorrect	correct	partially correct			
AI diagnosis	incorrect	0 (0.0)	6 (2.8)	0 (0.0)	8.068	2	0.018*
	correct	12 (100.0)	145 (67.4)	39 (52.0)			
	partially correct	0 (0.0)	64 (29.8)	36 (48.0)			

incorrect-correct χ^2 = 2.000. p = 0.157; **correct-partially correct** χ2 = 6.068. p = 0.014

According to the table presented, when comparing AI diagnoses, statistically significant differences were found depending on the doctor's diagnosis ($p < 0.05$) (applied method: McNemar-Bowker Test).

Analysis of AI's therapy was performed conditioning on doctor's therapy (Table 4).

Table 4. Analysis of AI's therapy conditioning on Doctor's therapy.

Variable	Categories	Doctor's diagnosis			χ^2	p
		incorrect	correct	partially correct		
AI's therapy	incorrect	9 (3.0)	159 (52.6)	103 (34.1)	4.492	< 0.001
	correct	6 (2.0)	25 (8.3)	0 (0.0)		

Statistically significant differences were revealed when comparing AI's therapy depending on Doctor's therapy ($p < 0.001$) (applied method: Marginal Homogeneity Test).

Discussion. Our study demonstrated that an adapted large language model based on ChatGPT is capable of achieving superior accuracy in therapeutic recommendations and demonstrating a lower rate of critical errors when assessing acid–base disorders compared to board-certified intensive care physicians.

According to the data we have received, the AI system demonstrated a significantly higher rate of correct therapeutic strategies (89.7% vs. 60.9% for physicians) and a lower rate of completely incorrect diagnostic conclusions (2.0% vs. 4.0%). Interestingly, board-certified physicians achieved a higher rate of fully correct diagnostic formulations (71.2% vs. 64.9%). This occurred primarily because the AI's highly detailed responses were often classified as 'partially correct' due to overly broad or redundant diagnostic wording, whereas experienced intensivists provided exact and concise clinical formulations. This highlights that while AI excels in standardizing therapy and calculating dosages, human clinical reasoning remains a significant role for precise diagnosis.

The advantages of AI lie in its ability to consistently process large volumes of data and rapidly analyze multiple parameters. However, it is crucial to acknowledge that AI systems are not free from algorithmic risks. As highlighted in recent studies, the use of large language models carries inherent limitations, including the potential for 'hallucinations', prompt sensitivity, automation bias, and the confident generation of incorrect clinical recommendations [19]. Therefore, AI should strictly remain an auxiliary tool. These findings add to the growing body of evidence that artificial intelligence-based systems can reduce variability in clinical decision-making and decrease the risk of iatrogenic errors [20,21].

A qualitative analysis revealed distinct differences between physician and AI responses. The physician reports were laconic, often without detailed explanations for specific decisions; specific dosages or ventilation parameters were sometimes omitted. In the physician group, the most common diagnostic errors involved missed mixed acid-base imbalances (e.g., failure to identify a secondary metabolic component in primary respiratory acidosis) and misinterpretation of compensatory mechanisms. Regarding therapeutic decisions, physicians primarily made errors related to the unnecessary administration of sodium bicarbonate when it was not clinically indicated, as well as inappropriate electrolyte correction.

In contrast, the AI model's responses were excessively detailed and specific, with each decision thoroughly justified. However, the AI occasionally provided questionable recommendations for certain drugs that had relative contraindications in the patient's condition, likely due to the AI's inability to contextualize unstated clinical nuances beyond the provided parameters.

From a theoretical perspective, the results support the possibility of modeling structured clinical reasoning using large language models. From a practical standpoint, implementing such systems as a “second opinion” tool may improve diagnostic reliability, promote standardization of treatment approaches, and provide support to less experienced clinicians [22]. This is particularly relevant in settings with limited resources and the need for rapid decision-making. Our findings are consistent with those of previous studies demonstrating the potential of large language models as clinical decision support tools, including improvements in diagnostic reasoning, treatment planning, and clinical decision-making support. [23,24]

Future research directions include prospective validation of these findings in real clinical settings, including randomized controlled trials to assess the impact of AI on patient outcomes.

Study limitations. It is important to note a number of limitations of this study. First, the retrospective design and single-center nature of the study may be associated with selection bias and limit the generalizability of the results. Second, the gold standard for evaluating the accuracy of diagnoses and therapeutic decisions was based solely on the expert assessment of a single senior intensivist. Furthermore, since the AI-generated responses differed in structure and were significantly more detailed than the concise responses from physicians, it was not possible to ensure a strict blinded assessment, which creates the potential for expert bias. In addition, the study evaluated text-based conclusions regarding clinical scenarios rather than actual clinical outcomes.

Conclusion. Our study shows that the use of an AI-based clinical decision support system is potentially able to complement physicians’ professional expertise and may help improve the accuracy of treatment recommendations. Although board-certified intensivists made more completely correct diagnoses for acid-base disorders, the AI system demonstrated a lower rate of completely incorrect conclusions. Furthermore, within the presented clinical scenarios, AI demonstrated greater accuracy in developing treatment plans for acid-base disorders, as evidenced by a higher number of completely correct conclusions compared to physicians, and no completely incorrect therapeutic recommendations were identified from the AI.

This study contributes to the development of AI technologies for use in intensive care. However, the results must be interpreted with specific study limitations in mind: the study had a retrospective, single-center design; a single expert without a strict blinded method conducted the evaluation; and the analysis focused exclusively on written responses to clinical scenarios rather than on actual clinical outcomes for patients. Thus, further prospective validation is required for the implementation of such AI-based systems into routine clinical practice.

Conflict of interest. The authors declare no potential conflicts of interest requiring disclosure in this article.

Authors' contribution. Writing – original draft preparation, Data curation: A.A.; Writing – review & editing, Conceptualization, Formal analysis: A.S.; Investigation, Software – S.T.; Methodology – Y.A.; Writing – review & editing, Supervision, Project administration, Validation, Visualization: S.T. All authors have read and

agreed to the published version of the manuscript. The authors declare that this material has not been previously published and is not under consideration by other publishers.

Funding. This research received no external funding.

Data availability statement. The data supporting the findings of this study are contained within the article. Additional data may be available from the corresponding author upon reasonable request.

LIST OF REFERENCES

1. Srivastava N, Verma S, Singh A, Shukla P, Singh Y, Oza AD, et al. Advances in artificial intelligence-based technologies for increasing the quality of medical products. Daru. 2024;33(1):1. https://doi.org/10.1007/s40199-024-00548-5.
2. Abdelmohsen SA, Al-Jabri MM. Artificial intelligence applications in healthcare: a systematic review of their impact on nursing practice and patient outcomes. J Nurs Scholarsh. 2025;57(6):957–66. https://doi.org/10.1111/jnu.70040.
3. Biesheuvel LA, Dongelmans DA, Elbers PWG. Artificial intelligence to advance acute and intensive care medicine. Curr Opin Crit Care. 2024;30(3):246–50. https://doi.org/10.1097/MCC.0000000000001150.
4. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716–20. https://doi.org/10.1038/s41591-018-0213-5.
5. Armoundas AA, Narayan SM, Arnett DK, Spector-Bagdady K, Bennett DA, Celi LA, et al. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the American Heart Association. Circulation. 2024;149(14):e1028–50. https://doi.org/10.1161/CIR.0000000000001201.
6. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. N Engl J Med. 2023;389(4):348–58. https://doi.org/10.1056/NEJMra2204673.
7. Moralez GM, Amado F, Liu VX, Tan SC, Meyfroidt G, Stevens RD, Pilcher D, Salluh JIF. Data-driven quality of care in the ICU: a concise review. Crit Care Med. 2025;53(12):e2720–28. https://doi.org/10.1097/CCM.0000000000006862.
8. Giri R, Firdhos SH, Vida TA. Artificial intelligence in anesthesia: enhancing precision, safety, and global access through data-driven systems. J Clin Med. 2025;14(19):6900. https://doi.org/10.3390/jcm14196900.
9. Kalimouttou A, Stevens RD, Pirracchio R. Harnessing AI in critical care: opportunities, challenges and key steps for success. Thorax. 2026;81(2):183–92. https://doi.org/10.1136/thorax-2024-222125.
10. Berkhout WEM, van Wijngaarden JJ, Workum JD, van de Sande D, Hilling DE, Jung C, Meyfroidt G, Gommers D, Buijsman SNR, van Genderen ME. Operationalization of artificial intelligence applications in the intensive care unit: a systematic review. JAMA Netw Open. 2025;8(7):e2522866. https://doi.org/10.1001/jamanetworkopen.2025.22866.
11. Pinsky MR, Bedoya A, Bihorac A, Celi L, Churpek M, Economou-Zavlanos NJ, Elbers P, Saria S, Liu V, Lyons PG, Shickel B, Toral P, Tscholl D, Clermont G. Use of artificial intelligence in critical care: opportunities and obstacles. Crit Care. 2024;28(1):113. https://doi.org/10.1186/s13054-024-04860-z.
12. Montomoli J, Hilty MP, Ince C. Artificial intelligence in intensive care: moving towards clinical decision support systems. Minerva Anestesiol. 2022;88(12):1066–72. https://doi.org/10.23736/S0375-9393.22.16739-8.
13. Fujii T, Udy AA, Nichol A, Bellomo R, Deane AM, El-Khawas K, et al. Incidence and management of metabolic acidosis with sodium bicarbonate in the ICU: an international observational study. Crit Care. 2021;25(1):45. https://doi.org/10.1186/s13054-020-03431-2.

14. Henrique LR, Souza MB, El Kadri RM, Boniatti MM, Rech TH. Prognosis of critically ill patients with extreme acidosis: a retrospective study. J Crit Care. 2023;78:154381. https://doi.org/10.1016/j.jcrc.2023.154381.
15. Mochizuki K, Fujii T, Paul E, Anstey M, Uchino S, Pilcher DV, Bellomo R. Acidemia subtypes in critically ill patients: an international cohort study. J Crit Care. 2021;64:10–17. https://doi.org/10.1016/j.jcrc.2021.02.006.
16. Canova TJ, Lipps K, Dahiya G, Hillerson DB, Kashani KB, Jentzer JC. Admission acid-base status and mortality in cardiac intensive care unit patients. J Intensive Care Med. 2025;8850666251399182. https://doi.org/10.1177/08850666251399182.
17. Huang Y, Ao T, Zhen P, Hu M. Association between the anion gap and mortality in critically ill patients with influenza: a cohort study. Heliyon. 2024;10(15):e35199. https://doi.org/10.1016/j.heliyon.2024.e35199.
18. Ciabattoni A, Chiumello D, Mancusi S, Pozzi T, Monte A, Rocco C, Coppola S. Acid-base status in critically ill patients: physicochemical vs. traditional approach. J Clin Med. 2025;14(9):3227. https://doi.org/10.3390/jcm14093227.
19. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med. 2025;5(1):330. https://doi.org/10.1038/s43856-025-01021-3.
20. Shi T, Ma J, Yu Z, Xu H, Yang R, Xiong M, et al. Large language models in critical care medicine: scoping review. JMIR Med Inform. 2025;13:e76326. https://doi.org/10.2196/76326.
21. Lovejoy CA, Buch V, Maruthappu M. Artificial intelligence in the intensive care unit. Crit Care. 2019;23(1):7. https://doi.org/10.1186/s13054-018-2301-9.
22. Turan Eİ, Baydemir AE, Balıtatlı AB, Şahin AS. Assessing the accuracy of ChatGPT in interpreting blood gas analysis results: ChatGPT-4 in blood gas analysis. J Clin Anesth. 2025;102:111787. https://doi.org/10.1016/j.jclinane.2025.111787.
23. Lu Y, Wu H, Qi S, Cheng K. Artificial intelligence in intensive care medicine: toward a ChatGPT/GPT-4 way? Ann Biomed Eng. 2023;51(9):1898–903. https://doi.org/10.1007/s10439-023-03234-w.
24. Jo E, Song S, Kim JH, Lim S, Kim JH, Cha JJ, Kim YM, Joo HJ. Assessing GPT-4's performance in delivering medical advice: comparative analysis with human experts. JMIR Med Educ. 2024;10:e51282. https://doi.org/10.2196/51282.

Information about authors

@Amrin Sabyrzhan Timuruly, 2nd-year master's student “KSPH” Kazakhstan Medical University LLP, Anesthesiologist and intensive care therapist of Talgar Central Regional Hospital, Almaty, Kazakhstan, e-mail: amrin.sabr@gmail.com, https://orcid.org/0000-0002-4347-1908.

Satbayeva Aknar Serikbaykyzy, resident, pediatric surgeon, S.D. Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan, e-mail: satbaeva.aknara@gmail.com, https://orcid.org/0000-0003-4545-114X.

Abdussemyatova Adilyam Abdussemyatovna, resident, anesthesiologist, Scientific Research Institute of Cardiology and Internal Diseases, Almaty, Kazakhstan, e-mail: adilyamabdussemyat@gmail.com, https://orcid.org/0000-0003-1145-7754.

Duisen Yerdos Adilbekuly, anesthesiologist, A.N. Syzganov National Scientific Center of Surgery, Almaty, Kazakhstan, e-mail: doseke195tc@gmail.com, https://orcid.org/0009-0009-0413-8430.

Авторлар туралы мәліметтер

@Амрин Сабыржан Тимурулы, «ҚДСЖМ» Қазақстан медициналық университетінің 2-курс магистранты, Талғар орталық аудандық ауруханасының анестезиолог-реаниматолог дәрігері, Алматы, Қазақстан, e-mail: amrin.sabr@gmail.com, https://orcid.org/0000-0002-4347-1908.
Сатбаева Акнар Серикбайкызы, резидент дәрігер, балалар хирургі, С.Ж. Асфендияров атындағы Қазақ ұлттық медицина университеті, Алматы, Қазақстан, e-mail: satbaeva.aknara@gmail.com, https://orcid.org/0000-0003-4545-114X.
Абдусемятова Адилям Абдусемятовна, резидент дәрігер, анестезиолог-реаниматолог, «Кардиология және ішкі аурулар ғылыми-зерттеу институты» АҚ, Алматы, Қазақстан, e-mail: adilyamabdussemyat@gmail.com, https://orcid.org/0000-0003-1145-7754.
Дуйсен Ердос Адилбекулы, анестезиолог-реаниматолог дәрігер, А.Н. Сызғанов атындағы Ұлттық ғылыми хирургия орталығы, Алматы, Қазақстан, e-mail: doseke195tc@gmail.com, https://orcid.org/0009-0009-0413-8430.

Сведения об авторах

@Амрин Сабыржан Тимурулы, магистрант 2-го курса Казахстанского медицинского университета «ВШОЗ», врач анестезиолог-реаниматолог Центральной районной больницы г. Талгар, Казахстан, e-mail: amrin.sabr@gmail.com, https://orcid.org/0000-0002-4347-1908.
Сатбаева Акнар Серикбайкызы, врач-резидент, детский хирург, Казахский национальный медицинский университет имени С.Д. Асфендиярова, Алматы, Казахстан, e-mail: satbaeva.aknara@gmail.com, https://orcid.org/0000-0003-4545-114X.
Абдусемятова Адилям Абдусемятовна, врач-резидент, анестезиолог-реаниматолог, АО «Научно-исследовательский институт кардиологии и внутренних болезней», Алматы, Казахстан, e-mail: adilyamabdussemyat@gmail.com, https://orcid.org/0000-0003-1145-7754.
Дуйсен Ердос Адилбекулы, врач анестезиолог-реаниматолог, Национальный научный центр хирургии имени А.Н. Сызганова, Алматы, Казахстан, e-mail: doseke195tc@gmail.com, https://orcid.org/0009-0009-0413-8430.

ҚЫШҚЫЛ-НЕГІЗДІК ТЕПЕ-ТЕҢДІКТІҢ БҰЗЫЛУЫНДАҒЫ ДИАГНОСТИКАЛЫҚ ДӘЛДІК ПЕН ТЕРАПИЯЛЫҚ ҰТЫМДЫЛЫҚ: ДӘРІГЕРЛЕР МЕН ЖАСАНДЫ ИНТЕЛЛЕКТІНІ САЛЫСТЫРМАЛЫ ТАЛДАУ

С.Т. АМРИН [1], А.С. САТБАЕВА [2], А.А. АБДУСЕМЯТОВА [3], Е.А. ДҮЙСЕН [4]

[1] «ҚДСЖМ» Қазақстан медициналық университеті»
[2] С.Ж. Асфендияров атындағы Қазақ ұлттық медицина университеті, Алматы, Қазақстан;
[3] Кардиология және ішкі аурулар ғылыми-зерттеу институты, Алматы, Қазақстан
[4] А.Н. Сызғанов атындағы Ұлттық ғылыми хирургия орталығы, Алматы, Қазақстан

Түйіндеме

Кіріспе. Қышқыл-сілтілік жағдайдың бұзылыстары қарқынды терапия бөлімшелерінде жиі кездесетін, жедел және дәл түзетуді талап ететін критикалық жағдайлар болып табылады. Бұл зерттеу ятерогендік қателерді азайту мақсатында клиникалық шешім қабылдауды қолдайтын қолжетімді жүйе ретінде үлкен тілдік модельдердің әлеуетін қарастырады.

Мақсаты. ChatGPT негізіндегі жасанды интеллект пен қарқынды терапия дәрігерлері ұсынған осы бұзылыстардың диагностикасының дәлдігін және терапиялық ұсынымдардың негізділігін салыстыру.

ISSN 2524 – 0684 e-ISSN 2524 – 0692

Материалдар мен әдістер. Зерттеу дизайны: ретроспективті, бір орталықты, салыстырмалы зерттеу. Талдауға 2024–2025 жылдар аралығында қарқынды терапия бөлімшесінде ем қабылдаған пациенттердің 302 клиникалық-зертханалық жағдайы енгізілді. Prompt engineering әдістерін қолдану арқылы «ReanimatorKZ» деп аталатын ChatGPT-тің бейімделген моделі әзірленді. Қышқыл-сілтілік тепе-теңдік бұзылыстары бойынша ChatGPT пен қарқынды терапия дәрігерлерінің қорытындыларына салыстырмалы сараптамалық бағалау жүргізілді. Статистикалық талдау StatTech v.4.12.7 және SPSS Statistics 27.0.1 бағдарламаларының көмегімен орындалды.

Нәтижелері. Дәрігерлер тобында диагностикалық дәлдік 71,2% толық дұрыс, 24,8% ішінара дұрыс және 4,0% қате қорытындыларды құрады. ChatGPT толық қате диагностикалық қорытындылардың төмен үлесін көрсетті (2,0%), ал толық дұрыс диагноздардың үлесі 64,9% болды. Дәрігерлердің емдік стратегиялары жағдайлардың 60,9%-ында толық дұрыс деп бағаланса, ChatGPT ұсынымдары 89,7% жағдайда толық дұрыс болды және толық қате ұсынымдар анықталған жоқ. Статистикалық тұрғыдан маңызды айырмашылықтар жұпталған тесттер арқылы расталды ($p < 0,05$).

Қорытынды. ChatGPT-тің бейімделген нұсқасы қышқыл-сілтілік тепе-теңдік бұзылыстарын анықтауда қарқынды терапия дәрігерлерімен салыстырмалы жоғары диагностикалық дәлдік көрсетті, сондай-ақ осы бұзылыстарға қатысты емдік ұсынымдарды қалыптастыруда жоғарырақ дәлдікке ие болды. Зерттеу нәтижелері кеңінен қолжетімді жасанды интеллект модельдеріне негізделген тиімді және оңай масштабталатын клиникалық шешім қабылдауды қолдау жүйелерін әзірлеу әлеуетін растайды. Алайда мұндай жүйелерді күнделікті клиникалық тәжірибеге енгізер алдында қосымша проспективті валидация қажет.

Түйінді сөздер: жасанды интеллект, қарқынды терапия, клиникалық шешім қабылдауды қолдау жүйелері, ChatGPT, диагностика дәлдігі.

ДИАГНОСТИЧЕСКАЯ ТОЧНОСТЬ И ТЕРАПЕВТИЧЕСКАЯ РАЦИОНАЛЬНОСТЬ ПРИ НАРУШЕНИЯХ КИСЛОТНО-ОСНОВНОГО СОСТОЯНИЯ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕЖДУ ВРАЧАМИ И ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ

С.Т. АМРИН [1], А.С. САТБАЕВА [2], А.А. АБДУСЕМЯТОВА [3], Е.А. ДҮЙСЕН [4]

[1] Казахстанский медицинский университет «ВШОЗ», Алматы, Казахстан
[2] Казахский национальный медицинский университет имени С.Д. Асфендиярова, Алматы, Казахстан
[3] Научно-исследовательский институт кардиологии и внутренних болезней, Алматы, Казахстан
[4] Национальный научный центр хирургии имени А.Н. Сызганова, Алматы, Казахстан

Аннотация

Введение. Нарушения кислотно-щелочного состояния являются критическими состояниями в отделениях интенсивной терапии, требующими быстрой и точной коррекции. Исследование рассматривает потенциал больших языковых моделей в качестве доступных систем поддержки принятия клинических решений для снижения ятрогенных ошибок.

Цель. Сравнить точность диагностики этих нарушений и обоснованность терапевтических рекомендаций, предлагаемых искусственным интеллектом на базе ChatGPT, и врачами интенсивной терапии.

ISSN 2524 – 0684 **e-ISSN 2524 – 0692**

Материалы и методы. Дизайн исследования: ретроспективное, одноцентровое, сравнительное исследование. Анализ включал 302 клинико-лабораторных случая пациентов, проходивших лечение в отделении интенсивной терапии в период с 2024 по 2025 годы. С использованием prompt engineering была настроена адаптированная модель ChatGPT под названием «ReanimatorKZ». Проведена сравнительная экспертная оценка заключений между ChatGPT и врачами интенсивной терапии при нарушениях кислотно-щелочного состояния. Статистический анализ выполнялся с использованием программ StatTech v.4.12.7 и SPSS Statistics 27.0.1.

Результаты. В группе врачей диагностическая точность составила: 71,2% полностью правильных, 24,8% частично правильных и 4,0% неправильных заключений. ChatGPT продемонстрировал более низкую долю полностью ошибочных диагностических заключений (2,0%), тогда как полностью правильные диагнозы составили 64,9%. Терапевтические стратегии врачей были полностью правильными в 60,9% случаев, тогда как рекомендации ChatGPT были полностью правильными в 89,7% случаев, при отсутствии полностью неправильных рекомендаций. Статистически значимые различия были подтверждены с использованием парных тестов ($p < 0,05$).

Заключение. Адаптированная версия ChatGPT продемонстрировала высокий уровень диагностической точности при выявлении нарушений кислотно-щелочного состояния, сопоставимый с таковым у врачей интенсивной терапии, а также более высокую точность в формировании терапевтических рекомендаций при данных нарушениях. Результаты нашего исследования подтверждают потенциал создания эффективных и легко масштабируемых систем поддержки принятия клинических решений на основе широкодоступных моделей искусственного интеллекта. Однако перед внедрением таких систем в рутинную клиническую практику необходима дополнительная проспективная валидация.

Ключевые слова: искусственный интеллект, интенсивная терапия, системы поддержки принятия клинических решений, ChatGPT, точность диагностики.